Studies in Choice and Welfare

Constanze Binder
Giulio Codognato
Miriam Teschl
Yongsheng Xu *Editors*

Individual and Collective Choice and Social Welfare

Essays in Honor of Nick Baigent

Springer

Studies in Choice and Welfare

More information about this series at
http://www.springer.com/series/6869

Constanze Binder • Giulio Codognato •
Miriam Teschl • Yongsheng Xu
Editors

Individual and Collective Choice and Social Welfare

Essays in Honor of Nick Baigent

Springer

Editors
Constanze Binder
Department of Philosophy
Erasmus University Rotterdam
Rotterdam
The Netherlands

Giulio Codognato
Department of Economics
University of Udine
Udine
Italy

Miriam Teschl
Aix-Marseille University
Aix-Marseille School of Economics,
CNRS & EHESS
Marseille
France

Yongsheng Xu
Department of Economics
Georgia State University
Atlanta
Georgia
USA

ISSN 1614-0311 ISSN 2197-8530 (electronic)
Studies in Choice and Welfare
ISBN 978-3-662-46438-0 ISBN 978-3-662-46439-7 (eBook)
DOI 10.1007/978-3-662-46439-7

Library of Congress Control Number: 2015938590

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media (www.springer.com)

Preface

This volume brings together papers, some of which were first presented at the Central European Program in Economic Theory (CEPET) Workshop held in honor of Nick Baigent at the University of Udine, Udine, Italy, on 2–4 June 2010. All the papers in the volume have gone through the usual process of review by anonymous referees. We have been helped by many individuals and institutions in organizing the Workshop and putting this volume together. We are grateful to the authors of this volume for contributing their papers and to the referees who reviewed the papers, and to Dr. Martina Bihn and Ms. Ruth Milewski of Springer-Verlag for their advice, help and patience. We also thank the University of Udine and CISM for hosting the workshop.

It has been a great pleasure and privilege for us to edit this volume to pay a tribute to Nick Baigent.

Rotterdam, The Netherlands — Constanze Binder
Udine, Italy — Giulio Codognato
Marseille, France — Miriam Teschl
Atlanta, USA — Yongsheng Xu
2014

Contents

Contributors

Ritxar Arlegi Economics Department, Public University of Navarre, Pamplona, Spain

Constanze Binder Faculty of Philosophy, Erasmus University Rotterdam, Rotterdam, The Netherlands

Walter Bossert Department of Economics and CIREQ, University of Montreal, Montreal, QC, Canada

Richard Bradley Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London, UK

Francesca Busetto Dipartimento di Scienze Economiche e Statistiche, Universitá degli Studi di Udine, Udine, Italy

Burak Can Department of Economics, School of Business and Economics, Maastricht University, MD Maastricht, The Netherlands

Giulio Codognato Dipartimento di Scienze Economiche e Statistiche, Universitá degli Studi di Udine, Udine, Italy

EconomiX, Université de Paris Ouest Nanterre la Défense, Nanterre, France

Marc Fleurbaey Princeton University, Princeton, NJ, USA

Elisabeth Gugl Department of Economics, University of Victoria, Victoria, BC, Canada

Mark R. Johnson Department of Finance and Economics, A. B. Freeman School of Business, Tulane University, New Orleans, LA, USA

Ipek Oezkal-Snaver Department of Economics, Istanbul Bilgi University, Murat Sertel Center for Advanced Economic Studies, Beyoğlu, Turkey

Pelin Pasin Department of Economics, Izmir Katip Celebi University, İzmir, Turkey

Prasanta K. Pattanaik Department of Economics, University of California, Riverside, CA, USA

Jonathan Riley Department of Philosophy, Tulane University, New Orleans, LA, USA

M. Remzi Sanver Department of Economics, Istanbul Bilgi University, Murat Sertel Center for Advanced Economic Studies, İstanbul, Turkey

Norman Schofield Center in Political Economy, Washington University in Saint Louis, Saint Louis, MO, USA

Richard Sturn University of Graz, Institute of Public Economics, Graz, Austria

Kotaro Suzumura School of Political Science and Economics, Waseda University, Tokyo, Japan

Miriam Teschl Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Centre de la Vieille Charité, Marseille, France

Simone Tonin Adam Smith Business School, University of Glasgow, Glasgow, UK

Anne van den Nouweland Department of Economics, University of Oregon, Eugene, OR, USA

Martin van Hees Department of Philosophy, VU University Amsterdam, HV Amsterdam, The Netherlands

Roberto Veneziani School of Economics and Finance, Queen Mary University of London, London, UK

Hanji Wu Department of Economics, Georgia State University, Atlanta, GA, USA

Yongsheng Xu Department of Economics, Georgia State University, Atlanta, Georgia, USA

Naoki Yoshihara The Institute of Economic Research, Hitotsubashi University, Tokyo, Japan

Zhen Zhong Research Institute of Finance and Banking, The People's Bank of China, Beijing, China

Introduction

Constanze Binder, Giulio Codognato, Miriam Teschl, and Yongsheng Xu

Nicholas Baigent received his doctoral degree in economics from the University of Essex in 1986. He has taught at various institutions, including Bedford College, University of London (1972–1974), University of Reading (1974–1975), Goldsmith College, University of London (1975–1976), University College of Swansea, University of Wales (1977–1981), University of Aarhus (1981–1982), University of Essex (1982–1984), Cornell University (1984–1985), Tulane University (1985–1993), and Graz University (1993–2011).

Nick is well-known mainly as a choice theorist. His research has focused on social choice theory, particularly topological theories of social choice, norms and rationality of choice, and normative public economics. His academic papers have appeared in leading economic journals including *Quarterly Journal of Economics*, *Journal of Mathematical Economics*, *Economic Theory*, *Japanese*

C. Binder (✉)
Faculty of Philosophy, Erasmus University Rotterdam, Campus Woudestein, Postbus 1738, 3000 DR Rotterdam, The Netherlands
e-mail: binder@fwb.eur.nl

G. Codognato
Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Udine, Via Tomadini 30, 33100 Udine, Italy

EconomiX, Université de Paris Ouest Nanterre la Défense, 200 Avenue de la République, 92001 Nanterre Cedex, France
e-mail: giulio.codognato@uniud.it

M. Teschl
Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Centre de la Vieille Charité, 2, rue de la Charité, 13002 Marseille, France
e-mail: miriam.teschl@ehess.fr

Y. Xu
Department of Economics, Georgia State University, Atlanta, GA, USA
e-mail: yxu3@gsu.edu

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_1

Economic Review, *Economics Letters*, *Analyse & Kritik*, *Mathematical Social Sciences*, *Social Choice and Welfare*, and *Theory and Decision*.

Nick has made invaluable contribution to the profession through his initiative and involvement in various activities that have inspired many young students and young researchers to be interested in doing research in economic theory and have guided them to flourish at the early but crucial stages of their careers. As a concrete initiative and involvement, Nick, together with Giulio Codognato, co-founded the Central European Program in Economic Theory (CEPET) in 1999. The program has been an excellent vehicle for young researchers to launch their academic careers through its annual summer workshop series.

The papers included in this volume try to represent the varied interests that reflect Nick Baigent's work, with a focus on the intersection of our research interests with his. For organizational purpose, we divide the volume into three parts marking Nick's main research interests.

Part I Individual Choice and Rationality

One of Nick's main research interests is to examine how norms, either individual or social, may be incorporated into rational choice theory to develop a satisfactory theory to explain an individual's choice behavior. It may be recalled that Nick, together with Wulf Gaertner, set the stage for the axiomatic study of norm-constrained choice behavior (Baigent and Gaertner 1996). The first part of the volume consists of five papers on rationality and theory of rational choice.

The paper, "Conflicts in decision making", by Ritxar Arlegi and Miriam Teschl argues that one must go "behind the veil of preference" following the argument by Baigent (1995) to be able to develop a satisfactory theory of rational behavior. They present an extensive overview over psychological research on intra-personal conflict, its influence on preference reversal and incoherent behavior on psychological well-being as well as on motivational and behavioral changes over time, and discuss a theory of choice under conflicting motivations based on Arlegi and Teschl (2012).

Richard Bradley's paper, "A note on incompleteness, transitivity and Suzumura consistency", contributes to the literature on rational choice theory. As Bradley points out, rationality does not require of preferences that they be complete or transitive. The paper examines the implications of these claims concerning rationality for the theory of rational choice. For this purpose, Bradley proposes a new choice rule-Strong Maximality-and argues that it better captures rational preference-based choice than other more familiar rules. A notion of rationality, Suzumura consistency of preferences, is shown to be both necessary and sufficient for non-empty strongly maximal choice. Finally, conditions on a choice function are stated that are necessary and sufficient for it to be rationalizable in terms of a Suzumura consistent preference relation.

In their paper, "Rationality and context-dependent preferences", Prasanta Pattanaik and Yongsheng Xu, take on the standard theory of rational choice in economics that considers an agent's choices to be rational if and only if the agent makes her choices in different choice situations on the basis of a fixed preference ordering defined over the set of all possible options, and explore the implication of the standard theory on a rational agent's preferences: a rational agent's preferences cannot be context-dependent. They outline a simple framework for defining context-dependence of preferences and for discussing relationships between context-dependent preferences and the notion of rationality, and examine some consequences.

The paper, "Moral responsibility and individual choice", by Constanze Binder and Martin van Hees examines the assessment of moral responsibility in individual choice situations. They define a responsibility function as a mapping that assigns to each choice situation those alternatives for which a person can be held responsible if she were to choose them in that situation. They then examine the conditions under which a responsibility function can be rationalized by a responsibility relation, that is, a relation describing the reasonableness of the various choice options. One result coming out of the analysis is that, when the standard of value underlying a person's responsibility relation coincides with a person's preferences, a person cannot be deemed responsible for choosing her uniquely most preferred option. They make use of a result obtained in Baigent and Gaertner (1996) to discuss and characterize one possible way out of this seemingly paradox.

Mark Johnson's paper, "A primer on economic choice automata", examines rational choice theory from a perspective of choice automata, and presents a development of the transformation semigroup of economic choice automata as a subgroup of the semigroup (monoid) of partial functions defined over the states of a finite state machine. The classes of consistency behavior considered are those rationalized by linear orders, weak orders, quasi-transitive relations and non-rationalizable path independent choice functions. For each of these classes of choice behavior, a particular class of lattice is identified as the action semigroup that drives the automaton. Given these characterizations, several features of the choice behavior are considered. In particular, the simplifying interval property of path independent choice, the importance of the distributive property of quasi-transitive rational choice in reducing the complexity of dynamic choice is addressed. Based on the algebraic structure of semi-automata implementing path independent choice functions it is possible to rank these semi-automata by the mathematical power required to implement a particular class of choice functions. This provides a means for ranking these machines by their "implementation complexity". Dually, the computation complexity of constructing a semi-automaton that implements a particular class of choice functions is investigated. It is seen that these complexities are inversely related.

Part II Collective Choice and Collective Rationality

The second part of the volume consists of seven papers on collective choice and collective rationality. These papers reflect Nick's broad research interests in social choice theory.

The paper by Walter Bossert and Kotaro Suzumura, *Multi-profile intertemporal social choice: a survey*, provides a brief survey of some literature on intertemporal social choice theory in a multi-profile setting. As is well-known, Arrows impossibility result hinges on the assumption that the population is finite. For infinite populations, there exist nondictatorial social welfare functions satisfying Arrows axioms and they can be described by their corresponding collections of decisive coalitions. Bossert and Suzumura review contributions that explore whether this possibility in the infinite-population context allows for a richer class of social welfare functions in an intergenerational model. Different notions of stationarity formulated for individual and for social preferences are examined.

The paper, "Minimal Maskin-monotonic extensions of tournament solutions", by Pelin Pasin, Ipek Ozkal-Sanver and Remzi Sanver, studies the minimal Maskin monotonic extensions of Condorcet consistent solutions. As it is known, given a neutral Condorcet consistent tournament solution the minimum number of alternatives that has to be beaten to be a winner at some tournament identifies the alternatives that are in the extension at this tournament. For the top-cycle, the uncovered set, the iterated uncovered set and the minimal covering set this number is equal to 1 which implies that at each tournament all the alternatives except the Condorcet looser is contained in the minimal monotonic extension. For the Copeland rule, however, this number depends on the number of alternatives over which the tournament is defined and is greater than 1 if there are 4 or more alternatives. They also determine the minimal Maskin monotonic extensions of the social choice rules that are generated by some Condorcet consistent solutions, namely, the top-cycle, the uncovered set, the iterated uncovered set and the minimal covering set.

The paper, *Single-profile choice functions and variable societies: characterizing approval voting* by Hanji Wu, Yongsheng Xu and Zhen Zhong, studies approval voting in a setting with a fixed profile of individuals choices and variable societies. They introduce four properties each linking choices made by a group of individuals to choices by its various subgroups, and use them to characterize approval voting. The paper extends the framework of studying approval voting in an environment of choice functions as introduced by Nick in his early study of approval voting.

Francesca Busetto, Giulio Codognato and Simone Tonin in their paper entitled "Nondictatorial Arrovian social welfare functions: an integer programming approach" explore new conditions on preference domains which make it possible to avoid Arrow's impossibility result as initiated by Kalai and Muller (1977). The paper provides a complete characterization of the domains admitting nondictatorial Arrovian social welfare functions with ties (i.e. including indifference in the range) by introducing a notion of strict decomposability. The main innovation lies in the proof where they use integer programming tools as first introduced by Sethuraman,

Teo and Vohra (2003, 2006) in the social choice literature. In the process, the paper generalizes Sethuraman et al.'s work and specifies integer programs in which variables are allowed to assume values in the set $\{0, 1/2, 1\}$. In particular, the paper shows that there exists a one-to-one correspondence between the solutions of an integer program defined on this set and the set of all Arrovian social welfare functions without restrictions on the range.

Burak Can's paper, "Distance rationalizability of scoring rules", examines collective decision making problems from the perspective of finding an outcome that is "closest" to a concept of "consensus". In particular, he shows that all non-degenerate scoring rules such as the Borda rule and the Dodgson rule can be distance-rationalized as "Closeness to Unanimity" procedures under a class of weighted distance functions introduced in the literature. Consequently, the results in this paper generalize the notion of "Closeness to Unanimity Procedure" introduced in the literature, and builds a connection between scoring rules and a generalization of the Kemeny distance, i.e. weighted distances.

Norman Schofield's paper, "Climate change and social choice theory", attempts to analyze the issue of climate change from the perspective of social choice theory. In particular, the paper surveys recent results in social choice which suggests that chaos rather than equilibrium is generic. In contrast to these results on chaos, Condorcet's Jury Theorem suggests that majority rule provides an ethical mechanism for a society to make a wise choice as long as every one perceives a common good underlying the choice. It is suggested that a belief equilibrium with regard to the appropriate response to climate change depends on the creation of a fundamental social principle of "guardianship of our planetary home."

In the paper, *Relevant irrelevance: the relevance of independence of irrelevant alternatives in family bargaining* by Elizabeth Gugl, she stresses the importance of the axiom of independence of irrelevant alternatives in family bargaining models when utility profiles do not lead to almost transferable utility. Independence of Irrelevant Alternatives (IIA) says that if a bargaining solution picks a point in the utility possibility set that is still available when some of the previously feasible points are removed, the bargaining solution applied to the new smaller set must again pick the point that was selected in the larger set. While IIA is not always persuasive and has been dropped in favor of other axioms as, for example, in the case of the Kalai-Smorodinsky solution, she demonstrates its appeal in the context of household decision making.

Part III Social Welfare and Equilibrium

The papers in Part III are concerned about some important issues relating to (1) the notion of utility or welfare in economics, and more generally various ethical evaluations of economic and public policies (the papers by Fleurbaey, Veneziani and Yoshihara, Sturn, and Riley), (2) the development of the Lindahl equilibrium in public economics (the paper by van den Nouweland).

In the paper, "Forced trades in a free market" by Marc Fleurbaey, he examines the idea that a free trade is always Pareto-improving. It argues that some "free trades" are actually forced in the sense that they reflect the trader's poverty rather than his or her preferences. He then proposes a rigorous concept of forced trade, and applies it to the ethical evaluation of Walrasian equilibria.

In their paper, "Unequal exchange, assets, and power: recent developments in exploitation theory", Roberto Veneziani and Naoki Yoshihara summarize and extend some recent contributions on the theory of exploitation as the unequal exchange of labour. They introduce a model of dynamic economies with heterogeneous optimising agents to encompass the models used in the literature as special cases, and show that the notion of exploitation is logically coherent and can be meaningfully analysed in such a general framework. They also show that the axiomatic approach of social choice theory can be adopted to explore the normative foundations of the notion of exploitation, and provide an argument against purely distributive approaches to exploitation.

Richard Sturn's paper, "The merits of merit wants", discusses the multi-faceted nature of the concept of merit wants by cutting through a complex array of problems associated with different levels of analysis. The concept of merit wants, according to Sturn, is considered as a shorthand notion for concerns that are respectable and important in a broadly individualist conception of welfare. He then questions why merit wants are not a firmly established part of modern normative economics, given that simplifying, but still meaningful notions are suitable as conceptual starting point for a research program. In this paper, he tries to link the answer to this question with making explicit three levels of problems (limits of reason, higher order preferences, collective choice) which may be useful to locate and scrutinize various interpretations of and approaches to merit wants.

Utilitarianism has a long tradition in economics. John Stuart Mill was an important and influential contributor and innovator in the development of classical utilitarianism. One of Mill's innovative argument for utilitarianism is his distinction of high pleasure from low pleasure. The notions of high pleasure and low pleasure are controversial, to say the least. Jonathan Riley's paper, "An extraordinary maximizing utilitarianism", defends Mill's version of utilitarianism. In particular, Riley reflects on Mill's view of higher pleasures and presents a defense of Mill. Given the controversy surrounding Mill's notion of high pleasure, the paper is a very representative interpretation of Riley's reading of Mill.

The last paper in this volume concerns about the development of the Lindahl equilibrium in public economics. Anne van den Nouweland, in her paper *Lindahl and Equilibrium*, presents a brief account of the development of the ideas expressed by Lindahl (1919) into an equilibrium concept for public good economies that is now known as Lindahl equilibrium. She also re-examines a seemingly forgotten equilibrium concept for public good economies known as ratio equilibrium and explains that from an axiomatic perspective this equilibrium concept is a better fit with the ideas expressed in Lindahl (1919).

Part IV An Interview with Nick Baigent

The volume concludes with an interview with Nick Baigent conducted by Constanze Binder, Miriam Teschl and Yongsheng Xu via email in the Summer and Fall of 2014.

Part I
Individual Choice and Rationality

Conflicts in Decision Making

Ritxar Arlegi and Miriam Teschl

Abstract Following Nick Baigent's argument that one must go "behind the veil of preference" (Baigent, Jpn Econ Rev 46(1):88–101, 1995) to be able to develop a satisfactory theory of rational behaviour, we propose to analyse potential intrapersonal conflicts caused by different reasons, goals or motivations to choose one option over another, which may make the development of a coherent preference impossible. We do this by presenting an extensive, but certainly not exhaustive overview of psychological research on intrapersonal conflict, its influence on preference reversal (and hence on incoherent behaviour), on psychological well-being and on motivational and behavioural changes over time. We then briefly describe our own theory of choice under conflicting motivations (Arlegi and Teschl, Working Papers of the Department of Economics DT 1208, Public University of Navarre, 2012), which is a first attempt at putting psychological insights into intrapersonal conflict into an axiomatic economic context.

Keywords Goals • Intrapersonal conflict • Motivations • Multiple self • Preference reversal • Want/should-self

1 Introduction

In "Behind the veil of preferences", Nick Baigent [3] makes a number of important observations about the plausibility of having or revealing preferences of which many economists themselves are not necessarily aware. This is probably so because the

R. Arlegi (✉)
Economics Department, Public University of Navarre, Campus de Arrosadia, 31006 Pamplona, Spain
e-mail: rarlegi@unavarra.es

M. Teschl
Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Centre de la Vieille Charité, 2, rue de la Charité, 13002 Marseille, France
e-mail: miriam.teschl@ehess.fr

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_2

common idea is that economics starts with given preferences, without questioning where those preferences come from. Students of economics are taught from early on that people *have* their own complete, transitive preferences, which enable a numerical *utility function* that represents such preferences to be defined, and that people choose what is best for them (i.e. they maximise their utility function) under the constraints that they face. It is assumed that people find out about their preferences through introspection.[1] Following Mas-Colell et al. [34], we call this the *preference-based approach*. Later, students are introduced to the concept of choice functions and to the idea of consistent behaviour, and that if people act consistently, such choices can be rationalised by a revealed preference ordering. Again according to Mas-Colell et al. [34], this is known as the *choice-based approach*. The "circle" seems to be closed: preferences are underlying choices, and can be revealed from those choices. Samuelson himself said: "The complete logical equivalence of [the revealed preference] approach with the regular Pareto-Slutsky-Hicks-Arrow ordinal preference approach has essentially been established. So in principle there is nothing to choose between the formulations. There is, however, the question of convenience of different formulations." [44, p. 1]. This "logical equivalence" however (and for that matter the term "preference") is a factor that may cause confusion. For example, as Baigent [3] highlights: "It is important to emphasize though, that the preference so revealed is not a preference that exists as a separate entity, distinct from the choices that reveal it. In fact, such a preference is only a description of choice and not an entity that has any independent existence." (p. 90). This means that while the preference-based approach assumes that preferences exist in the person in terms of their own "tastes",[2] the choice-based approach does not suggest that revealed preferences need to be such *person-inherent* or *intrinsic* preferences, in the sense of reflecting the person's tastes (which may include her interests, personal goals, etc.). Revealed preferences are just an ordering of alternatives, which may be based on *intrinsic* preferences or tastes, but also on other reasons such as norms, rules, obligations, etc. that make the person act consistently. However, economists usually do not say anything about what the reasons for "revealed preferences" may be. As Ken Binmore [9] would say: "The theory of revealed preference therefore makes a virtue of assuming nothing whatever about the psychological causes of our choice behavior." (p. 8).

But, as Amartya Sen [48] has pointed out, if nothing is assumed about the causes of behaviour, what is the rationale of imposing consistency on people's choices? Sen argues that there is no such thing as an "internal consistency of choice", which may be translated as consistency for its own sake or logical consistency. On the

[1] In regard to introspection, Mas-Colell et al. [34] write: "Introspection quickly reveals how hard it is to evaluate alternatives that are far from the realm of common experience. It takes work and serious reflection to find out one's own preferences." (p. 6).

[2] Mas-Colell et al. [34] for example write: "The [preference-based approach] treats the decision maker's tastes, as summarized in her *preference relation*, as the primitive characteristic of the individual." (p. 5).

contrary, choices are often made with respect to some reason, which Sen calls an "external reference".[3] For example, if choices are induced by maximising an "intrinsic" preference which, as Baigent [3] stresses, "exists separately from the choices it induces" (p. 90), then there is good reason to think that choices will satisfy consistency requirements, and more precisely those that are usually assumed by the choice-based approach. But there is no good reason to assume that people only want to maximise their "intrinsic" preferences. Hence, "the problem with the revealed-preference approach is that some choices do not reveal a *preference*"[4] [3, p. 89]. This may actually be true in two senses: first that people act consistently according to standard economics, but are motivated by reasons other than their "intrinsic" preferences, though they happen to satisfy those same consistency axioms. Second, people may not act consistently in the standard sense, which is usually considered to be irrational behaviour, but on the basis of something else, which happens not to satisfy those consistency axioms. In this case, one would need to see what people are trying to do—and if possible establish the consistency axioms that represent those reasons (see e.g. Baigent and Gaertner [4], Gaertner and Xu [19]).

The next question obviously is what happens if a person acts on grounds of several reasons, i.e. if she had multiple choice criteria? This, in principle, is no problem to economists. As Baigent [3] notes, it is probably most widely assumed, yet seldom fully articulated, that preferences in economics are considered to be "all-things-considered" (ATC). That is, a person may have multiple cares and concerns expressed in terms of different rankings and "[n]o doubt in economics many are inclined to think that a rational agent would aggregate the underlying rankings by weighting them and by making trade-offs, thus obtaining an ATC-preference." [3, p. 92]. This would also mean that choice necessarily implies the existence of a trade-off that can be used to determine an ATC-preference ordering. But Baigent shows with a simple example that this argument is false and concludes "[s]ince, therefore, choices need not reveal a preference at all, they certainly need not reveal a trade-off." (p. 93). Moreover, to establish an ATC-ordering, it is normally assumed that the underlying rankings are complete. However, this may not necessarily be the case. If a person acts on the basis of several concerns, neither they themselves, nor eventually the ATC-ordering, need be complete. Baigent takes

[3] Sen [48] says: "Statements *A* and *not-A* are contradictory in a way that choosing x from $\{x, y\}$ and y from $\{x, y, z\}$ cannot be. If the latter pair of choices were to entail respectively the statements (1) x is a better alternative than y, and (2) y is a better alternative than x, then there would indeed be a contradiction here (assuming that the content of "being better requires asymmetry). But those choices do not, *in themselves*, entail any such statements. *Given* some ideas as to what the person is trying to do (this is an external correspondence), we might be able to "interpret" these actions as implied statements. But we cannot do that without invoking such an external reference. There is no such thing as *purely* internal consistency of choice." (p. 499).

[4] What we call here an *intrinsic* preference.

the example of a person who has to choose between several job options in different locations and who takes into account the happiness of her family members as well as the variety of leisure activities in those locations. In both cases, there may not be a complete ranking for each of these criteria and thus an ATC-ranking is impossible to achieve by means of trade-offs.

This analysis clearly indicates that economists, who start their models with a given utility function, and especially those who use functions that contain multiple concerns such as fairness considerations or social norms etc. in addition to people's own intrinsic preferences are taking an extreme shortcut. Nothing guarantees that such utility functions actually exist. They may exist in certain cases if the individual happens to have complete orderings underlying their different concerns and has been able to assign weights and/or to form trade-offs to obtain a complete ATC-ordering, which is represented by that particular utility function; but this may not hold for many cases. This way of proceeding is comparable with the idea of first throwing the dart and then drawing the dartboard. It certainly works, but the question is in how far it really depicts and explains human behaviour. It is for this reason that we have always been susceptible to Baigent's suggestion that "[...] the starting point for the theory of rational choice should not generally be an ATC-preference, but the underlying cares and concerns that lie well behind the veil of an ATC-preference." (p. 95) and his follow-up question: "How are individuals to be characterized, given that, [...] characterization in terms of an ATC-preference is not generally satisfactory?" (p. 95). Baigent himself suggests for example characterising people as norm-holding individuals. Consider a cake cut into different pieces that can be ranked from the smallest to the largest. Standard economic theory would suggest that people prefer to eat the largest piece of cake. However, the norm says *not to choose the single largest one* and thus goes against their intrinsic preferences. Their choice problem, if they give lexical priority to the norm over intrinsic preferences, can then be represented as a "norm constrained (intrinsic) preference optimisation", which however is not consistent with the optimisation of any preferences. But the question is, why should people *always* give lexical priority to the norm? Why should they "prefer" the norm over their intrinsic preferences in all circumstances? One may rather think that if an ATC-preference is not generally a good characterisation of individuals it is because they may experience a conflict between their intrinsic preferences and obeying norms and may therefore not know how best to attribute weights to each of these concerns. That is, people may be torn between different reasons for choosing one option over another and thus experience some kind of internal conflict. The question, which interests us in particular in this paper, then, is how best to characterise individuals if they experience conflict and may not be able to say to which of the reasons and motives they would give priority. This is a largely unexplored question in economics, but it has received substantial attention in psychology in terms of "motivational" or "intrapsychic" conflict.

In the next section therefore, we introduce some of the psychological research on conflict. The section is subdivided into three main parts: the first describes the consequences of intrapersonal conflict on behaviour and choice, the second describes the influence of conflict on well-being and the third presents a more

general view in psychology of human development, which is assumed to be marked by different kinds of conflicts that induce people to change their behaviour or their concerns over time. In section three, we briefly introduce our theory of choice under internal conflict [1] as a first attempt to integrate this psychological research into an economic, axiomatic framework. We do this in a rather descriptive way, but add formal notations where necessary. The last section presents our conclusions.

2 Conflict in Psychology

2.1 Conflict and Preference Reversal

Of course, intrapersonal conflict is not unknown in economics. The main intrapersonal conflict discussed in economics is one where an individual acts against her longer-term interest by engaging in pleasant, enjoyable or more satisfactory actions in the present, which may however harm her well-being in the future. Strotz's [50] classic paper seems to have been the first to discuss this phenomenon in terms of a non-exponential discount function, which predicts impulsive and myopic behaviour in the present but more "considered" behaviour in the future. Such inconsistency has also been framed in terms of a dual or multiple self problem, where the myopic acting self has to be controlled by the more informed planning self [8, 15, 18, 45, 46, 51, etc.]. The source of the conflict in such models is the passage of time. These models have however been criticised for several reasons. One, as Loewenstein [31] points out, is that multiple self models are metaphorical and not an actual description of what happens within individuals. This, of course, may not bother economists too much because, as Schelling admits, only when talking to economists does he feel secure using the terminology of selves [47, p. 74], thus suggesting that economists have less difficulties in thinking about the economic agent as a succession of different selves as may be the case in other social sciences. This may be so because, as Loewenstein also points out, "[t]he strength of multiple self models is that they transfer insights from a highly developed field of research on interpersonal interactions to the less studied topic of intraindividual conflict." (p. 288). But the analogy of intrapersonal conflict with interpersonal conflict does not always fully capture the "nature" of the former. People are able to punish or control each other to avoid conflict in a way that is not possible among "multiple selves". Loewenstein himself sees conflicts more in terms of visceral factors such as hunger, thirst and sex drive, but also emotional states such as anger or fear affecting people's decisions. This has the advantage of explaining in what situations impulsive behaviour occurs, whereas non-exponential discounting literature has difficulties in explaining situations or reward-specific outbursts of impulsiveness. This is also the case because the only source of the problem in non-exponential discounting and multiple self models is time delay, whereas physical proximity and sensory contact can also be associated with impulsiveness. "[I]t is difficult to explain the impulsive

behavior evoked by cookie shops that vent baking smells into shopping malls in terms of hyperbolic discounting." [31, p. 279].[5]

The research group around Max Bazerman (e.g. [5–7, 35–37, 42]) has also published a number of papers, including reports on experiments, to highlight conflictual decision making and argue that conflict has more than just a temporal dimension. They observe that many people do not *want* to exercise, but know that they *should* do so. Eve *wanted* the apple, but knew that she *should* not eat it [6, p. 225]. Hence they argue that many decisions can be described as situations in which a *want* self and a *should* self struggle with each other. Broadly in line with Loewenstein [31], they see the *want* self as being more emotional and impulsive, and the *should* self as more rational and thoughtful. The above examples are compatible with a multiple or dual self view with the typical short-term and long-term interest conflict, but Bazerman and colleagues argue that their want/should self distinction may encompass more decision problems and preference reversal phenomena than can be explained with the temporal perspective of the multiple self model. One preference reversal that their want/should model can explain is that which is observed in joint versus separate evaluation problems. It has been noticed that people tend to choose one option if a decision problem presents them with a single choice option, but another when they are confronted with several possibilities at once. For example, Bazerman et al. [5] offered subjects (second-year MBA students) the choice either between six single job offers (separate evaluation) or three pairs of offers (joint evaluation). The job offers were set up to create a conflict between procedural justice concerns and the maximisation of their salary. The results of this experiment and others have consistently brought to light that people tend to choose the *want* option in separate evaluations (which in this case is said to be the job with the justice aspect),[6] but the *should* option in joint evaluations (the maximising salary option). O'Connor et al. [37] tested this theory with respect to the ultimatum game by creating two particular conditions, namely one in which subjects had to answer the question "What do you want to do?" (*want* condition) and another in which they were asked "What do you should do?" (*should* condition) either before, during or after responding to a 1$ offer (out of 10$). In agreement with their hypothesis, they observed that more individuals rejected the offer in the *want* conditions than in the *should* condition. In another experiment, O'Connor et al. [37] also sought to learn which of the two responses (*want* or *should* response) people preferred and found out that most people would rather like to act more thoughtfully and follow their own insights about what they *should* do in conflictual

[5]It should be clear by now that psychologists are far from imposing a strict preference structure in the economic sense on the individual (no consistency is imposed on people's choices, from which their preferences are revealed). Psychologists usually assume much simpler behavioural factors, such as different motivations or impulses, sometimes triggered and changed by varying contextual effects. These are therefore on a much more elementary level than the concept of "preferences" in economics.

[6]Bazerman at al. [6] note that it has been argued in procedural justice literature that procedural injustice creates an emotive (want) response.

situations, but when they are caught in a particular conflictual situation they respond with their *want* option. These and other experiments thus underline two particular issues, namely that in the heat of the moment people tend to give *want* responses, although in most cases they would have liked to give a *should* response, and that the availability of more options invites a more rational reflection and allows more people to choose their *should* option. Similarly, Rogers and Bazerman [42] find that people report stronger support for *should* policies when those policies are to be implemented in the distant future rather than in the near future. They call this finding the "future lock-in effect".[7]

Interestingly, the want/should explanation of the preference reversal concerning separate versus joint evaluations seems to be reversed when people know whether or not they will engage in a sequence of similar decision problems. Khan and Dhar [29] conducted an experiment in which they observed that a larger number of people tended to choose the vice (or *want*) good rather than the virtuous (or *should*) good (e.g. a lowbrow entertainment film versus a highbrow documentary) if they know that they face repeated choices of the same kind than if they have to make a one-shot decision.

Khan and Dhar [29] explain their results by arguing that people tend to be overly optimistic about their future behaviour in a repeated choice situation. These findings stand in contrast to the observation expressed by others that fragmenting a stream of activity into isolated choices encourages impatient choices [32]. Related to this, Prelec and Herrnstein [40] refer to this kind of problem as one where there is a "scale mismatch" (p. 322), which means that one element in an evaluation appears to have an impact only in the aggregate. For example, one may have decided to "always buckle up" in the car, but in fact the decision whether to use the seat-belt has to be faced each time one drives a car and it is not self-evident that it is followed every time. What emerges clearly from these studies is that choices made in isolation differ from those made sequentially. But they may also differ in connection with other decisions. With regard to this, Khan and Dhar [28] have shown that subjects are significantly more likely to choose a vice good if they have earlier engaged in a virtuous behaviour in a separate domain. For example, they show that subjects who are asked to help a foreign student to better understand a lecture subsequently donate less to charity. Related to this observation, Sachdeva et al. [43] argue that people seem to have a particular self-concept of their moral self-worth, which implies that people do not always behave in the same way but tend rather to use their self-concept as a reference point around which they can move. Hence if they perceive themselves as having acted morally, they feel licensed to act immorally on a subsequent occasion and vice versa. These results clearly indicate that people are aware of their underlying, often competing motivations and find

[7] With respect to these results, Milkman et al. [35] reflect on the possibility of "empowering" the *should* self and mention that their results give indications as to what people believe is better for them, rather than, as libertarian paternalism promotes, propose policies that facilitate the selection of options policy makers think are welfare-promoting (p. 336).

different ways to accommodate both or all of them over the stream of their actions. As Khan and Dhar [29] therefore point out, it would be interesting to study when people connect their current choice with future (or other) choices and when they do not do so. For example, "while deciding whether to attend a party or to prepare for an exam, people are often aware of another upcoming party next week. Similarly, while deciding what to have for lunch, people are aware of having to make the same decision later at dinner." (p. 287). In fact, choices are often seen as connected if they serve a particular goal. In particular, in the case where a choice involves a trade-off between two goals Dhar and Simonson [13] have shown that people prefer "balancing" the two goals, rather than "highlighting" one of them. For example, if the two goals are pleasure and good health, then Mr. A's dessert choice after dinner at a nice restaurant will be dependent on his previous main course choices. If he had a "tasty but unhealthy New York steak" he would rather opt for the "low-fat seasonal fruit salad", while if he had a "healthy but not so tasty low-fat pasta dish", he would rather choose a "great tasting but high-fat chocolate cake" (p. 32). He would not choose the chocolate cake after the steak, as "a neglect of one goal spoils the value of a peak experience on the other goal, for example, by creating guilt feelings" [13, p. 41]. Hence, the idea of *balancing* could be seen as inconsistent behaviour as the person may be unable to decide to which of the two goals she gives more weight. In fact, Dhar and Simonson [13] consider such behaviour as a form of self-control tactic, because by balancing one does not give in to any particular (possibly harmful) goal.[8]

2.2 *Conflict and Individual Well-Being*

In the above examples of intrapersonal conflict, researchers conducted experiments to test an underlying theory or general pattern of behaviour. However, there are also studies in psychology that have attempted to understand better particular kinds of conflict and their consequences on an individual's well-being. In each of these cases, intrapersonal conflict is generally understood as "[...] a situation in which one goal striving is seen by an individual as interfering with the achievement of other strivings in the individual's striving system." [16, p. 1041].

One intrapersonal conflict that has received considerable attention is the conflict between education or schooling and leisure (e.g. [17, 24, 30, 41]). Most of these authors acknowledge that young people, especially during their years at school or at university, have more than one goal. Many students are involved in extracurricular

[8] At least since Daniel Kahneman's book "Thinking Fast and Slow" [26], a particular kind of conflict, namely the one between, as Kahneman describes it, *System 1* and the *System 2* has become more well-known among economists. However, these are conflicts that have a cognitive origin most of the time and do not therefore correspond perfectly to the kind of psychological conflicts that we seek to consider here.

activities for various reasons (e.g. making friends, becoming more athletic, contributing to the school newspaper or radio station, etc.) that interfere with their academic work. In quite a number of cases, this "activity overload" impinges on their academic success, which may contribute to an overall decline in motivation to study, to concentrate and to be willing to continue attending school. Ratelle et al. [41], following Ryan and Deci's [14] "self-determination theory", distinguish between two kinds of motivation: one is "self-determined motivations", in which a person engages in an activity for its own sake or for the pleasure and satisfaction that she receives from such activity. Non-self-determined motivations on the other hand imply that a person engages in an activity for controlled reasons. That is, she does so to attain a reward or to avoid a punishment. Ratelle et al. [41] find that the interplay of two different conflicting motivations can be negative when motivations are non-self-determined. A school-leisure conflict, for example, predicts poor concentration, academic hopelessness and little intention to persist at school. These effects could also have negative consequences on psychological health. They therefore stress the need of students feeling pleasure and importance in pursuing school activities, because this may act as a protective factor against conflict with leisure.

Hofer [24] provides similar results. He uses the term *motivational conflict* when pupils strive for mutually exclusive goals at the same time, such as achievement goals but also a number of social and age-specific goals (connected to their body development, family, identity), and notes that in such cases engagement in school may decline and academic achievement is at risk. He refers to "goal switches to off-task behaviour" (p. 30) when pupils start doing something other than concentrating on their school activity while in class, e.g. day-dreaming, becoming angry or experiencing other negative feelings. He concludes that "[d]iscipline problems are not a failure in pupils' behaviour, rather they are a failure in the coordination of multiple goals" (p. 34). Different goals therefore need to be coordinated, a process Hofer calls "goal synthesis". One way of doing this is to put goals on a time line and to create a form of habitual behaviour, which has some self-regulatory benefits because each goal is allotted fixed time slots. Hofer also suggests the realignment of goals if inextricable goal conflicts continue to exist. In this case, one should look for new goals to replace inappropriate goals. In some other cases it would also mean downgrading specific goals to facilitate personal adjustment. An experiment run by Kilian et al. [30] on motivational conflict between learning and another enjoyable activity shows that if studying is associated with pleasure students will be much less distracted from following this goal and will in fact value the experience, i.e. there are ways to avoid a negative experience arising from competing motivations.

Another area of conflict that has been studied is the work/family context. In an overview article, Greenhaus and Beutell [20] describe this intrapersonal conflict in terms of "interrole" conflict. "Interrole conflict is experienced when pressures arising in one role are incompatible with pressures arising in another role." (p. 77). Obviously, multiple roles compete for a person's time and it has been shown that work/family conflict is positively related to the number of hours worked per week. These conflicts become even more important when, as above, they "motivationally

interfere" [17] with each other, in the sense that while one is physically attempting to meet the demands of one role, she is preoccupied with the pressures of another.

Other sources of conflict arise for example when behavioural styles that males (still quite often) exhibit at work (such as impersonality, logic, power, or authority) are not suited to the behaviour desired by their children. It has therefore been suggested that male managers may feel caught between two incompatible behaviour or value systems (see for references [20]). Scheduling conflicts arise when people do not manage to go to particular scheduled events, such as a concert, play, movie or a party). Pleck et al. [39] report that this is particularly a problem for women. Holahan and Gilbert [25] look at interrole conflict for working women who hold bachelor's degrees and are married with children. They hypothesise that women who perceive their employment as a career may experience greater interrole conflict than those who view it as just a job (even if they have the same level of education). However, they find exactly the contrary, i.e. greater involvement and personal investment in pursuit of a career does not seem to cause greater interrole conflict. In fact, the career group also stated that they received significantly more life satisfaction both from work and with respect to their own self-esteem, whereas the job group reported much less satisfaction from their work and family roles. Pleck et al. [39] report that being a parent increases the incidence of moderate or severe conflict by some 7 % points among husbands in two-earner families, but by twice as much among breadwinning husbands. This is about the same as the increase reported in conflict among the wives of employed husbands. Staines and O'Connor [49] report that parents of children under six experience greater work/family conflict than parents of school-age children, who again report greater conflict than childless couples. Clearly, interrole conflict and for that matter intrapersonal conflict is a cause of particular psychological strain and thus affects personal well-being. It causes emotional stress and lowers people's life satisfaction. As Emmons and King [16] report, conflict can even cause psychosomatic illnesses. In fact, in one of their studies they find a positive association between conflict and health centre visits.

2.3 Conflict and the Self

The above sections summarise, though certainly not exhaustively, a number of findings in psychology with respect to intrapersonal conflict and its consequences in terms of behaviour and people's psychological well-being. What seems to emerge clearly is that intrapersonal conflict is a pervasive phenomenon that interferes in many different, important life contexts and situations in which people are unable to attribute a clear priority ranking to their different concerns, motivations, goals or strivings, which would supposedly help to solve the conflict once and for all. Some studies have also indicated ways to alleviate the conflict experience in individuals, which generally entail an improvement in well-being.

"It has long been believed that reconciliation of opposing tendencies is a premier goal of human development" state Emmons and King [16, p. 1046],

summarising a great deal of research in psychology. In fact, research in child development has found that a child has to go through different developmental stages in which each stage is a more adequate way of understanding moral problems and resolving the conflicts encountered. This means that increased conflict is a condition for development [52, 53]. Higgins [23] reminds us that in the course of their development children learn at various age-stages to deal with egocentric and non-egocentric thought and to acquire a perspective-taking ability. That is, they come to understand that other people have different reactions to their behaviour and that they themselves prefer certain reactions to others; they then learn to adapt their behaviour accordingly. Over time, however, they learn to construct their "own standpoint", which may be distinct from the standpoint of significant "others" and these standpoints may come into conflict with one another because they learn to be more than just a "good boy" or a "good girl". In fact, Higgins is known for having developed the "self-discrepancy theory" [22], in which he postulates that people may experience conflicts between their "actual self", their "ideal self" and their "ought self". These discrepancies cause discomfort, and in particular he shows that a conflict between actual and ideal self causes depression, whereas a conflict between actual and ought self may lead to anxiety. Brim and Kagan [10] argue that throughout their lives people undergo change and that there are two fundamental dramatic conflicts inherent in the process of that change. "The first is the conflict between the person's wish to change while maintaining a sense of identity. The second is the conflict between the person and society; the person may wish to change, yet society may demand constancy, or the person may wish to remain the same, yet society may demand that the person change." (p. 17). That is, while society first transforms "the raw material of individual biology into persons suitable for the activities and requirements of society" [10, p. 19], people may then also start resisting societal demands and rebel against them. On the other hand, people may notice a difference between their actual and ought selves, to use Higgins' terms, and wish to conform rather than rebel. Such changes, according to Brim and Kagan, are usually supported by society. Conflicts of this kind are experienced throughout people's lifetimes as they move through a variety of positions in society.

Psychologists study people's self-concept and have long come to agree on the fact that there is no such thing as *one* single self-concept, but rather a multidimensional, multifacted dynamic structure [33]. Self-regulation, i.e. how a person controls her own behaviour, is therefore an important aspect of people's lives. Carver and Scheier [11, 12] in particular claim for example that people tend to compare their current state with a particular standard of behaviour. If they notice a discrepancy between the two they will attempt to reduce it. To use Brim and Kagan's [10] words, "[...] each person is, by nature, a purposeful, striving organism with a desire to be more than he or she is now." (p. 18). While to want to be more than one currently is causes a person to experience conflict and discrepancy, self-regulatory processes can help to achieve desired goals. Carver and Scheier [12] find that if people manage to make steady progress toward reducing this discrepancy they experience positive feelings and confidence. If they do not make any progress or progress only very slowly they experience doubt and negative effects. Hence, contrary to Higgins' view that all

discrepancies cause some form of negative experience, Carver and Scheier argue that what matters is the rate of progress in reducing the discrepancy and moving towards one's ideal. Thus actions in this context are not only choices that bring satisfaction to the individual (as economists may see them): in Carver and Scheier's terms they also imply changes between states (p. 22). Hence, action implies change but change may involve conflict, and vice versa.

3 A Theory of Choice with Conflicting Motivations

As the previous section indicates, motivational conflict is considered to be a widespread phenomenon in psychology and has been associated with inconsistent behaviour and "preference reversals". When, say, parents experience conflict between work and family and students between their academic and other social goals, such as making friends or, for instance, being politically active, then, in our opinion, this means more than their merely not being able to do all that they want within 24 h and thus being faced with a time constraint. In fact, if it were simply a time constraint then they could rank their alternatives and give more weight to those options that they like better or think are more important (e.g. work over family), maximise their utility and choose the time-distribution that best fits their own preferences or "tastes". In that case they would not experience any conflict. However, when parents say that they suffer from interrole and thus intrapsychic conflict, they may feel competing demands from the different life-spheres (family versus work) and even though they may like to work and like to be with their families, they have difficulties in deciding how much weight to give to each of these demands or how best to live with the pressures and concerns from competing domains. In economic terms, this means that they are unable to compare those "likings" and thus unable to form an *all-things-considered* preference ordering. In fact, as explained in the introduction, when talking about preferences, it is either assumed that people have already decided their *all-things-considered* preferences (which means that they have already been able to solve any potential conflict), or that preferences are simply a description of their consistent choices, and the reasons for those choices are not necessarily known to economists. As we have just observed, the assumption in the former case does not always make sense because, as psychological research shows, people do experience conflicts and are thus unable to determine their preferences. But with regard to the latter assumption there is also research, as summarised above, which highlights that inconsistent behaviour is associated with the experience of conflict and consequently no preferences can be revealed from such behaviour. In such cases, people may try for example to "balance" their different goals (e.g. students who have been partying often in one week are suddenly seen studying hard for a few days, only to go out more often afterwards once again, etc.). These kinds of conflict can also easily be seen as conflicts between what the person *wants* to do (e.g. spend more time with her children), and what she *should* do (e.g. work during weekends), and her behaviour

may differ, as pointed out above, according to the number of options available to her. This, however, is not the only way to describe possible conflicts. They may also arise as a consequence of the discrepancy between what a person would actually like to do (e.g. work full-time) and what other significant people she cares about expect her to do (e.g. work part-time and spend more time with the kids). In any case, competing demands or competing motivations may make it impossible to establish a unique *all-things-considered* preference ordering and as a consequence the person may "try by doing", that is, she may attempt several possible ways of reconciling her motivations and reasons for doing certain things in order to solve or alleviate the intrapsychic conflict. In economics, this would be considered as irrational behaviour (because it is inconsistent), but clearly it is not.

Following Baigent [3] we thus ask how such individuals can be characterised in an economic context. What follows is a short description of our theory of choice under conflicting motivations [1] in which we show that inconsistent behaviour is in fact associated with an underlying conflict between competing motivations. To simplify matters, we assume that people may experience conflict between two motivations, for example, in line with the spirit of some of the literature described above, between what a person *wants to do* and what that person thinks she *should be doing* (e.g. according to the goals that she has set for herself, or what parents or other significant people wish her to do, etc.). We refrain from calling these two motivations *want self* and *should self* as Bazerman et al. [6] do, because we do not want to attribute any "human-like features" to them such as the idea that the *want self* is more emotional, hot-headed or impulsive than the *should self*. The motivations are considered to be of equal status: we do not take any moral stand or assume that one is superior to the other. However, we do assume that motivations are more elementary, or more basic, than the standard idea of preferences. If the economic concept of preferences is taken seriously, then it is much more complex and more structured than motivations because they can be revealed from consistent choices. Motivations, as we understand them following psychological literature can be described as particular drives and forces that push an individual to do certain actions. Motivations could be visceral factors, as proposed by Loewenstein [31], but in the current context we think of them not as something that is triggered, say, by the smell of fresh cake, but as somewhat more permanent (e.g. following a career plan, being a good parent, eating healthy, being fit, etc.). That is why we consider it plausible to represent them in terms of single-peaked ordinal orderings of actions over a single dimension.[9] There is one or more particular actions which the individual is most motivated to choose, and any action further away from that peak will be less *wanted* or will satisfy less what the person *should* be doing. The set of actions along the dimension, normalised between 0 and 1, will depend on the problem at stake. For example, 0 could represent a student who enjoys her status

[9] As mentioned above, Dhar and Simonson [13] talk of "peak experiences" of people's goals. It does not therefore seem strange to think of particular experiences and motivations as single-peaked orderings.

as such, but more for the freedom and the social activities that she has available to her than for the studying in itself, while 1 is the opposite extreme. In between lies her *want*-peak ($\hat{W}$) at, say, 0.4, which indicates that she would enjoy a fair amount of studying, but still a substantial amount of time doing other activities, while her *should*-peak ($\hat{S}$) lies at 0.8, which would mean that she believes she should mainly study, and take only some limited time off to engage in extra-curricular activities. What we assume is that the individual is faced repeatedly with the same choice problem: for example the student has to make up her mind every day how much time to spend studying and how much on going out. The working mother has to decide every day whether to try to finish work early and go to the park with her kids, or to work overtime and make more progress with her workload. The person who wishes to be fit and healthy has to decide on each occasion whether to choose the tasty but very sweet dessert or the low-calory but less tasty cake in her favourite restaurant.

We assume that those two peaks do not overlap, which is a precondition for the experience of conflict. Whether a person experiences a conflict or not will depend on her *status quo* (SQ), which we define as the action currently chosen. Depending on the action currently chosen, which could lie either to the left or the right of either peak or between the two, the person may be confronted with different *types* of actions. We call these *A*-type actions if they satisfy both motivations more, *B*-type actions if they satisfy the *want* more and the *should*-motivation less, *C*-type actions if they satisfy the *should* less and the *want*-motivation more, and finally *D*-type actions, if they satisfy both the *want* and the *should*-motivation less than the action currently chosen, i.e. the SQ. We say that a person is confronted with a conflictual choice if she is faced with a choice between actions that satisfy one of the motivations but not both, that is with *B* or *C*-type actions. Figure 1 represents this characterisation of the individual.

The fundamental decision problem in such a situation is that the person is unable to compare each of the two motivations with the other. She is unable to establish how

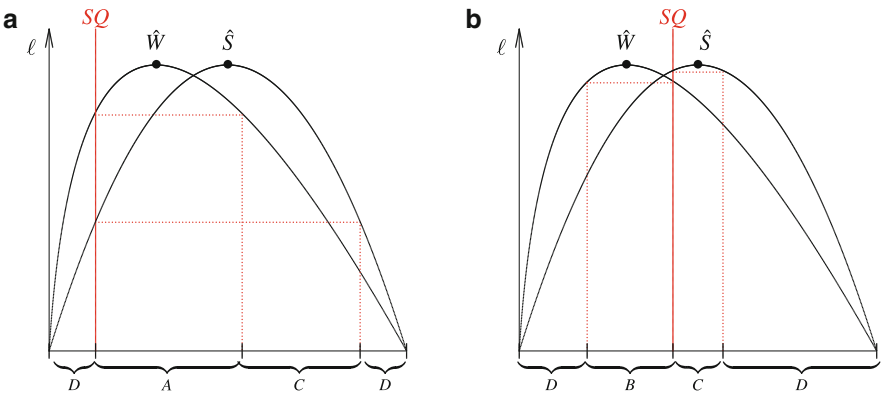


Fig. 1 Different types of actions (*curly brackets* indicate the respective range of actions). (**a**) Status quo to the left of $\hat{W}$. (**b**) Status quo in between the two peaks

to prioritise or weight either motivation and thus to choose an action based on an *all-things-considered* preference.[10] The question then is what reasonable conditions can be imposed on an individual's behaviour. We give just one example here, but discuss more conditions in Arlegi and Teschl [1]. It seems reasonable to assume that in such a situation the person would not choose a dominated action. We thus propose a condition called *Dominance* (DOM), according to which if the set of options includes two options such that one provides a lower level of fulfilment of both motivations than the other, i.e. that one option is *dominated* by the other, then the dominated option will never be chosen. DOM has as the consequence that (if it is assumed for instance that all actions over the course of the dimension are available) the individual restricts her choice to those actions that lie between the two peaks. Incomparability between the two motivations, however, makes any further condition difficult to justify. Consequently, once the person has chosen an action between the two peaks, which then becomes the SQ, she will only be faced with B and C-type actions, thus with conflictual actions that satisfy one but not both motivations with respect to the SQ. It is this circumstance that may lead people to act inconsistently.

For example, a well-known consistency condition that ensures that a preference relation can be found that rationalises a choice function is *Independence of Irrelevant Alternatives* (IIA) (see, for example, [27] or [9]). In words, IIA imposes that if an option x is chosen when y is feasible, then in another situation where the set of options is the same or more restricted but both options are still available, y will never be chosen. In our context, if we assume that DOM holds then we obtain that if a pair of actions (x, y) leads to an *IIA violation* (first x was chosen against y, but later y was chosen even though x was also available), then y can only be either a B or C-type action. Basically, the reason is that if DOM holds then x is an action between the two peaks and thus the new SQ. It is not difficult to check that if the SQ is between the two peaks then there are no A or D-type actions between the two peaks. Therefore any new action y that is taken will be either a B or a C-type action.[11] What we therefore show is that inconsistent behaviour is necessarily associated with conflict.

The next point to consider is that there is no reason to assume that motivations will not change over time. As seen in the previous section, psychologists think that motivations may change not only with the passage of time but also with the physical presence of particular objects (such as the smell of a cake). Of course, motivations may also change when people learn that they are in fact less or more important to them than they first thought. It has been suggested that goals (or motivations in our case) need to be realigned if they continue to cause irresolvable conflicts, and

[10]Pattanaik and Xu [38], inspired by Hare [21], propose a general model of multi-attribute choice where the different attributes are prioritised in one or another way depending on the occurrence of certain contextual characteristics of the decision problem. In our theory we do not presuppose the existence of such exogenous information.

[11]The formal proof, which can be found in Arlegi and Teschl [1] is a little more sophisticated and distinguishes between several particular cases.

in certain cases it has even been observed that goals are "downgraded" to reduce conflict. In fact, managing one's goals and motivations is a big part of research on self-regulation strategies.

Following the dictum of Carver and Scheier [12] that an "action implies change between states" (p. 22), we assume that motivations change with the actions chosen. We propose two different kinds of motivation change in the form of axioms, namely *reinforcement* (RF) and *dissonance reduction* (DR). Other motivation changes could be imagined, but for the moment we limit our analysis to these two. We do not necessarily assume that the person is aware of these motivation changes, i.e. for the moment we assume a rather myopic individual who does not have the knowledge of her motivation change required in order to, say, strategically choose actions to modify her motivations. We do however consider a more forward-looking person who may be aware of her motivation changes in Arlegi and Teschl [2]. The *reinforcement* axiom means that the individual will come to like or to want the chosen action more. Graphically, this is represented as the peak of the *want*-motivation, $\hat{W}$, moving towards the action chosen x to become $\hat{W}'$. Obviously, if the chosen action is the option that the person is most motivated to choose, the *want*-motivation does not change. The *dissonance reduction* axiom means that if the person chooses an action that lowers the fulfilment of what she *should* be doing, she experiences "dissonance", that is an unpleasant feeling that she would like to alleviate or to get rid of. This triggers a change in the *should*-motivation, in the sense that what the person *should* be doing is made more consistent with the action chosen. That is, the person accommodates what she *should* be doing with what she *wants* to do in order to restore some "consonance". Graphically, this means that the peak of the *should*-motivation, $\hat{S}$ moves towards the action chosen x and becomes $\hat{S}'$. Figure 2 represents the effects of the two axioms.

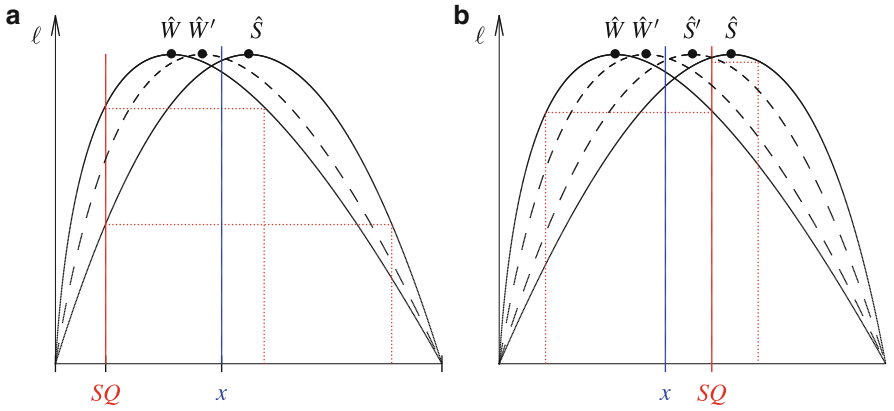


Fig. 2 Two psychological axioms. (**a**) Reinforcement: $\hat{W}$ moves towards x. (**b**) Dissonance reduction: $\hat{W}$ and $\hat{S}$ move towards x

If DOM, RF and DR are imposed, then it turns out that any sequence of different actions in which the individual engages will cause the peaks of the two motivations to move towards each other. As above, if DOM holds then the first action chosen will necessarily become the new SQ between the two peaks. From then on the individual will only be left with B or C-type actions. If the person chooses C-type actions, the $\hat{W}$ peak will continuously move towards the $\hat{S}$ peak, which does not move, whereas if she chooses a sequence of B-type actions, or a sequence of B and C-type actions, both peaks move and the distance between them is reduced. Consequently, engaging in a sequence of conflictual actions, DOM, together with RF and DR will reduce the set of undominated actions until eventually only one option may be left. In this case, the peaks converge and the individual has fully solved her conflict and may from then onwards "reveal" a preference in the standard economic sense. Hence, contrary to standard economic assumptions where changing preferences imply inconsistency, changing motivations here may actually lead to consistent behaviour. However, nothing in our analysis suggests that this needs to be case. In fact, in order to solve her conflict the person needs to engage in a series of conflictual choices, which may of course affect her psychological state and personal well-being. It may not always be easy to reduce the fulfilment of one motivation, even to gain satisfaction in another. It is therefore imaginable that the person might, for example, consistently choose a given SQ, which would increase her liking of this option because of RF but may not fully solve the conflict, i.e. the peaks would not fully converge and there may therefore always be a possibility of the person changing her behaviour as long as other undominated actions are available.

4 Conclusion

The role of motivations in human behaviour and the importance of conflict between different motivations is a well-known, well-reported issue among psychologists, but it has received only limited attention from economists. In our opinion, a careful formal analysis of the meaning of conflict between motivations and the effect of that conflict on an individual's behaviour and well-being constitutes a genuine exercise of what Baigent would understand as lifting the *veil of preferences*.

We extensively report psychological theories and experimental evidence of the fact that motivational conflict influences the consistency and well-being of decision makers, and of the importance of endogenous change in motivations. We then present the main ideas of our theory, which takes these aspects into account. We show that under a particular but not implausible way of representing conflict between motivations, there is a close connection between intrapersonal conflict and inconsistency in choice. Moreover, we show that when a more dynamic perspective of the problem is taken and endogenous changes in motivations are considered the interesting conclusion is reached that motivation change helps to reduce the possibility of inconsistencies. Finally, an interesting lesson that we have learnt from the theory that we propose is that conflict is a crucial aspect to be considered when

making judgements of well-being, and that this is an important, unexplored field that merits further research.

Acknowledgements We would like to thank two anonymous referees for very helpful comments on an earlier version of this paper.

References

1. Arlegi R, Teschl M (2012) A theory of choice under internal conflict. Working papers of the Department of Economics DT 1208, Public University of Navarre
2. Arlegi R, Teschl M (2014) Conflict, commitment and well-being. In: Søraker J, van der Rijt J-W, de Boer J, Wong P-H (eds) Well-being in contemporary society, Springer, Cham
3. Baigent N (1995) Behind the veil of preferences. Jpn Econ Rev 46(1):88–101
4. Baigent N, Gaernter W (1996) Never choose the uniquely largest: a characterization. Econ Theory 8(2):239–249
5. Bazerman M, Schroth H, Shah PP, Diekmann K, Tenbrunsel A (1994) The inconsistent role of comparison others and procedural justice in reactions to hypothetical job descriptions: implications for job acceptance decisions. Organ Behav Hum Dec Process 60(3):326–352
6. Bazerman M, Tenbrunsel A, Wade-Benzoni K (1998) Negotiating with yourself and losing: making decisions with competing internal preferences. Acad Manage Rev 23(2):225–241
7. Bazerman M, Moore D, Tenbrunsel A, Wade-Benzoni K, Blount S (1999) Explaining how preferences change across joint versus separate evaluation. J Econ Behav Organ 39:41–58
8. Bénabou R, Tirole J (2002) Self-confidence and personal motivation. Q J Econ 117(3):871–915
9. Binmore K (2009) Rational decisions. Princeton University Press, Princeton
10. Brim O, Kagan J (eds) (1980) Constancy and change in human development. Harvard University Press, Cambridge
11. Carver C, Scheier M (1981) Attention and self-regulation: a control theory approach to human behavior. Springer, New York
12. Carver C, Scheier M (1990) Origins and functions of positive and negative affect: a control-process view. Psychol Rev 97(1):19–35
13. Dhar R, Simonson I (1999) Making complementary choices in consumption episodes: highlighting versus balancing. J Mark Res 36:29–44
14. Deci E, Ryan R (2000) The what and why of goal pursuits: human needs and the self-determination of behavior. Psychol Inq 11:227–268
15. Elster J (1987) The multiple self. Cambridge University Press, Cambridge
16. Emmons R, King L (1988) Conflict among personal strivings: immediate and long-term implications for psychological and physical well-being. J Pers Soc Psychol 54(6):1040–1048
17. Fries S, Dietz F, Schmid S (2008) Motivational interference in learning: the impact of leisure alternatives on subsequent self-regulation. Contemp Educ Psychol 33:119–133
18. Fudenberg D, Levine D (2006) A dual self model of impulse control. Am Econ Rev 96(5):1449–1476
19. Gaertner W, Xu Y (1999) Rationality and external reference. Ration Soc 11(2):169–185
20. Greenhaus J, Beutell N (1985) Sources of conflict between work and family roles. Acad Manag Rev 10(1):76–88
21. Hare C (2007) Rationality and the distant needy. Philos Public Aff 35(2):161–178
22. Higgins ET (1987) Self-discrepancy: a theory relating self and affect. Psychol Rev 94(3):319–340
23. Higgins ET (1996) The self-digest: self-knowledge serving self-regulatory functions. J Pers Soc Psychol 71(6):1062–1083
24. Hofer M (2007) Goal conflicts and self-regulation: a new look at pupils' off-task behaviour in the classroom. Educ Res Rev 2:28–38

25. Holahan C, Gilbert L (1979) Interrole conflict for working women: careers versus Jobs. J Appl Psychol 64(1):86–90
26. Kahneman D (2011) Thinking fast and slow. Farrar, Straus and Giroux, New York
27. Kalai G, Rubinstein A, Spiegler R (2002) Rationalizing choice functions by multiple rationales. Econometrica 70(6):2481–2488
28. Khan U, Dhar R (2006) The licensing effect in consumer choice. J Mark Res 43(2):259–266
29. Khan U, Dhar R (2007) Where there is a way, is there a will? The effect of future choices on self-control. J Exp Psychol Gen 136(2):277–288
30. Kilian B, Hofer M, Kuhnle C (2012) The influence of motivational conflicts on personal values. J Educ Dev Psychol 2(1):57–68
31. Loewenstein G (1996) Out of control: visceral influences on behavior. Organ Behav Hum Decis Process 65(3):272–292
32. Loewenstein G, Prelec D (1991) Negative time preference. Am Econ Rev 81(2):347–352
33. Markus H, Wurf E (1987) The dynamic self-concept: a social psychological perspective. Annu Rev Psychol 38:299–337
34. Mas-Colell A, Whinston MD, Green JR (1995) Microeconomic theory. Oxford University Press, Oxford
35. Milkman K, Rogers T, Bazerman M (2008) Harnessing our inner angels and demons: what we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. Perspect Psychol Sci 3:324–338
36. Milkman K, Chugh D, Bazerman M (2009) How can decision making be improved. Perspect Psychol Sci 4:379–383
37. O'Connor K, De Dreu C, Schroth H, Barry B, Lituchy T, Bazerman M (2002) What we want to do versus what we think we should do: an empirical investigation of intrapersonal conflict. J Behav Decis Making 15:403–418
38. Pattanaik PK, Xu Y (2012) On dominance and context-dependence in decisions involving multiple attributes. Econ Philos 28:117–132
39. Pleck J, Staines G, Lang L (1980). Conflicts between work and family life. Mon Labor Rev 103:29–32
40. Prelec D, Herrnstein RJ (1991) Preferences or principles: alternative guidelines for choice. In: Zeckhauser R (ed) Strategy and choice. MIT Press, Cambridge, pp 321–340
41. Ratelle C, Vallerand R, Senècal C, Provencher P (2005) The relationship between school-leisure conflict and educational and mental health indexes: a motivational analysis. J Appl Soc Psychol 35(9):1800–1823
42. Rogers T, Bazerman M (2008) Future lock-in: future implementation increases selection of should choices. Organ Behav Hum Dec Process 106(1):1–20
43. Sachdeva S, Iliev R, Medin, D (2009) Sinning saints and saintly sinners. Psychol Sci 20(4):523–528
44. Samuelson PA (1953) Consumption theorems in terms of overcompensation rather than indifference comparisons. Economica 20(77):1–9
45. Schelling T (1960) The strategy of conflict. Harvard University Press, Cambridge
46. Schelling T (1984) Choice and consequence: perspectives of an errant economist. Harvard University Press, Cambridge
47. Schelling T (2006) Strategies of commitment. Harvard University Press, Cambridge
48. Sen A (1993) Internal consistency of choice. Econometrica 61(3):495–521
49. Staines G, O'Connor P (1980) Conflicts among work, leisure, and family roles. Mon Labor Rev 103(8):35–39
50. Strotz R (1955/56) Myopia and inconsistency in dynamic utility maximization. Rev Econ Stud 23(3):165–180
51. Thaler RH, Shefrin HM (1981) An economic theory of self control. J Polit Econ 89(2):392–406
52. Turiel E (1974) Conflict and transition in adolescent moral development. Child Dev 45(1): 14–29
53. Turiel E (1977) Conflict and transition in adolescent moral development, II: the resulution of disequilibrium through structural reorganiszation. Child Dev 48(2):634–637

A Note on Incompleteness, Transitivity and Suzumura Consistency

Richard Bradley

Abstract Rationality does not require of preferences that they be complete. Nor therefore that they be transitive: Suzumura consistency suffices. This paper examines the implications of these claims for the theory of rational choice. I propose a new choice rule—Strong Maximality—and argue that it better captures rational preference-based choice than other more familiar rules. Suzumura consistency of preferences is shown to be both necessary and sufficient for non-empty strongly maximal choice. Finally conditions on a choice function are stated that are necessary and sufficient for it to be rationalisable in terms of a Suzumura consistent preference relation.

Keywords Choice function • Incomplete preferences • Rationalisability • Suzumura consistency • Transitivity

1 Preference and Choice

This note concerns two questions about reason-based choice: What is required of the agent who makes her choices on the basis of her preferences? What can be inferred about an agent's preferences from the choices she makes? On both of these questions, I have learnt a great deal from Nick Baigent: from his writings, of course, but even more from discussions with him. If I could achieve even a small fraction of the clarity that he does when addressing these topics, I would be very happy indeed.

When thinking about the relation between preference and choice, it is worth distinguishing between the choices that are *permissible* given the agent's preferences, those that are *mandatory* and those that she *actually* makes. Rationality does not generally require that agents have strict preferences over all alternatives, so it is to be expected that these sets of choices will not coincide. For instance if she is indifferent between two alternatives or is unable to compare them it might be permissible for

R. Bradley (✉)
Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London, UK
e-mail: R.Bradley@lse.ac.uk

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*, Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_3

her to choose both of them, not mandatory to pick either, while in fact choosing only one of them.

There are two implications of this point. Firstly, preference-based explanations and/or rationalisations are necessarily limited in scope. Invoking someone's preferences will suffice to explain why some choices were not made (i.e. in terms of rational impermissibility) but not typically why some particular choice was made. To take up the slack, explanations must draw on factors other than preference: psychological ones such as the framing of the choice problem or the saliency of particular options, or sociological ones such as the existence of norms or conventions governing choices of the relevant kind. Some work has been done on how to rationalise choice when it has more than one determinant (see, for instance, Baigent [2]), but in general it is an insufficiently studied problem.

Secondly, observations of actual choices will only partially constrain preference attribution. For instance, that someone chooses a banana when an apple is available does not allow one to conclude that the choice of an apple was ruled out by her preferences, only that her preferences ruled the banana in. In this simple observation lies a serious obstacle to the ambition of Revealed Preference theory to give conditions on observed choices sufficient for the existence of a preference relation that rationalises them. For the usual practice of inferring the completeness of the agent's preferences from the fact that she always makes a choice when required to is clearly illegitimate if more than one choice is permitted by her preferences.

The upshot is that the usual focus on the case where an agent has complete preferences is quite unjustified. The aim of this note is therefore to explore the two opening questions without assuming completeness, building on the work of Sen [9], Richter [7] and especially the recent work of Bossert and Suzumura [4, 5]. I argue that when incompleteness of preference is reasonable then rationality does not require full transitivity of preferences. Instead it requires it that they be Suzumura consistent—roughly that there be no cycles of weak preference containing a strong preference. In a similar vein I argue for a choice rule—Strong Maximality—that is roughly intermediate between optimisation and maximisation and show that Suzumura consistency of preference is sufficient to ensure that this choice rule picks a non-empty set of alternatives from any given non-empty set of them. Finally, I investigate the rationalisability of choice functions in terms of Suzumura consistent preferences and strong maximal choice.

1.1 Preference

In the usual fashion we introduce a reflexive binary relation $\succeq$ (called the weak preference relation) on a set of alternatives X, with symmetric part $\approx$ (indifference) and anti-symmetric part $\succ$ (strict preference). In contrast to the way these terms are often used, we do not assume that in general any two alternatives are comparable under these preference relations. Instead we define a comparability relation $\bowtie$ on alternatives by: $\alpha \bowtie \beta$ iff $\alpha \succeq \beta$ or $\beta \succeq \alpha$. When all alternatives are comparable the preference relation is said to be complete. (Hence it is incomplete iff there is a pair of alternatives α and β such that $\alpha \not\bowtie \beta$.)

1.1.1 Transitivity

A number of different forms of transitivity-like properties of preference relations will be of interest. We say that $\succeq$ is:

1. *Transitive* iff for all $\alpha, \beta, \gamma \in X$, $\alpha \succeq \beta$ and $\beta \succeq \gamma$ implies that $\alpha \succeq \gamma$ (and intransitive otherwise)
2. *Incompletely transitive* iff for all $\alpha, \beta, \gamma \in X$, $\alpha \succeq \beta$ and $\beta \succeq \gamma$ implies that $\gamma \nsucc \alpha$
3. *PI-transitive* iff for all $\alpha, \beta, \gamma \in X$, $\alpha \succ \beta$ and $\beta \approx \gamma$ implies that $\alpha \succ \gamma$
4. *Quasi-transitive* iff $\succ$ is transitive

Transitivity implies incomplete transitivity, PI-transitivity and quasi-transitivity. On the other hand, a reflexive relation is transitive if it is either both complete and incompletely transitive or both PI-transitive and quasi-transitive [8, Theorem I.6]. But in general a relation can be incompletely transitive without being PI-transitive or quasi-transitive, and vice versa: they constitute alternative weakenings of transitivity.

The view taken here is that completeness is not a rationality requirement on preference. This is not in itself very controversial. Much more so is something that follows rather naturally from this view, namely that transitivity is too strong a requirement to impose on preferences. The problem is that transitivity imposes comparability even when it is not appropriate to do so. The following example serves to illustrate this point.

Suppose that Ann, Bob and Carol have interval scores in Maths and English as follows:

- (Ann) Maths: 80–90, English: 60–70
- (Bob) Maths: 56–65, English: 66–75
- (Carol) Maths: 75–85, English: 55–65

The teacher decides to rank them in each subject using the heuristic that two students with overlapping intervals scores in a subject should be regarded as on a par in that subject, but one is ranked higher than the other if the lower bound of their interval score is greater than the upper bound of the interval score of the other. So Ann ranks higher than Bob because she is definitely better at Maths and not comparably worse at English, Bob and Carol are ranked the same because each is better at one of the subjects and Ann and Carol are unranked relative to each other because neither is comparably better than the other in either subject.

The teacher's ranking of her students does not satisfy transitivity, but it is not obvious that her ranking is irrational given her inability to discriminate between Ann and Carol on the basis of their performances. It is not that the teacher should not infer that Ann is better than Carol, but rather that she is not rationally *compelled* to do so. This suggest that in situations in which a preference relation is not complete the requirements of rationality (with regard to preferences between pairs of a triple of alternatives) are more appropriately expressed by the condition of incomplete transitivity, than by that of full transitivity.

1.1.2 Consistency

In addition to the basic conditions on preference listed above, which are defined in terms of pairs or triples of alternatives, we are also interested in a number of derived consistency properties of the preference relation that can be defined in terms of these basic ones.

The weak preference relation $\succeq$ will be said to be:

1. *Strongly consistent* iff $\succeq$ is transitive
2. *Suzumura consistent* iff for all $\alpha_1, \alpha_2, \ldots, \alpha_n \in X, \alpha_1 \succeq \alpha_2, \alpha_2 \succeq \alpha_3, \ldots, \alpha_{n-1} \succeq \alpha_n$ implies that $\alpha_n \not\succ \alpha_1$
3. *Weakly consistent* iff $\succ$ is acyclic iff for all $\alpha_1, \alpha_2, \ldots, \alpha_n \in X, \alpha_1 \succ \alpha_2, \alpha_2 \succ \alpha_3, \ldots, \alpha_{n-1} \succ \alpha_n$ implies that $\alpha_n \not\succ \alpha_1$

These properties are in descending order of strength: strong consistency implies Suzumura consistency which implies weak consistency. Suzumura consistency strengthens incomplete transitivity, by extending it to arbitrary sets of alternatives.[1] As Bossert and Suzumura [4] point out there are three notable characteristics of Suzumura consistency. Firstly it rules out cycles with at least one strict preference and so preferences that satisfy it are not vulnerable to money pumps. Secondly, Suzumura consistency of a weak preference relation is necessary and sufficient for the existence of a complete and transitive extension of it.[2] And thirdly, any preference relation that is both Suzumura consistent and complete is strongly consistent. So there is good reason to think of Suzumura consistency as being the appropriate consistency condition for incomplete preferences.

2 Preference-Based Choice

Let $\mathcal{C}$ be a choice function on $\wp(X) - \emptyset$: a mapping from non-empty subsets $A \subseteq X$ to subsets $\mathcal{C}(A) \subseteq A$. Intuitively $\mathcal{C}(A)$ is the set of objects from the set A that *could* be chosen: could permissibly be so in normative interpretations, could factually be so in descriptive ones. If $\mathcal{C}(A)$ is always non-empty then it is said to be *decisive*. (Decisiveness is often built into the definition of a choice function, but it will prove more convenient here to make it a separate assumption.)

We are especially interested in the case when a choice function $\mathcal{C}$ can be said to be based on or determined by a preference relation. A natural condition for this being the case is that an object is chosen from a set only if no other object in the set is strictly preferred to it. Formally:

SPBC: (Strict Preference Based Choice)

$$\alpha \succ \beta \Rightarrow \forall(A : \alpha \in A), \beta \notin \mathcal{C}(A)$$

[1] The concept of Suzumura consistency was introduced in Suzumura [11].

[2] See Suzumura [11] for a proof.

SPBC certainly seems necessary for preference-based choice. But is it sufficient? I think not. A further requirement is that two alternatives that are regarded indifferently should always either both be chosen or both not chosen. Formally:

IBC: (Indifference Based Choice)

$$\alpha \approx \beta \Rightarrow \forall(A : \alpha, \beta \in A), \alpha \in \mathcal{C}(A) \Leftrightarrow \beta \in \mathcal{C}(A)$$

SPBC and IBC are not generally sufficient to determine choice because they don't settle the question of how to handle incomparability. Let us therefore consider three possible preference-based rules of choice that do fully determine what may be chosen and consider how they relate to these conditions. To do so it is useful to consider the transitive closure of $\approx$ on $A \subseteq X$, denoted $\hat{\approx}^A$ and defined by, for all $\alpha, \beta, \gamma \in A$: (1) $\alpha \hat{\approx}^A \alpha$, and (2) if $\alpha \hat{\approx}^A \beta$ and $\gamma \approx \beta$ then $\alpha \hat{\approx}^A \gamma$. Note that if $\alpha \hat{\approx}^A \beta$ then there exists a sequence of elements in X, $\alpha_1, \alpha_2, \ldots, \alpha_n$ linking α and β in the sense that $\alpha \approx \alpha_1$, $\alpha_1 \approx \alpha_2$, $\ldots$, and $\alpha_n \approx \beta$. It follows that $\hat{\approx}^A$ is transitive and symmetric and hence an equivalence relation on A. We call the set of $\beta \in A$ such that $\alpha \hat{\approx}^A \beta$, the indifference class of α in A.

The three rules of interest are the following:

Optimality: An object is chosen from a set if and only if it is weakly preferred to all others in the set. Formally, for all A such that $\alpha \in A$:

$$\alpha \in \mathcal{C}(A) \Leftrightarrow \forall(\beta \in A), \alpha \succeq \beta$$

Maximality: An object is chosen from a set if and only if no alternative in the set is strictly preferred to it. Formally, for all A such that $\alpha \in A$:

$$\alpha \in \mathcal{C}(A) \Leftrightarrow \neg\exists(\beta \in A : \beta \succ \alpha)$$

Strong Maximality: An object α is chosen from a set A iff there is no alternative in A strictly preferred to any alternative in α's indifference class in A. Formally, for all A such that $\alpha \in A$:

$$\alpha \in \mathcal{C}(A) \Leftrightarrow \neg\exists(\beta, \gamma \in A : \alpha \hat{\approx}^A \gamma \text{ and } \beta \succ \gamma)$$

Of these three rules, Optimality is the one that is most commonly taken to express rational preference-based choice (see, for instance, Arrow [1] and Sen [9]). But although Optimality satisfies both SPBC and IBC, it is clearly too strong a condition on *permissible* choice. This is because it implies that if $\alpha \not\bowtie \beta$ then $\mathcal{C}(\{\alpha, \beta\}) = \varnothing$. But even if there are situations in which no choice is permissible (contrary to the usual assumption of decisiveness), this is not a consequence of incomparability. If two alternatives are incomparable it should normally be permissible to choose either of them.

For this reason Maximality is often seen as the more appropriate rule of rational choice when the possibility of incomparability is not ruled out (see Sen [10]).

But Maximality is also not quite right, as the following schematic version of our earlier example shows. Suppose that $\alpha \succ \beta$ and $\beta \approx \gamma$ but $\alpha \not\bowtie \gamma$. Then it would not be unreasonable for $\mathcal{C}(\{\alpha, \gamma\}) = \{\alpha, \gamma\}$ because the two alternatives are incomparable and $\mathcal{C}(\{\beta, \gamma\}) = \{\beta, \gamma\}$ because the two alternatives are equally preferred, but $\mathcal{C}(\{\alpha, \beta, \gamma\}) = \{\alpha\}$ because β should not be chosen when a strictly preferred alternative—α—is available and γ should not be chosen if β is not, given that $\gamma \approx \beta$. But these choices are inconsistent with Maximality which requires that $\mathcal{C}(\{\alpha, \beta, \gamma\}) = \{\alpha, \gamma\}$.

The problem with Maximality is that it leads to violations of IBC. Since Maximality requires that $\mathcal{C}(\{\alpha, \beta, \gamma\}) = \{\alpha, \gamma\}$, it is not the case that β is chosen whenever γ is, even though $\beta \approx \gamma$. So just as admitting the possibility of incompleteness required a shift from Optimality to Maximality, so too recognition of the rational permissibility of incompletely transitive preferences requires a shift from Maximality to Strong Maximality.

Let us consider a reformulation of Strong Maximality that will make its implications clearer. For any $A \subseteq X$ let $\mathbf{A} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \ldots\}$ be the set of equivalence classes in A induced by the relation $\hat{\approx}^A$. Define a weak preference relation $\geq$ on $\mathbf{A}$ by $\forall \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbf{A}$:

$$\boldsymbol{\alpha} \geq \boldsymbol{\beta} \Leftrightarrow \exists(\alpha \in \boldsymbol{\alpha}, \beta \in \boldsymbol{\beta} : \alpha \succeq \beta)$$

Then choosing from any A in accordance with Strong Maximality is equivalent to choosing the $\geq$-maximal element of the set $\mathbf{A}$ of equivalence classes in A induced by the equivalence relation $\hat{\approx}^A$.

Now it might be objected that adopting Strong Maximality as a principle of rational choice is tantamount to smuggling transitivity of indifference back in via the equivalence classes under $\hat{\approx}^A$. But there is another way for formulating the rule which should serve to alleviate this worry. Let us define a sequence of choice functions $\langle \bar{\mathcal{C}}^\tau_\succeq(A) \rangle^\infty_{\tau=1}$ as follows[3]:

1. $\bar{\mathcal{C}}^0_\succeq(A) = \{\alpha \in A : \exists \beta \in A \text{ such that } \beta \succ \alpha\}$
2. $\bar{\mathcal{C}}^\tau_\succeq(A) = \{\alpha \in A : \exists \beta \in A \text{ such that } \beta \approx \alpha \text{ and } \beta \in \bar{\mathcal{C}}^{\tau-1}_\succeq(A)\}$

Then we define the set of impermissible alternatives by:

$$\bar{\mathcal{C}}_\succeq(A) = \bigcup_{\tau=0}^{\infty} \bar{\mathcal{C}}^\tau_\succeq(A)$$

Intuitively $\bar{\mathcal{C}}_\succeq(A)$ is the set of alternatives in A that must not be chosen. Then Strong Maximality is equivalent to the rule of choosing any alternative that is not impermissible, i.e. to the rule:

Non-Elimination: $\alpha \in \mathcal{C}(A) \Leftrightarrow \alpha \notin \bar{\mathcal{C}}(A)$

[3] I am grateful to an anonymous referee for suggesting this formulation.

To apply this rule it suffices that the agent iteratively eliminates alternatives from her choice set by removing any dominated alternatives; then checking if any alternatives that are left are indifferent to any eliminated ones and, if so, removing them as well; then checking if any alternatives that are left are indifferent to any eliminated ones, and so on.

2.1 Properties of Preference-Based Choice

Each of the three choice rules under examination expresses a view on the relationship between preference and choice. To examine what these are and how they differ for the three choice rules, let us denote the choice function determined by the weak preference relation $\succeq$ together with Maximality, Optimality or Strong Maximality by $\mathcal{C}_{\succeq}^{Max}$, $\mathcal{C}_{\succeq}^{Op}$ and $\mathcal{C}_{\succeq}^{SM}$ respectively, where these are defined as follows. For any $A \subseteq X$:

$$\mathcal{C}_{\succeq}^{Op}(A) = \{\alpha \in A : \forall \beta \in A, \alpha \succeq \beta\}$$
$$\mathcal{C}_{\succeq}^{Max}(A) = \{\alpha \in A : \forall \beta \in A, \beta \not\succ \alpha\}$$
$$\mathcal{C}_{\succeq}^{SM}(A) = \{\alpha \in A : \forall (\gamma \in A : \gamma \hat{\approx}^{A} \alpha), \neg \exists (\beta \in A : \beta \succ \gamma)\}$$

For the rest of this section, I will drop the subscript on the choice function as the preference relation is fixed throughout the discussion.

The first thing to note is that the set of permissible choices according to $\mathcal{C}^{Max}$ is always at least as large as those determined by $\mathcal{C}^{Op}$ or $\mathcal{C}^{SM}$. Furthermore when the preference relation is Suzumura consistent then the set of choices that are permissible according to Strong Maximality contain those that are permissible according to Optimality (as well as being contained by those determined by Maximality). On the other hand when the preference relation is complete $\mathcal{C}^{Op}$ coincides with $\mathcal{C}^{Max}$ and when it is transitive, $\mathcal{C}^{SM}$ coincides with $\mathcal{C}^{Max}$. These relationships are summarised in Fig. 1, where arrows indicate implications given the indicated conditions, and proven below as Theorem 1.

It is well known that for finite sets of alternatives $\mathcal{C}^{Op}$ is decisive iff $\succeq$ is complete and weakly consistent and that $\mathcal{C}^{Max}$ is decisive iff $\succeq$ is weakly consistent. Theorem 2 below establishes a corresponding result for choices that are strongly maximal, namely that a choice function based on Strong Maximality is decisive iff the underlying preference relation is Suzumura consistent. The main significance of this result for our argument is that Suzumura consistency is thereby shown to be both necessary and sufficient for decisive, strongly maximal choice.

Theorem 1

1. $\mathcal{C}^{Op} \subseteq \mathcal{C}^{Max}$ *and* $\mathcal{C}^{SM} \subseteq \mathcal{C}^{Max}$
2. *If* $\succeq$ *is complete then* $\mathcal{C}^{Op} = \mathcal{C}^{Max}$

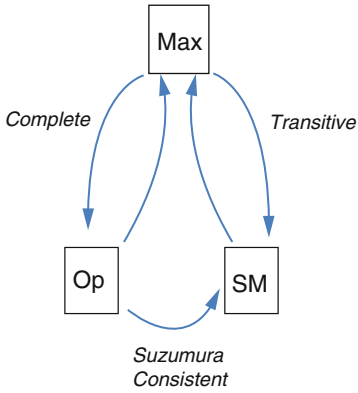


Fig. 1 Relations between choice rules

3. *If* $\succeq$ *is transitive, then* $\mathcal{C}^{SM} = \mathcal{C}^{Max}$
4. *If* $\succeq$ *is Suzumura consistent, then* $\mathcal{C}^{Op} \subseteq \mathcal{C}^{SM}$.
5. *If* $\succeq$ *is complete and Suzumura consistent then* $\mathcal{C}^{SM} = \mathcal{C}^{Op} = \mathcal{C}^{Max}$

Proof

(1) Suppose $\alpha \in \mathcal{C}^{Op}(A)$. Then $\forall \beta \in A, \alpha \succeq \beta$. But then $\forall \beta \in A, \beta \not\succ \alpha$. So $\alpha \in \mathcal{C}^{Max}(A)$. Similarly suppose $\alpha \in \mathcal{C}^{SM}(A)$. Now by the symmetry of indifference $\alpha \approx \alpha$, so it follows that $\neg \exists \beta \in A$ such that $\beta \succ \alpha$). So $\alpha \in \mathcal{C}^{Max}(A)$.

(2) Suppose that $\succeq$ is complete. Then for any $\beta \in A$ if $\beta \not\succ \alpha$ then $\alpha \succeq \beta$. Hence if $\alpha \in \mathcal{C}^{Op}(A)$ then $\alpha \in \mathcal{C}^{Max}(A)$. So $\mathcal{C}^{Op} = \mathcal{C}^{Max}$.

(3) Suppose that $\succeq$ is transitive but that there exists $\alpha \in A$ such that $\alpha \in \mathcal{C}^{Max}(A)$ but $\alpha \notin \mathcal{C}^{SM}(A)$. Now if $\alpha \notin \mathcal{C}^{SM}(A)$ then there exists $\beta, \gamma \in A$ such that $\alpha \hat{\approx}^A \gamma$ and $\beta \succ \gamma$. By transitivity, if $\alpha \hat{\approx}^A \gamma$ then $\alpha \approx \gamma$ and so by transitivity again, $\beta \succeq \alpha$. But if $\alpha \in \mathcal{C}^{Max}(A)$ then $\beta \not\succ \alpha$. So $\beta \approx \alpha$ and by transitivity, $\beta \approx \gamma$. Hence, contrary to assumption, $\beta \not\succ \gamma$. It follows that if $\alpha \in \mathcal{C}^{Max}(A)$ then $\alpha \in \mathcal{C}^{SM}(A)$ and hence that $\mathcal{C}^{SM} = \mathcal{C}^{Max}$.

(4) Suppose that $\succeq$ is Suzumura consistent and that $\alpha \in \mathcal{C}^{Op}(A)$. Then $\forall \beta \in A$, $\alpha \succeq \beta$. Let $\gamma \in A$ be such that $\alpha \hat{\approx}^A \gamma$. Then there exists a sequence of elements in A, $\alpha_1, \alpha_2, \ldots, \alpha_n$ linking γ, α and β in the sense that $\gamma \approx \alpha_1, \alpha_1 \approx \alpha_2, \ldots, \alpha_n \approx \alpha$ and $\alpha \succeq \beta$. Hence by Suzumura consistency $\beta \not\succ \gamma$. It follows that $\alpha \in \mathcal{C}^{SM}(A)$.

(5) Follows from 2, 3 and 4. ■

Theorem 2 *Suppose that the set of alternatives X is finite. Then:*

1. $\mathcal{C}^{Op}$ *is decisive iff* $\succeq$ *is complete and weakly consistent*
2. $\mathcal{C}^{Max}$ *is decisive iff* $\succeq$ *is weakly consistent*
3. $\mathcal{C}^{SM}$ *is decisive iff* $\succeq$ *is Suzumura consistent*

Proof (2) Suppose $\succeq$ is not weakly consistent. Then there exists $A = \{\alpha_1, \alpha_2, \ldots, \alpha_n\} \subseteq X$, such that $\alpha_0 \succeq \alpha_1, \alpha_1 \succeq \alpha_2, \ldots, \alpha_{n-1} \succeq \alpha_n$ and $\alpha_n \succ \alpha_0$. But then for $n \geq i \geq 1$, $\alpha_i \notin \mathcal{C}^{Max}(A)$ because $\alpha_{i-1} \succ \alpha_i$. And

$\alpha_0 \notin \mathcal{C}^{Max}(A)$ because $\alpha_n \succ \alpha_0$. So $\mathcal{C}^{Max}(A) = \varnothing$. Hence $\mathcal{C}^{Max}$ is not decisive. For the converse see Kreps [6].

(1) Suppose $\succeq$ is either not weakly consistent or incomplete. Suppose it is not weakly consistent. Then since by Theorem 1(1), $\mathcal{C}^{Op} \subseteq \mathcal{C}^{Max}$ it follows from 2. that $\mathcal{C}^{Op}(A) \neq \varnothing$. Now suppose that $\succeq$ is incomplete. Then there exists $\alpha, \beta \in X$ such that $\alpha \not\succ\!\!\prec \beta$ and hence such that $\mathcal{C}^{Op}(\{\alpha, \beta\}) = \varnothing$. So $\mathcal{C}^{Op}$ is not decisive. The converse follows from 2. and Theorem 1(2).

(3) Suppose that $\mathcal{C}^{SM}$ is decisive but that $\succeq$ is not Suzumura consistent, i.e. for some set $A = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ it is the case that $\alpha_1 \succeq \alpha_2, \alpha_2 \succeq \alpha_3, \ldots, \alpha_{n-1} \succeq \alpha_n$ but that $\alpha_n \succ \alpha_1$. We prove by induction on i that it then follows that for all $\alpha_i \in A$, $\alpha_i \notin \mathcal{C}^{SM}(A)$ and hence that $\mathcal{C}^{SM}(A) = \varnothing$. First Strong Maximality implies that $\alpha_1 \notin \mathcal{C}^{SM}(A)$ because $\alpha_n \succ \alpha_1$. Now assume that for some $k > 1$, $\alpha_k \notin \mathcal{C}^{SM}(A)$. Then there exists some α_j and $\alpha_{j'}$ such that $\alpha_k \hat{\approx}^A \alpha_j$ but $\alpha_{j'} \succ \alpha_j$. Now consider α_{k+1}. Either $\alpha_k \succ \alpha_{k+1}$ in which case it follows by Strong Maximality that $\alpha_{k+1} \notin \mathcal{C}^{SM}(A)$. Or $\alpha_k \approx \alpha_{k+1}$, in which case $\alpha_k \hat{\approx}^A \alpha_{k+1}$. But then $\alpha_{k+1} \hat{\approx}^A \alpha_j$ and so by Strong Maximality $\alpha_{k+1} \notin \mathcal{C}^{SM}(A)$. But this implies that $\mathcal{C}^{SM}(A) = \varnothing$ in contradiction to the assumption of decisiveness. So $\succeq$ must be Suzumura consistent.

For the other direction, suppose that $\succeq$ is Suzumura consistent, but that for some set A, $\mathcal{C}^{SM}(A) = \varnothing$. If $\mathcal{C}^{SM}(A) = \varnothing$ then $\mathcal{C}^{Max}_{\succeq}(\mathbf{A}) = \varnothing$ and so by (2), $>$ is cyclic i.e. there exist subsets of A such that $\mathbf{A}_1 > \mathbf{A}_2, \mathbf{A}_2 > \mathbf{A}_3 \ldots, \mathbf{A}_{n-1} > \mathbf{A}_n$, and $\mathbf{A}_n > \mathbf{A}_1$. So by definition there exists $\alpha_1, \alpha'_1, \alpha_2, \ldots, \alpha'_n, \alpha_n \in A$ such that $\alpha_1 \hat{\approx}^A \alpha'_1, \ldots, \alpha'_n \hat{\approx}^A \alpha_n$ and $\alpha'_1 \succ \alpha_2$, $\alpha'_2 \succ \alpha_3, \ldots, \alpha'_{n-1} \succ \alpha_n$, but $\alpha'_n \succ \alpha_1$. But by Suzumura consistency, if $\alpha_1 \hat{\approx}^A \alpha'_1, \alpha'_1 \succ \alpha_2$, $\alpha_2 \hat{\approx}^A \alpha'_2, \alpha'_2 \succ \alpha_3, \ldots, \alpha'_{n-1} \succ \alpha_n$ and $\alpha_n \hat{\approx}^A \alpha'_n$ then $\alpha'_n \not\succ \alpha_1$. So $\mathcal{C}^{Max}_{\succeq}(\mathbf{A}) \neq \varnothing$. Hence $\mathcal{C}^{SM}$ is decisive. ■

2.2 *Properties of Choice Functions*

What features of choice functions are induced by our choice rules? The following properties—Sen's alpha, beta and gamma conditions—have figured prominently in the existing literature. Let $\alpha \in A$ and $\gamma \in C$. Then:

Set Contraction: If $\alpha \in \mathcal{C}(B)$ and $A \subseteq B$ then $\alpha \in \mathcal{C}(A)$
Set Expansion: If $\alpha, \beta \in \mathcal{C}(B)$, $B \subseteq A$ and $\beta \in \mathcal{C}(A)$, then $\alpha \in \mathcal{C}(A)$
Set Union: If $\alpha \in \mathcal{C}(A)$ and $\alpha \in \mathcal{C}(B)$, then $\alpha \in \mathcal{C}(A \cup B)$

It is well known that Optimality-based choice will satisfy both Set Contraction and Set Expansion so long as the underlying weak preference relation is weakly consistent (see Sen [9]). In fact choice based on weakly consistent preferences will satisfy Set Contraction given any one of the three choice rules under examination. Set Expansion on the other hand need not be satisfied by maximal or strongly maximal choice. This is as it should be. Suppose, for example, that the agent cannot compare α and β, but that no alternative in A is preferred to either. So both are

permissible choices. Now suppose that $B = A \cup \{\gamma\}$ and that $\gamma \succ \alpha$ but $\gamma \not\rhd \beta$. Then β is a still permissible choice but not α. So Set Expansion is violated.

More interesting perhaps is that $\mathcal{C}^{SM}$, unlike the other two rules, does not satisfy Set Union, for Sen [9] has shown that satisfaction of this condition along with Set Contraction is essentially equivalent to the choice function being binary in composition. Instead it is both necessary and sufficient that strongly maximal choices be based on PI-transitive preferences for $\mathcal{C}^{SM}$ to satisfy Set Union.

If $\mathcal{C}^{SM}$ does not generally satisfy Set Union, what properties are characteristic of it? Two weaker principles than Set Union turn out to be significant. The first is a very weak consequence of Set Union.

Element Union: If $\forall \alpha \in A, \alpha \in \mathcal{C}(\{\alpha, \beta)\})$ then for some $\alpha^* \in A, \alpha^* \in \mathcal{C}(A \cup \{\beta\})$

To state the second condition, we need to introduce a choice theoretic analogue of the notion of the indifference class of an alternative α in some set A—called α's choice equivalence class. Intuitively the choice equivalence class of α is the set of elements that are chosen whenever α is, in any set containing both. To state it formally, we first define a sequence of functions $\langle \tilde{\mathcal{C}}^{\tau}(A, \alpha)\rangle_{\tau=1}^{\infty}$, induced by a given choice function $\mathcal{C}$, as follows:

1. $\tilde{\mathcal{C}}^{0}(A) = \{\alpha\}$
2. $\tilde{\mathcal{C}}^{\tau}(A) = \{\beta \in A$: For some $\gamma \in \tilde{\mathcal{C}}^{\tau-1}(A, \alpha)$, $\beta \in \mathcal{C}(B) \Leftrightarrow \gamma \in \mathcal{C}(B)$ for all $B \subseteq X$ such that $\beta, \gamma \in B\}$

Then we define α's *choice equivalence class* in A induced by $\mathcal{C}$, as follows:

$$\tilde{\mathcal{C}}(A, \alpha) = \bigcup_{\tau=0}^{\infty} \tilde{\mathcal{C}}^{\tau}(A)$$

Note that if $\beta \hat{\approx}^{A} \alpha$ then $\beta \in \tilde{C}^{SM}(A, \alpha)$. For if $\beta \hat{\approx}^{A} \alpha$, then there exists a sequence of elements in A, $\alpha_1, \alpha_2, \ldots, \alpha_n$ linking α and β in the sense that $\alpha = \alpha_1, \alpha_1 \approx \alpha_2, \ldots, \alpha_{n-1} \approx \alpha_n$, and $\alpha_n = \beta$. And for all α_i in this sequence, $\alpha_i \in \mathcal{C}^{SM}(B) \Leftrightarrow \alpha_{i+1} \in \mathcal{C}^{SM}(B)$ for all $B \subseteq X$ such that $\alpha_i, \alpha_{i+1} \in B$. Since $\alpha \in \tilde{C}^{SM}(A, \alpha)$, it follows that $\alpha_2 \in \tilde{C}^{SM}(A, \alpha)$ and hence $\alpha_3 \in \tilde{C}^{SM}(A, \alpha)$ and hence $\ldots \beta \in \tilde{C}^{SM}(A, \alpha)$.

Now we can state the final condition of interest:

Equivalence Class Union If $\tilde{\mathcal{C}}(A \cup B, \alpha) \subseteq \mathcal{C}(A)$ and $\tilde{\mathcal{C}}(A \cup B, \alpha) \subseteq \mathcal{C}(B)$, then $\tilde{\mathcal{C}}(A \cup B, \alpha) \subseteq \mathcal{C}(A \cup B)$

This condition, like Element Union, *is* satisfied by choice in accordance with Strong Maximality.

Theorem 3

1. *$C_{\succeq}^{Op}/\mathcal{C}^{Max}/\mathcal{C}^{SM}$ all satisfy Set Contraction*
2. *$\mathcal{C}^{\widetilde{Op}}$ and $\mathcal{C}^{Max}$ satisfy Set Union, but $\mathcal{C}^{SM}$ need not.*

3. If $\mathcal{C}^{SM}$ is decisive then $\mathcal{C}^{SM}$ satisfies Set Union iff $\succeq$ is PI-transitive
4. $\mathcal{C}^{SM}$ satisfies Element Union and Equivalence Class Union

Proof

(1) Suppose $B \subseteq A$ and $\alpha \in \mathcal{C}^{SM}(A)$. Then $\forall \beta, \gamma \in A$ such that $\alpha \hat{\approx}^{A} \gamma$, $\beta \not\succ \gamma$. Hence $\forall \beta, \gamma \in B$ such that $\alpha \hat{\approx}^{A} \gamma$, $\beta \not\succ \gamma$. So $\alpha \in \mathcal{C}^{SM}(B)$. Similarly for $\mathcal{C}^{Op}$ and $\mathcal{C}^{Max}$.

(2) Suppose that $\alpha \in \mathcal{C}^{Max}(A)$ and $\alpha \in \mathcal{C}^{Max}(B)$. Then $\forall \beta \in A$, $\beta \not\succ \alpha$ and $\forall \beta \in B$, $\beta \not\succ \alpha$. Hence $\forall \beta \in A \cup B$, $\beta \not\succ \alpha$. It follows that $\alpha \in \mathcal{C}^{Max}(A \cup B)$. Similarly for $\mathcal{C}^{Op}$. However consider a case in which $\alpha \not\bowtie \beta$, $\alpha \approx\!\!\!\!\approx \gamma$ and $\beta \succ \gamma$. Then $\mathcal{C}^{SM}(\{\alpha, \beta\}) = \{\alpha, \beta\}$, $\mathcal{C}^{SM}(\{\alpha, \gamma\}) = \{\alpha, \gamma\}$ but $\alpha \notin \mathcal{C}^{SM}(\{\alpha, \beta, \gamma\})$ because $\alpha \hat{\approx}^{A} \gamma$ and $\beta \succ \gamma$.

(3) Suppose $\succeq$ is PI-transitive, that $\alpha \in \mathcal{C}^{SM}(A)$ and that $\alpha \in \mathcal{C}^{SM}(B)$, but that $\alpha \notin \mathcal{C}^{SM}(A \cup B)$. Then there exists $\beta, \gamma \in A \cup B$ such that $\alpha \hat{\approx}^{A} \gamma$ and $\beta \succ \gamma$. But then by repeated applications of PI-transitivity it follows that $\beta \succ \alpha$. Hence $\alpha \notin \mathcal{C}^{SM}(A)$ or $\alpha \notin \mathcal{C}^{SM}(B)$, depending on whether $\beta \in A$ or $\beta \in B$. Now suppose that $\succeq$ is not PI-transitive. Then there exists $\alpha, \beta, \gamma \in X$ such that $\alpha \succ \beta$ and $\beta \approx \gamma$ but $\alpha \not\succ \gamma$. Then either $\gamma \succ \alpha$, $\alpha \approx \gamma$ or $\alpha \not\bowtie \gamma$. Suppose $\gamma \succ \alpha$ or $\alpha \approx \gamma$. Then $\mathcal{C}^{SM}(\{\alpha, \beta, \gamma\}) = \varnothing$, contrary to the assumption that $\mathcal{C}^{SM}$ is decisive. So suppose that $\alpha \not\bowtie \gamma$. Then $\gamma \in \mathcal{C}^{SM}(\{\alpha, \gamma\})$ and $\gamma \in \mathcal{C}^{SM}(\{\beta, \gamma\})$, but $\gamma \notin \mathcal{C}^{SM}(\{\alpha, \beta, \gamma\})$ because $\alpha \succ \beta$ and $\gamma \approx \beta$. So Set Union is violated.

(4) Suppose that $\forall \alpha \in A$, $\alpha \in \mathcal{C}(\{\alpha, \beta\})$. Then if $\mathcal{C}^{SM}(A \cup \{\beta\}) = \{\beta\}$, there must exist some $\alpha^* \in A$, such that $\beta \succ \alpha^*$ and hence, contrary to supposition, $\alpha^* \notin \mathcal{C}^{SM}(\{\alpha^*, \beta)\})$. So Element Union is satisfied. Now suppose that $\tilde{\mathcal{C}}^{SM}(A \cup B, \alpha) \subseteq \mathcal{C}^{SM}(A)$ and $\tilde{\mathcal{C}}^{SM}(A \cup B, \alpha) \subseteq \mathcal{C}^{SM}(B)$. Suppose that $\tilde{\mathcal{C}}^{SM}(A \cup B, \alpha) \not\subseteq \tilde{\mathcal{C}}^{SM}(A \cup B)$. Then in particular, $\alpha \notin \mathcal{C}^{SM}(A \cup B)$. Then there exists $\gamma, \delta \in A \cup B$ such that $\alpha \hat{\approx}^{A \cup B} \gamma$ and $\beta \succ \gamma$. But if $\alpha \hat{\approx}^{A \cup B} \gamma$ then $\gamma \in \tilde{C}^{SM}(A \cup B, \alpha)$. So $\gamma \in A \cap B$ since $\gamma \in \mathcal{C}^{SM}(A)$ and $\gamma \in \mathcal{C}^{SM}(B)$. But $\delta \in A$ or $\delta \in B$. So $\gamma \notin \mathcal{C}^{SM}(A)$ or $\gamma \notin \mathcal{C}^{SM}(B)$, in contradiction to what we have just established. It follows that $\tilde{\mathcal{C}}^{SM}(A \cup B, \alpha) \subseteq \mathcal{C}^{SM}(A \cup B)$. ■

3 Rationalisability

A question that naturally arises is whether, and under what conditions, the choices that are formally represented by a choice function can be rationalised or explained in terms of an underlying preference relation that, together with some choice rule, determines it. To tackle it, let us say that a choice function $\mathcal{C}$ is *rationalisable* by a consistent weak preference relation $\succeq$ iff $\mathcal{C}$ is generated by $\succeq$ together with a given choice rule R, i.e. iff $\mathcal{C} = \mathcal{C}^{R}_{\succeq}$. This definition of rationalisability contains two unspecified parameters: the type of consistency to be required of preference and the type of choice rule to be used in the determination of the choice function.

Different kinds of rationalisability will be associated with different values for these parameters and a particular choice function may be rationalisable relative to some combination of consistency property and choice rule but not another. Here we will require that the preference relation be at least weakly consistent in order to speak of rationalisation, and differentiate between O-, M- and SM-rationalisations of a choice function in accordance with the choice rule that determines it.

3.1 Revealed Preference

In the literature on revealed preference the question of rationalisability is typically approached by defining the weak preference relation $\dot{\succeq}_{\mathcal{C}}$ 'revealed' by a choice function $\mathcal{C}$ in the following way:

Revealed Preference: $\alpha \dot{\succeq}_{\mathcal{C}} \beta \Leftrightarrow \exists A \subseteq X$ such that $\alpha, \beta \in A$ and $\alpha \in \mathcal{C}(A)$

It is then possible to ask what properties of the revealed preference relation $\dot{\succeq}_{\mathcal{C}}$ are implied by various assumed properties of the choice function $\mathcal{C}$. It is well known for instance that if $\mathcal{C}$ satisfies both Set Contraction and Set Expansion then $\dot{\succeq}_{\mathcal{C}}$ so defined is both complete and transitive (see Sen [9, Theorem II]). In this case, as we learnt from Theorem 1(5), our three choice rules coincide and so it is reasonable to speak without further qualification of the revealed preference relation $\dot{\succeq}_{\mathcal{C}}$ as rationalising or explaining the choices represented by $\mathcal{C}$. But when either transitivity and completeness fails for the choice function then this neat relationship breaks down. Indeed in the absence of grounds for presuming completeness, the underlying conception of revealed preference becomes much less compelling.

The fundamental problem with the usual definition of the revealed preference relation is that it does not allow for any distinction between an attitude of indifference between two alternatives and an inability to compare them. Indeed the effect of Revealed Preference is to collapse the two since it entails that $\alpha \dot{\approx}_{\mathcal{C}} \beta \Leftrightarrow \exists A, B \subseteq X$ such that $\alpha, \beta \in A \cap B$, $\alpha \in \mathcal{C}(A)$ and $\beta \in \mathcal{C}(B)$ and so ascribes to the agent an attitude of indifference between any two alternatives that can permissibly be chosen from some set containing both—in particular to any alternatives α, β such that $\mathcal{C}(\{\alpha, \beta\}) = \{\alpha, \beta\}$—irrespective of whether they are comparable or not.

To allow for incomparability we need to build a revealed weak preference relation up from its component revealed strict preference and indifference relations. I suggest that the following definitions encode the correct way to do so from a choice function $\mathcal{C}$.

RSP: $\alpha \succ_{\mathcal{C}} \beta \Leftrightarrow \forall (A : \alpha \in A), \beta \notin \mathcal{C}(A)$
RI: $\alpha \approx_{\mathcal{C}} \beta \Leftrightarrow \forall (A : \alpha, \beta \in A), \alpha \in \mathcal{C}(A) \Leftrightarrow \beta \in \mathcal{C}(A)$
RWP: $\alpha \succeq_{\mathcal{C}} \beta \Leftrightarrow \alpha \succ_{\mathcal{C}} \beta$ or $\alpha \approx_{\mathcal{C}} \beta$

RSP strengthens SPBC into a biconditional that mandates the inference that one prospect is strictly preferred to another iff the latter is never chosen when the former is available. Note that RSP implies that $\beta \notin \mathcal{C}(\{\alpha, \beta\})$ if $\alpha \succ_{\mathcal{C}} \beta$. The converse

is only true however if $\mathcal{C}$ satisfies Set Contraction. Put more positively, if a choice function satisfies Set Contraction then the revealed strict preference relation based on it is binary in composition.

RI similarly strengthens IBC into a biconditional, but the inference it mandates is more controversial; namely that two alternatives are indifferent iff they are either both chosen or both not. The intuition underlying RI is that what distinguishes indifference between two alternatives from their incomparability is that in the former case (indifference) no third alternative should be strictly preferred to one, but not the other, of the pair, while in the latter case (incomparability) such a third alternative could exist. The problem is that if the set of alternatives is sufficiently sparse such a third alternative might not in fact exist and then RI would mandate an inference of indifference when the case is actually one of incomparability. On the other hand, when the underlying set of alternatives contains, for every pair of alternatives α and β, a third alternative α^+, that is comparably better than α, or alternative β^- that is comparably worse than β, then RI will be applicable.

RWP defines $\succeq_{\mathcal{C}}$ in terms of the relations of strict preference, $\succ_{\mathcal{C}}$, and indifference, $\approx_{\mathcal{C}}$, that are revealed by the choice function $\mathcal{C}$ in accordance with RSP and RI. So defined $\succeq_{\mathcal{C}}$ is not necessarily complete, since it can be the case that there are sets A and B such that $\alpha, \beta \in A, B$ but $\alpha \in \mathcal{C}(A)$ and $\beta \in \mathcal{C}(B)$. This would arise when α and β are incomparable and A and B contain elements that respectively dominate β and α but not the other. Furthermore, although $\approx_{\mathcal{C}}$ must be symmetric and $\succeq_{\mathcal{C}}$ reflexive, in the absence of any further assumptions about $\mathcal{C}$ it is not assured that $\succeq_{\mathcal{C}}$ is a weak preference relation, nor that $\succ_{\mathcal{C}}$ and $\approx_{\mathcal{C}}$ are its symmetric and anti-symmetric parts. For this we must assume that $\mathcal{C}$ is decisive.

Theorem 4 *Suppose that $\mathcal{C}$ is decisive and that $\succeq_{\mathcal{C}}$ is defined from $\mathcal{C}$ in accordance with RSP, RI and RWP. Then $\succeq_{\mathcal{C}}$ is a weakly consistent weak preference relation with symmetric and anti-symmetric parts $\succ_{\mathcal{C}}$ and $\approx_{\mathcal{C}}$.*

Proof RI implies the symmetry of $\approx_{\mathcal{C}}$ and, together with RWP, the reflexivity of $\succeq_{\mathcal{C}}$. Note firstly that it is not possible that both $\alpha \succ_{\mathcal{C}} \beta$ and that $\alpha \approx_{\mathcal{C}} \beta$. For if $\alpha \succ_{\mathcal{C}} \beta$, then by RSP $\beta \notin \mathcal{C}(\{\alpha, \beta\})$. So by decisiveness $\alpha \in \mathcal{C}(\{\alpha, \beta\})$ and hence by RI $\alpha \not\approx_{\mathcal{C}} \beta$. Similarly if $\alpha \approx_{\mathcal{C}} \beta$ then by RI and decisiveness $\mathcal{C}(\{\alpha, \beta\}) = \{\alpha, \beta\}$. So by RSP, $\alpha \not\succ_{\mathcal{C}} \beta$. To establish the anti-symmetry of $\succ_{\mathcal{C}}$, let $\Gamma := \{A \subseteq X : \alpha, \beta \in A\}$. Suppose that $\alpha \succ_{\mathcal{C}} \beta$ so that by RSP, $\forall A \in \Gamma, \beta \notin \mathcal{C}(A)$. Then $\beta \notin \mathcal{C}(\{\alpha, \beta)\}$ and hence by decisiveness, $\alpha \in \mathcal{C}(\{\alpha, \beta)\}$. So it is not the case that $\forall A \in \Gamma, \alpha \notin \mathcal{C}(A)$, i.e. $\beta \not\succ_{\mathcal{C}} \alpha$. Finally suppose that, contrary to hypothesis, $\succeq_{\mathcal{C}}$ is not weakly consistent. Then there exists a sequence of alternatives $\alpha_1, \alpha_2, \ldots, \alpha_n$, such that, $\alpha_1 \succ_{\mathcal{C}} \alpha_2, \alpha_2 \succ_{\mathcal{C}} \alpha_3, \ldots, \alpha_{n-1} \succ_{\mathcal{C}} \alpha_n$ and $\alpha_n \succ_{\mathcal{C}} \alpha_1$. Then by RSP, $\mathcal{C}(\{\alpha_1, \alpha_2, \ldots, \alpha_n\}) = \varnothing$ contrary to decisiveness. So $\succeq_{\mathcal{C}}$ must be weakly consistent. ■

3.2 Conditions for Rationalisability

Let us now turn to the question of whether it is possible in general to rationalise an arbitrary choice function $\mathcal{C}$ in terms of the revealed weak preference relation $\succeq_{\mathcal{C}}$ defined by RWP. As is to be expected, without some restrictions on $\mathcal{C}$ and/or the set of alternatives, the answer is negative for each of the three types of rationalisability under consideration.

1. *O-rationalisability*: Consider $\mathcal{C}$ and set of alternatives $\{\alpha, \beta, \gamma\}$ such that $\mathcal{C}(\{\alpha, \beta\}) = \{\alpha, \beta\}$ but $\mathcal{C}(\{\alpha, \beta, \gamma\}) = \{\beta, \gamma\}$. Then by RWP, $\alpha \not\succeq_{\mathcal{C}} \beta$ and $\beta \not\succeq_{\mathcal{C}} \alpha$. So $\mathcal{C}^{Opt}_{\succeq_{\mathcal{C}}}(\{\alpha, \beta\}) = \varnothing \neq \mathcal{C}(\{\alpha, \beta\})$.
2. *M-rationalisability*: Consider $\mathcal{C}$ and set of alternatives $\{\alpha, \beta, \gamma\}$ such that $\mathcal{C}(\{\alpha, \beta\}) = \{\alpha\}$, $\mathcal{C}(\{\beta, \gamma\}) = \{\beta, \gamma\}$, $\mathcal{C}(\{\alpha, \gamma\}) = \{\alpha, \gamma\}$ but $\mathcal{C}(\{\alpha, \beta, \gamma\}) = \{\alpha\}$. Then by RWP, $\alpha \succ_{\mathcal{C}} \beta$, $\beta \approx_{\mathcal{C}} \gamma$ but $\gamma \not\trianglerighteq_{\mathcal{C}} \alpha$. So $\mathcal{C}^{Max}_{\succeq_{\mathcal{C}}}(\{\alpha, \beta, \gamma\}) = \{\alpha, \gamma\} \neq \mathcal{C}(\{\alpha, \beta, \gamma\})$.
3. *SM-rationalisability*: Consider $\mathcal{C}$ and set of alternatives $\{\alpha, \beta, \gamma\}$ such that $\mathcal{C}(\{\alpha, \beta\}) = \{\alpha\}$, $\mathcal{C}(\{\beta, \gamma\}) = \{\beta\}$, and $\mathcal{C}(\{\alpha, \gamma\}) = \{\gamma\}$. So by RWP, $\alpha \not\trianglerighteq_{\mathcal{C}} \beta$, $\beta \not\trianglerighteq_{\mathcal{C}} \gamma$ and $\gamma \not\trianglerighteq_{\mathcal{C}} \alpha$. But then $\mathcal{C}^{SM}_{\succeq_{\mathcal{C}}}(\{\alpha, \beta\}) = \{\alpha, \beta\} \neq \mathcal{C}(\{\alpha, \beta\})$.

What conditions on $\mathcal{C}$ are sufficient to ensure rationalisability? Our earlier observation that satisfaction of Set Contraction and Set Expansion is sufficient for O-rationalisability extends to both M- and SM-rationalisability: this is a consequence of Theorem 1(5). This result is of marginal interest however since these conditions are very restrictive and indeed suffice to ensure the completeness of the revealed preference relation.

It is possible to do better. Blair et al. [3] prove that a choice function satisfies Set Union and Set Contraction iff there exists a weakly consistent preference weak relation that rationalises it. Since both conditions are also implied by Maximality, this theorem provides the required characterisation of consistent maximal choice. Below, in Theorem 5, we establish that the weak preference relation defined by RWP, RSP and RI is just such a rationalising relation.

Set Contraction and Set Union are in fact also jointly sufficient for a SM-rationalisation, but in this case it does not give us the characterisation that we seek since Set Union is not necessary for preference-based strongly maximal choice. What is required, it turns out, are the two weaker conditions we introduced: Element Union and Equivalence Class Union. In Theorem 6 below we show that it is sufficient that the choice function be decisive and satisfy these two conditions along with Set Contraction, for it to have a Suzumura consistent SM-rationalisation. This gives us the characterisation of consistent, strongly maximal choice that we want, namely that *it is necessary and sufficient that a choice function be decisive and satisfy Set Contraction, Element Union and Equivalence Class Union that it be SM-rationalisable by a Suzumura consistent weak preference relation*. This is proved below as a corollary of Theorem 6. (In the proofs that follow, we omit the subscripts from the relations induced by the choice function $\mathcal{C}$.)

Theorem 5 *Suppose that $\mathcal{C}$ is a decisive choice function satisfying Set Contraction and Set Union. Let $\succeq$ be defined from $\mathbf{C}$ by RWP, RSP and RI. Then $\succeq$ is a weakly consistent weak preference relation that M-rationalises $\mathcal{C}$.*

Proof Suppose that $\alpha \notin \mathcal{C}^M_{\succeq}(A)$. Then by RSP and RI there exists $\beta \in A$ such that $\forall(B : \alpha \in B)$, $\alpha \notin \mathcal{C}(B)$. So in particular, $\alpha \notin \mathcal{C}(A)$. Now suppose that $\alpha \in \mathcal{C}^M_{\succeq}(A)$. Then by RSP there does not exist any $\beta \in A$ such that $\forall(B \subset X : \beta \in B)$, $\alpha \notin \mathcal{C}(B)$. Hence for all $\beta_i \in A$, there exists a set $B_i \subseteq X$ such that $\beta \in B_i$ and $\alpha \in \mathcal{C}(B_i)$. But then by Set Union, $\alpha \in \mathcal{C}(\cup B_i) = \mathcal{C}(A)$. The weak consistency of $\succeq$ then follows from Theorem 2(2). ■

Lemma 1 *Suppose that choice function $\mathcal{C}$ and indifference relation $\approx$ are related in accordance with RI. Then for all $\gamma \in X$:*

$$\gamma \hat{\approx}^A \alpha \Leftrightarrow \gamma \in \tilde{\mathcal{C}}(A, \alpha)$$

Proof Suppose that $\gamma \hat{\approx}^A \alpha$. Then then there exists a sequence of elements in A, α_1, $\alpha_2, \ldots, \alpha_n$ linking α and γ in the sense that $\alpha = \alpha_1, \alpha_1 \approx \alpha_2, \ldots,$ and $\alpha_n = \gamma$. Hence by RI, for all α_i in the sequence, $\alpha_i \in \mathcal{C}(B) \Leftrightarrow \alpha_{i+1} \in \mathcal{C}(B)$ for all $B \subseteq X$ such that $\alpha_i, \alpha_{i+1} \in B$. Now $\alpha \in \tilde{\mathcal{C}}(A, \alpha)$. And so since $\alpha_1 \in \tilde{\mathcal{C}}(A, \alpha)$, it follows by the definition of $\tilde{\mathcal{C}}$ that $\alpha_2 \in \tilde{\mathcal{C}}(A, \alpha)$ and hence $\ldots$ $\alpha_n \in \tilde{\mathcal{C}}(A, \alpha)$. We conclude that $\gamma \in \tilde{\mathcal{C}}(A, \alpha)$.

We establish the other direction by proving by induction that if $\gamma \in \tilde{\mathcal{C}}^\tau(A, \alpha)$ then $\gamma \hat{\approx}^A \alpha$ for all $\tau \geq 0$. Suppose that $\tau = 0$. Then $\gamma = \alpha$ and so it follows from the symmetry of $\approx$ that $\gamma \hat{\approx}^A \alpha$. Next assume true for $\tau = k$, i.e. if $\gamma \in \tilde{\mathcal{C}}^k(A, \alpha)$ then $\gamma \hat{\approx}^A \alpha$. Now we prove the hypothesis for $\tau = k + 1$. Suppose that $\gamma \in \tilde{\mathcal{C}}^{k+1}(A, \alpha)$. Then there exists $\beta \in \tilde{\mathcal{C}}^k(A, \alpha)$ such that $\beta \in \mathcal{C}(B) \Leftrightarrow \gamma \in \mathcal{C}(B)$ for all $B \subseteq X$ such that $\beta, \gamma \in B$. Hence by RI, $\gamma \approx \beta$. But by assumption $\beta \hat{\approx}^A \alpha$. So $\gamma \hat{\approx}^A \alpha$. ■

Corollary 1 *If $\gamma \hat{\approx}^A \alpha$ then $\tilde{\mathcal{C}}(A, \gamma) = \tilde{\mathcal{C}}(A, \alpha)$*

Theorem 6 *Suppose that $\mathcal{C}$ is a decisive choice function satisfying Set Contraction, Element Union and Equivalence Class Union. Let $\succeq$ be defined from $\mathbf{C}$ by RWP, RSP and RI. Then $\succeq$ is a Suzumura consistent weak preference relation that SM-rationalises $\mathcal{C}$.*

Proof Suppose that $\alpha \notin \mathcal{C}^{SM}_{\succeq}(A)$. Then by RSP and RI there exists $\beta, \gamma \in A$ such that $\beta \succ \gamma$ and $\gamma \hat{\approx}^A \alpha$. This implies that there exists $\alpha_1, \alpha_2, \ldots, \alpha_{n-1}, \alpha_n \in A$ such that $\gamma = \alpha_1, \alpha_1 \approx_C \alpha_2, \ldots, \alpha_{n-1} \approx_C \alpha_n$, and $\alpha_n = \alpha$. Now since $\beta \in A$, we know that $\gamma \notin \mathcal{C}(A)$. So by RI, $\alpha_2 \notin \mathcal{C}(A)$. Hence by RI, $\alpha_3 \notin \mathcal{C}(A)$ .... and hence by RI, $\alpha \notin \mathcal{C}(A)$.

Now suppose that $\alpha \in \mathcal{C}^{SM}_{\succeq}(A)$. Let $\Gamma = \{\gamma \in A : \gamma \hat{\approx}^A \alpha\}$. Then $\forall \gamma \in \Gamma$, there exists no $\delta \in A$ such that $\beta \succ \gamma$. Hence also for all such γ there exists no $\gamma', \delta' \in A$ such that $\gamma \hat{\approx}^A \gamma'$ and $\delta' \succ \gamma'$. So $\forall \gamma \in \Gamma$, $\gamma \in \mathcal{C}^{SM}_{\succeq}(A)$. Now by RSP, $\forall \gamma \in \Gamma$, there does not exist any $\beta \in A$ such that $\forall(B \subseteq X : \beta \in B)$, $\gamma \notin \mathcal{C}(B)$. Hence for all $\beta \in A$, there exists a set $B \subseteq X$ such that $\beta \in B$ and $\gamma \in \mathcal{C}(B)$. But then by Set Contraction, $\gamma \in \mathcal{C}(\{\bigcirc, \beta\})$. So by Element Union, there exist $\gamma^* \in \Gamma$

such that $\gamma^* \in \mathcal{C}(\Gamma \cup \{\beta\})$. Now by Lemma 1, since $\approx$ is constructed from $\mathcal{C}$ in accordance with RI, $\tilde{\mathcal{C}}(A, \gamma^*) = \tilde{\mathcal{C}}(A, \alpha) = \Gamma$. Hence $\tilde{\mathcal{C}}(A, \alpha) \subseteq \mathcal{C}(\Gamma \cup \{\beta\})$ for all $\beta \in A$. Hence by Equivalence Class Union, $\tilde{\mathcal{C}}(A, \alpha) \subseteq \mathcal{C}(A)$. But then it follows from the fact that $\alpha \in \tilde{\mathcal{C}}(A, \alpha)$, that $\alpha \in \mathcal{C}(A)$. So $\mathcal{C} = \mathcal{C}^{SM}_{\succeq}$. Finally the Suzumura consistency of $\succeq$ follows from Theorem 2(3). ■

Corollary 2 *$\mathcal{C}$ is a decisive choice function satisfying Set Contraction, Element Union and Equivalence Class Union iff there exists a Suzumura consistent weak preference relation $\succeq$ that SM-rationalises $\mathcal{C}$.*

Proof Follows from Theorems 6 and 3(1) and (4). ■

4 Conclusion

When preferences are incomplete, as they often are, they will not suffice to determine a unique choice from all sets of alternatives. Nonetheless, it is useful to know what choices an agent's preferences permit her to make. In this paper I have proposed a new choice rule—Strong Maximality—and argued that it better characterises rational preference-based choice than the more familiar rules of Maximality and Optimality. Only Strong Maximality respects both the requirement that an alternative never be chosen when something strictly preferred to it is available (PBC) and the requirement that two alternatives that are comparably indifferent to one another must either both be chosen or both not chosen from any set containing both (IBC).

When preferences are transitive, Strong Maximality will yield the same prescriptions as Maximality, when preferences are also complete both will coincide with the prescriptions of Optimality. But just as recognition of the rational permissibility of incompleteness motivates a move from Optimality to Maximality, so the recognition that transitivity is too strong a requirement motivates a move from Maximality to Strong Maximality.

Strong Maximality is closely linked to the requirement that preferences be Suzumura consistent; in particular Suzumura consistency is both necessary and sufficient for decisive strongly maximal choice. These two concepts are thus mutually supportive in the same way as are the concepts of maximal choice and weak consistency and the concepts of optimal choice and transitivity. And just as weak consistency is too weak and transitivity too strong, so too is maximal choice too permissive and optimal choice too demanding. Strong Maximality and Suzumura consistency are, like small bear's porridge, just right.

Acknowledgements I would like to thank the audience of the LSE Choice Group seminar for their usual robust questioning of a presentation of this paper. I owe special thanks to David Makinson, Wulf Gaertner, Silvia Milano and two anonymous referees for very helpful written comments on earlier drafts.

References

1. Arrow KJ (1959) Rational choice functions and orderings. Economica 26:121–127
2. Baigent N (2007) Choices, norms and preference revelation. Analyse & Kritik: Zeitschrift fur Sozialtheorie 2:139–145
3. Blair DH, Bordes G, Kelly JS, Suzumura K (1975) Impossibility theorems without collective rationality. J Econ Theory 13:361–379
4. Bossert W, Suzumura K (2010) Consistency, choice, and rationality. Harvard University Press, Cambridge
5. Bossert WY, Sprumont Y, Suzumura K (2005) Consistent rationalisability. Economica 72:185–200
6. Kreps DM (1988) Notes on the theory of choice. Westview Press, Boulder and London
7. Richter RK (1966) Revealed preference theory. Econometrica 34:635–645
8. Sen AK (1969) Choice functions and revealed preference. Rev Econ Stud 36(3):381–391
9. Sen AK (1971) Choice functions and revealed preference. Rev Econ Stud 38(3):307–317
10. Sen AK (1997) Maximization and the act of choice. Econometrica 65:745–780
11. Suzumura K (1976) Remarks on the theory of collective choice. Economica 43:381–390

Rationality and Context-Dependent Preferences

Prasanta K. Pattanaik and Yongsheng Xu

Abstract The standard theory of rational choice in economics considers an agent's choices to be rational if and only if the agent makes her choices in different choice situations on the basis of a fixed preference ordering defined over the set of all possible options. This implies that a rational agent's preferences cannot be context-dependent. This paper outlines a simple framework for defining context-dependence of preferences and for discussing relationships between context-dependent preferences and the notion of rationality.

Keywords Context-dependence • Context-independence • Preference • Rationality • Standard theory

1 Introduction

An important criticism of the economists' theory of rational choice is that it rules out the possibility that an agent's preferences, which constitute the basis of her choices, may be context-dependent. The purpose of this paper is to introduce a simple framework to define the concept of context-dependent preferences of an agent, to indicate how a number of phenomena described in the literature can be fitted in this framework, and to explore the significance of context-dependent preferences for the notion of rational choice in economics.

To see how the concept of context-independence of preferences is embedded in the concept of rational choice in economics, it may be useful to note that, methodologically, there are two distinct approaches in the economists' theory of choice, namely, the preference-based approach which starts with preference as a

P.K. Pattanaik (✉)
Department of Economics, University of California, Riverside, CA 92521, USA
e-mail: prasanta.pattanaik@ucr.edu

Y. Xu
Department of Economics, Georgia State University, Atlanta, GA 30302, USA
e-mail: yxu3@gsu.edu

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_4

primitive concept in the model, and the choice-based approach (also widely known as the revealed preference approach), which starts with choice, but not preference, as the primitive concept. In this paper, we focus on the preference-based approach, but what we have to say about rational choice and context-dependent preferences in this approach can be readily adapted to the choice-based approach.

In the preference-based approach, an agent is said to be rational (or, equivalently, an agent is said to be choosing rationally) if she has a given preference ordering defined over the universal set of options (i.e., the set of all conceivable options) and if, for every possible set of feasible options that may confront her, she always chooses from that set of options[1] which are the best options in that set, "best" being defined in terms of her given preference ordering over the universal set of options. This notion of rational choice by an agent can be conceptually split into two components. The first component is the idea that a rational agent has a fixed binary weak preference relation ("at least as good as") over the set of all conceivable options, which constitutes the sole basis of her choices from different sets of feasible options; the second component is the assumption that this fixed binary weak preference relation over the universal set, which serves as the basis of the agent's choices in all contexts, is an ordering (i.e., it satisfies reflexivity, connectedness (or completeness), and transitivity). Intuitively, the existence of a fixed binary weak preference relation over the set of all conceivable options, which determines the agent's choice(s) from every set of feasible options in all contexts, implies that the specific context in which the agent chooses from a set of feasible options does not matter at all for her preferences over the options in that feasible set and that all the information that she considers to be relevant for her preferences are already contained in the specification of the options at the outset; this feature has sometimes been called the axiom of context-independence [13]. In this paper, we are primarily concerned with the first constituent component, to wit, context-independence of preferences, of the economists' conception of rational choice, though we also comment on the second constituent component, namely, the assumption that the fixed binary weak preference relation over the universal set of options is an ordering.

Section 2 develops a simple framework in which the notion of context-dependent preferences can be defined precisely and explicitly. Also, in Sect. 2, we present several examples that are well-known in the literature and show how they fit in our framework. In Sect. 3, we discuss the formal link between context-independence and the conventional concept of rational choice. Finally, in Sect. 4, we comment on the significance of context-dependent preferences for the economic theory of rational choice.

[1] Following a fairly common practice in the literature, we are permitting the choice set for a given set of feasible options to have multiple elements.

2 The Framework

Let X be a given, non-empty universal set of *conventionally defined* options. For example, in a competitive consumer's choice problem in economics, X is the set of all commodity bundles, a commodity bundle being a specification of the quantities of all commodities under consideration. If the choice of an entree for dinner is the problem at hand, then the set of all possible alternative entrees constitutes the universal set of (conventionally defined) options. Y denotes a non-empty subclass of the class of all non-empty subsets of X, the elements of Y being denoted by A, B, etc. The interpretation of Y is that Y is the class of all possible alternative menus or sets of feasible conventionally defined options from which the agent may have to choose. Y is permitted, but not constrained, to be the class of all non-empty subsets of X. In the standard economic theory of rational choice, an agent makes her choices rationally if and only if: (1) she has a fixed binary weak preference relation $\succeq$ over X, such that, for all $A \in Y$, the agent's set of chosen elements in A are given by $\{x \in A : \text{for all } y \in A, x \succeq y\}$; and (2) this fixed binary weak preference relation $\succeq$ is an ordering. There are, however, numerous examples in the literature where the agent does not seem to make her choices in this fashion. In such examples, either the agent directly tells us that she strictly prefers an option x in X to another option y in X under certain circumstances but she considers option y to be at least as good as x under certain other circumstances or we observe the agent choosing in a fashion, which is not consistent with the assumption that she has a fixed binary weak preference relation $\succeq$ over X, such that from every $A \in Y$, she chooses the $\succeq -$greatest elements in A.

2.1 Two Examples

It may be useful to start with two well-known examples, where the choice behavior of the agent could not have been generated by any fixed binary relation over the universal set of options (later we consider several other examples, which have been discussed in the literature and which have the same feature).

Example 1(a): Menu-Dependence

The first example here is a slightly modified version of an example due to Sen [17]. Consider the following choices of an individual. When a fruit basket on the dinner table contains many apples and many oranges, the agent, who is one of several guests at the dinner, chooses an apple rather than any of the oranges, but, had the fruit basket on the table contained the same oranges but exactly one apple,

the individual would have chosen an orange rather than the single apple. It is clear that there cannot be any fixed binary weak preference relation $\succeq$ over the fruits that can induce the choice of an apple over the other fruits from the first fruit basket and the choice of an orange over the apple from the second fruit basket. The intuitive explanation given by Sen for such choices is that choosing the single apple from the second basket would be rude while no rudeness is involved in choosing one of several available apples.

Example 1(b): Menu-Dependence

Consider another example, which is due to Luce and Raiffa [14]. The waiter in a restaurant gives a customer, who does not know much about the restaurant, a menu for the day's dishes. The menu contains two items: steak and fish. The customer chooses fish. A little later, the waiter reports that, because of a mistake, frog's legs have been omitted from the day's menu but they are available. The customer then chooses steak from this expanded menu. Again, there cannot be any fixed binary weak preference relation $\succeq$ over all conceivable dishes, such that, in terms of $\succeq$, fish is the most preferred dish in the initial menu consisting of fish and steak while steak is the most preferred dish in the expanded menu consisting of fish, steak, and frog's legs. The explanation that Luce and Raiffa [14, p. 288] give for the agent's choices is as follows. The customer would prefer steak to fish if she had reasonable assurance that the restaurant was good so that steak will be well prepared, but, in the absence of any information about the quality of the restaurant, she prefers fish to steak. The customer's past experience tells her that frog's legs are served only in good restaurants. So, when she knows that frog's legs are in the menu, she believes that the restaurant must be good and accordingly prefers steak to fish (she prefers steak to frog's leg in any case).

In Example 1(a), whether or not a particular fruit is the only one of its kind in the fruit basket is a relevant consideration for the agent, but no information about this is available in the description of the an apple or the description of an orange as such. Only when the context for the agent's choice is given in the form of a specific fruit basket in front of her, does she get additional information about whether an apple or an orange is the only fruit of its kind available to her. Similarly, in Example 1(b) the customer's preference over fish and meat depends on contextual information regarding the availability of frog's legs in the menu. In both cases, the relevance, for the agent, of the information, which is to be found in the contextual features but not in the descriptions of the options in X, causes problem for the standard theory of rational choice formulated in terms of a fixed binary weak preference relation defined over the options. Given this, one possibility may be to reformulate explicitly the theory so as to allow for the fact that, in assessing the relative desirability of any two options, the agent takes into account the contextual circumstances, in which the options are presented to her, besides the information contained in the specification

of those options themselves.[2] But what exactly is the demarcating line between the features which are parts of the specification of options and the features which constitute parts of the descriptions of contexts? If the theorist constructing the model of choice knows that certain features affect the relative desirability, for the agent, of options as specified by her (i.e., the theorist), then why should the theorist not re-specify the options so as to make these features a part of the descriptions of these re-specified options? These are questions which arise naturally. In fact, Savage's [16, Chap. 2] discussion of the largeness or the smallness of the "world" (i.e., "the objects about which the person is concerned") is essentially concerned with this issue. As one switches progressively from narrower conceptions of options to broader conceptions of options, more and more of the features which can be regarded as contextual features under narrower conceptions will be subsumed in the description of options conceived more broadly. There are clearly advantages in starting with a very broad notion of options (or a "sufficiently broad world", to use Savage's [16] terminology). But as Savage [16, p. 9] points out, "the use of modest little worlds, tailored to particular contexts is often a simplification, the advantage of which is justified by a considerable body of mathematical experience with related ideas". It seems to us that, in many areas of economics, conventional specifications of the agent's options are simplifications of this type. Though it may be known that certain features outside such narrow but simple specifications may be considered relevant by the agent, there may not be sufficient agreement about the significance of such features, and so the conventional and narrow specifications of options continue to be used. An example is the theory of consumers' behavior in economics, where commodity bundles are assumed to be the objects with which the consumer is concerned. While it may be recognized that a consumer's preferences over a pair of commodity bundles may depend on certain features of the "context" in which the consumer faces the task of comparing two commodity bundles, there may not be any general consensus about the importance of such contextual features, and, therefore, the conventional specification of options as commodity bundles continues to be used. Also, often the theorist may not know what matters for an agent beyond the features captured by a certain specification of options; in such cases, the contextual features will represent the gap between what features the agent perceives to be relevant for her own preferences and what the theorist initially perceives to be relevant for the agent. Once the theorist becomes aware of this gap, she has the option of re-specifying the options in her theory, incorporating the features missed out in her initial specification; we shall take up this issue again in Sect. 4.

[2] See Baigent [1], Baigent and Gaertner [2], Bhattacharyya et al. [5], Bossert and Suzumura [6, 7], Gaertner and Xu [8], and Xu [19] for some axiomatic studies along this line of research.

2.2 Context-Dependent Preferences

Let O be a set of all (mutually exclusive) contextual features (or, simply, contexts), which may affect the relative desirability or undesirability of conventionally defined options for the agent under consideration but which do not constitute a part of the descriptions of the options in X. Given X and O, we now introduce a simple but general framework which permits (without making it mandatory) context-dependence of the agent's preferences. For all $A \in Y$, let $O(A)$ be the (non-empty) set all contexts $o \in O$, such that the agent may have to choose from A given the context o (note that there may be many different contexts in which x may be chosen from A, so that $O(A)$ may have more than one element). Let S be the set of all (x, o) in $X \times O$, such that, for some A in Y, $x \in A$ and $o \in O(A)$. It is possible that there may be $x \in X$ such that, for all $A \in Y$, $x \notin A$; let X' be the set of all such x.

Let R be a binary relation over S. For all $(x, o), (y, o') \in S$, $(x, o)R(y, o')$ will be interpreted as "the agent considers x in the context o to be at least as desirable as y in the context o'". We will use I and P, respectively, to denote the symmetric and asymmetric parts of R, that is, for all $(x, o), (y, o') \in S$, $[(x, o)I(y, o')$ iff $((x, o)R(y, o')$ and $(y, o')R(x, o))]$ and $[(x, o)P(y, o')$ iff $((x, o)R(y, o')$ and not$(y, o')R(x, o))]$. For all $A \in Y$ and all $o \in O(A)$, when the agent chooses from A given the context o, the set of her chosen options is $\{x \in A : (x, o)R(y, o)$ for all $(y, o) \in A \times \{o\}\}$. Essentially, this simple framework augments the informational basis of the theory of choice by introducing explicitly into the formal framework information about the different contexts in which the agent may have to make her choices. For the sake of convenience, we shall refer to this framework as the *context-inclusive framework*.

We now introduce three notions of context-independence of the preferences represented by R in our context-inclusive framework.

Context-independence (I). For all $(x, o), (y, o'), (x, \bar{o}), (y, \bar{o}') \in S$, $[(x, o)R(y, o')$ iff $(x, \bar{o})R(y, \bar{o}')]$ and $[(y, o')R(x, o)$ iff $(y, \bar{o}')R(x, \bar{o})]$.

The following is a somewhat weaker notion of context-independence, which stipulates that the ranking of any two options, x and y given the contextual features o is exactly analogous to the ranking of x and y given any other contextual features o'.

Context-independence (II). For all $(x, o), (y, o), (x, o'), (y, o') \in S$, $[(x, o)R(y, o)$ iff $(x, o')R(y, o')]$ and $[(y, o)R(x, o)$ iff $(y, o')R(x, o')]$.

Our last formulation of context-independence stipulates that the desirability of an option does not depend on the context in which it is chosen.

Context-independence (III). For all $(x, o), (x, o') \in S, (x, o)I(y, o')$.

We say that R is *context-dependent (I)* (resp. *context-dependent (II), resp. context-dependent (III)*) iff R does not satisfy context-independence (I) (resp. context-independence (II), resp. context-independence (III)).

Our first result examines the relationship between these notions of context-independence.

Proposition 1

(1.i) If R is context-independent (I), then it is context-independent (II), and, if R reflexive and context-independent (I), then it is context-independent (III).

(1.ii) If R is transitive, then context-independence (III) of R implies context-independence (I) of R and context-independence (II) of R

Proof

(1.i) It is obvious that, if R is context-independent (I), then R is context-independent (II). We show that, if R is reflexive and context-independent (I), then R is also context-independent (III). Suppose R is reflexive and context-independent (I). Let $(x,o),(x,o') \in S$. By context-independence (I), for all $(z,o^1),(w,o^2),(z,o^3),(w,o^4) \in O$, $[(z,o^1)R(w,o^2)$ iff $(z,o^3)R(w,o^4)]$ and $[(w,o^2)R(z,o^1)$ iff $(w,o^4)R(z,o^3)]$. Then, letting $z = w = x, o^1 = o^2 = o^3 = o$, and $o^4 = o'$, and noting $(x,o)I(x,o)$, which follows from reflexivity of R, we have $(x,o)I(x,o')$.

(1.ii) Suppose R is transitive and context-independent (III). Then we show that R is context-independent (I). Let $(x,o),(y,o'),(x,\bar{o})$, and $(y,\bar{o}')$ be in S. Suppose $(x,o)R(y,o')$. By context-independence(III) of R, we have $(x,o)I(x,\bar{o})$ and $(y,o')I(y,\bar{o}')$. Then, by the transitivity of R, we obtain $(x,\bar{o})R(y,\bar{o}')$. Similarly, we can show that, if $(x,\bar{o})R(y,\bar{o}')$ then $(x,o)R(y,o')$, if $(y,o')R(x,o)$ then $(y,\bar{o}')R(x,\bar{o})$, and, if $(y,\bar{o}')R(x,\bar{o})$ then $(y,o')R(x,o)$. Therefore, we have $[(x,o)R(y,o')$ iff $(x,\bar{o})R(y,\bar{o}')]$ and $[(y,o')R(x,o)$ iff $(y,\bar{o}')R(x,\bar{o})]$. Thus, R is context-independent (I).

Since, in the presence of transitivity, context-independence (III) implies context-independence (I), by Proposition (1.i), it follows that, in the presence of transitivity, context-dependence (III) implies context-dependence (II). ■

As a corollary to Proposition 1, we have the following result.

Corollary 1 *Suppose R is reflexive and transitive. Then, (i) context-independence (I) of R and context-independence (III) of R are equivalent; and (ii) context-independence (III) of R, as well as context-independence (I) of R, implies context-independence (II) of R.*

In the presence of reflexivity and transitivity of R, context-independence (II) of R is strictly weaker than each of context-independence (I) of R and context-independence (III) of R as shown by the example in the following remark.

Remark 1 Suppose $O = \{o_1, \cdots, o_n\}$ with $n \geq 2$ and $S = X \times O$. Let $\succeq$ over X be an ordering, and define R^* over S such that, for all $(x,o_i),(y,o_j) \in S$, if $i > j$ then $(x,o_i)P^*(y,o_j)$, and if $i = j$ then $(x,o_i)R^*(y,o_j)$ iff $x \succeq y$. Then, R^* is context-independent (II) since, for all $(x,o_i),(y,o_i),(x,o_j),(y,o_j) \in S$, from the definition of R^*, we have $[(x,o_i)R^*(y,o_i)$ iff $x \succeq y$ and $x \succeq y$ iff $(x,o_j)R^*(y,o_j)]$ implying that $(x,o_i)R^*(y,o_i)$ iff $(x,o_j)R^*(y,o_j)$. That R^*

is an ordering follows from the fact that $\succeq$ is an ordering and the definition of R^*. On the other hand, R^* is not context-independent (III) by noting, for example, $(x, o_1)P(x, o_2)$. Given that R^* is an ordering, that R^* is not context-independent (I) follows from Corollary 1 and that R^* is not context-independent (III). Thus, the intuition that context-independence (II) of R is strictly weaker than either context-independence (I) or (III) of R seems to point to the fact that, in context-independence (II) of R, contexts may contribute to R independently of the outcomes, while such possibly independent values of contexts do not exist in context-independence (I) or (III) of R.

Remark 2 Proposition 1, Corollary 1, and Remark 1 spell out the formal relations between the three concepts of context-independence. It may, however, be helpful to comment on the intuitive difference between these concepts. Intuitively, what context-independence (I) says is that, to compare two outcomes x and y, the agent does not need any information about the context where x is to be chosen and the context where y is to be chosen. Context-independence (II) says something significantly weaker: it says that, so long as the agent is comparing outcomes x and y in a fixed and unchanging context, the agent's comparison of outcomes is not affected by any information about what that fixed context happens to be. Context-independence (III) is concerned with the agent's assessment of the desirability of a given outcome in two different contexts: it says that the desirability of a given outcome x remains exactly the same even if the context changes. By Proposition 1.i, if R is reflexive, then context-independence (I) implies context-independence (III); and, by Proposition 1.ii, if R is transitive, then context-independence (III) implies context-independence (I). It is difficult to think of any circumstance where R may not be reflexive. So, intuitively, the difference between context-independence (I) and context-independence (III) matters only in the absence of transitivity of R.

2.3 Some Further Examples

The literature on the theory of preference and choice discusses several types of choice behavior which are clearly incompatible with the idea that the agent has a fixed binary weak preference relation, which constitutes the basis of her choices in different choice situations. In what follows, we will consider several such examples and interpret them in terms of our formal framework.

First consider Examples 1(a) and 1(b) above. In Example 1(a), let x be the physically specified option of having an apple and y be the physically specified option of having an orange. The two contexts under consideration in this example can be specified as o = "the fruit basket contains more than one apple and more than one orange" and o' = "the fruit basket contains exactly one apple and several oranges". Consider (x, o), (x, o'), (y, o), and (y, o') in S. Suppose the agent has the binary relation R over S, such that $(x, o)P(y, o)$ and $(y, o')P(x, o')$. It is then clear that: (1) R violates context-independence (II) and, hence, context-independence

(I); and (2) given the set $\{x, y\}$, the agent will choose only x if the context is o and she will choose only y if the context is o'. In Example 1(b), the options are x = fish, y = steak, and z = frog's legs, and the contexts are o = "frog's legs are available in the restaurant" and o' = "frog's legs are not available in the restaurant". R over S is such that $(x, o)\,P\,(y, o)$, $(y, o')\,P\,(x, o')$, and $(y, o')\,P\,(z, o')$. Then R is context-dependent (II), and, hence, context-dependent (I). Given the context o, the agent chooses only x from $\{x, y\}$, and, given the context o', the agent chooses only y from $\{x, y, z\}$.

Example 2: Procedural Considerations

An individual expresses a preference for reading the government's official newspaper when a spectrum of newspapers, from the left to the right and from pro-government to anti-government, is available. However, after the government cracks down on dissenting newspapers by allowing only pro-government newspapers to remain published, the individual changes his preference and now prefers not to read any newspaper [9, 10]. The change of the individual's preference is due to the individual's concerns about procedures that have brought about the change in the availability of newspapers. Contexts in this example are possible procedures that are used to bring about the options under consideration. R is such that:

> (reading the official newspaper; there is no government interference with the publication of newspapers) P (not reading any newspaper; there is no government interference with the publication of newspapers)

and

> (not reading any newspaper; the government has shut down the dissenting newspapers) P (reading the official newspaper; the government has shut down the dissenting newspapers).

Example 3: States of Nature as Contexts

A consumer expresses a preference for a scoop of ice cream over a cup of chicken noodle soup when the weather is hot, while she prefers a cup of chicken noodle soup to a scoop of ice cream when it is cold [3, 4]. It is easy to interpret this problem in our framework by treating hot weather and cold weather as the two contexts The consumer's weak preference relation R in our framework will be as follows:

> (a scoop of ice cream; hot weather) P (a cup of chicken noodle soup; hot weather), and (a cup of chicken noodle soup; cold weather) P (a scoop of ice cream; cold weather).

3 Context-Independence of Preferences and Rationality

In this section, we examine the formal connection between our concepts of context-independence and the notion of rational choice in standard economic theory.

In the standard economic theory, an agent's observed choices are said to be *rationalizable in terms of a binary weak preference relation* if there exists a binary weak preference relation, $\succeq$, over X such that, for all $A \in Y$, the set of $\succeq$ −greatest elements in A coincides with the observed set of options chosen by the agent from A, and the agent's observed choices are said to be rational if they are rationalizable in terms of a preference ordering $\succeq$ over X. Adapting this terminology, we shall say that, an agent's observed choices are *rationalizable in terms of a binary relation*, R, over S in the context-inclusive framework if, for all $A \in Y$ and all $o \in O(A)$, the set $\{x \in A : (x, o)$ is $R-$greatest in $A \times \{o\}\}$ coincides with the observed set of options chosen by the agent from A; and we say that the agent's observed choices are $rational$ in the context-inclusive framework if they are rationalizable in terms of an ordering R over S in that framework.

Proposition 2

(2.i) Suppose, in the context-inclusive framework, an agent's observed choices are rationalizable in terms of an ordering, R, over S, such that R is context-independent (I) or context-independent (III). Then the agent's observed choices are rational in the sense of standard economic theory.

(2.ii) Suppose, in the context-inclusive framework, an agent's observed choices are rationalizable in terms of an ordering, R, over S, such that R is context-independent (II). Then the agent's observed choices are rational in the sense of standard economic theory.

(2.iii) If an agent's observed choices are rational in the sense of standard economic theory, then, in the context-inclusive framework, the agent's observed choices are rationalizable in terms of an ordering R over S, such that R is context-independent (I), context-independent (II), and context-independent (III).

Proof

(2.i) By Proposition 1, if a binary relation R over S is an ordering, then context-independence (I) of R is equivalent to context-independence (III) of R. So we give the proof only for the case where R is context-independent (I).

Suppose, in the context-inclusive framework, the agent's observed choices are rationalizable in terms of an ordering R over S, such that R is context-independent (I). Define a binary weak preference relation $\succeq$ over X as follows:

(i) For all $x \in X - X'$ and all $y \in X'$, $x \succ y$, and, for all $y, z \in X'$, $y \sim z$, where $\succ$ and $\sim$ are, respectively, the strict preference relation and indifference relation corresponding to $\succeq$ (that is, for all $a, b \in X$, $[a \succ b$ iff $a \succeq b$ and not$(b \succeq a)]$ and $[a \sim b$ iff $a \succeq b$ and $b \succeq a]$);

(ii) For all $x, y \in X - X'$, $x \succeq y$ iff $(x, o)R(y, o')$ for some $o, o' \in O$.

Note that, by context independence (I) of R, for all (x,o), (y,o'),$(x,\bar{o})$, $(y,\bar{o}') \in S$, $[(x,o)R(y,o')$ iff $(x,\bar{o})R(y,\bar{o}')]$ and $[(y,o')R(x,o)$ iff $(y,\bar{o}')R(x,\bar{o})]$. Therefore, by (ii), $x \succeq y$ is well-defined for all $x, y \in X - X'$. Given (i), it then follows that $x \succeq y$ is well-defined for all $x, y \in X$.

Given that R is reflexive, connected, and transitive, it can be easily checked that, by (ii), the restriction of $\succeq$ to $X - X'$ is reflexive, connected, and transitive over $X - X'$. Hence, noting (i), $\succeq$ is an ordering over X.

R being context independent (I), for all $A \in Y$, the observed set of options chosen by the agent from A is the same as $\{x \in A : (x,o)R(y,o)$ for all $(y,o) \in A \times \{o\}\}$ for all $o \in O(A)$. From the definition of $\succeq$, it follows that, for all $A \in Y$ and all $o \in O(A)$, $\{x \in A : (x,o)R(y,o)$ for all $(y,o) \in A \times \{o\}\} = \{x \in A : x \succeq y$ for all $y \in A\}$. Therefore, for all $A \in Y$, $\{x \in A : x \succeq y$ for all $y \in A\}$ is the same as the observed set of options chosen by the agent from A.

Thus, the observed choices of the agent are rationalizable in terms of the ordering $\succeq$ in the standard economic framework.

(2.ii) Suppose, in the context-inclusive framework, the agent's observed choices are rationalizable in terms of an ordering R over S, such that R is context-independent (II). Define a binary weak preference relation $\succeq$ over X as follows:

(iii) For all $x \in X - X'$ and all $y \in X'$, $x \succ y$, and, for all $y, z \in X'$, $y \sim z$, where $\succ$ and $\sim$ are, respectively, the strict preference relation and indifference relation corresponding to $\succeq$;

(iv) For all $x, y \in X - X'$, $x \succeq y$ iff $(x,o)R(y,o)$ for some $o \in O$.

Note that, by context independence (II) of R, for all (x,o), (y,o),(x,o'), $(y,o') \in S$, $[(x,o)R(y,o)$ iff $(x,o')R(y,o')]$ and $[(y,o)R(x,o)$ iff $(y,o')R(x,o')]$. Therefore, by (iv), $x \succeq y$ is well-defined for all $x, y \in X - X'$. Given (iii), it then follows that $x \succeq y$ is well-defined for all $x, y \in X$.

Given that R is reflexive, connected, and transitive, it can be easily checked that, by (iv), the restriction of $\succeq$ to $X - X'$ is reflexive, connected, and transitive over $X - X'$. Hence, noting (i), $\succeq$ is an ordering over X.

R being context independent (II), for all $A \in Y$, the observed set of options chosen by the agent from A is the same as $\{x \in A : (x,o)R(y,o)$ for all $(y,o) \in A \times \{o\}\}$ for all $o \in O(A)$. From the definition of $\succeq$, it follows that, for all $A \in Y$ and all $o \in O(A)$, $\{x \in A : (x,o)R(y,o)$ for all $(y,o) \in A \times \{o\}\} = \{x \in A : x \succeq y$ for all $y \in A\}$. Therefore, for all $A \in Y$, $\{x \in A : x \succeq y$ for all $y \in A\}$ is the same as the observed set of options chosen by the agent from A.

Thus, the observed choices of the agent are rationalizable in terms of the ordering $\succeq$ in the standard economic framework.

(2.iii) Suppose an agent's observed choices are rational in the sense of standard economic theory. Then, there exists a preference ordering $\succeq$ over X such that, for all $A \in Y$, the observed set of options chosen by the agent from A coincides with the set $\{x \in A : x \succeq y$ for all $y \in A\}$. Define a binary relation R over S as follows: for all (x,o), $(y,o') \in S$, $(x,o)R(y,o')$ iff $x \succeq y$. Clearly, R is well-defined. Since $\succeq$ is reflexive, R is reflexive. Similarly, the connectedness

and transitivity of R follow from the connectedness and transitivity of $\succeq$. It is clear that R is context-independent (I). Then, by Proposition (1.i), R is context-independent (II), and, noting reflexivity of R, by Proposition (1.i) again, R is context-dependent (III). Finally, note that, by the construction of R, for all $A \in Y$ and all $o \in O(A)$, $\{x \in A : (x, o)R(y, o)$ for all $(y, o) \in A \times \{o\}\}$ coincides with $\{x \in A : x \succeq y$ for all $y \in A\}$, which, in turn, coincides with the observed set of options chosen by the agent from A.

Thus, the agent's observed choices are rationalizable in terms of the ordering R over S in the context-inclusive framework and R is context-independent in each of the three senses under consideration. ■

Remark 3 Proposition 2 formally confirms that the following two concepts can be thought of as being exact counterparts of each other: (1) the concept of rational choice in standard economic theory; and (2) the concept, in the context-inclusive framework, of observed choices being rationalizable in terms of an ordering R, which is defined over S and which satisfies any of the three types of context-independence introduced earlier in this paper.

Remark 4 Should context-dependence of R in the context-inclusive framework be regarded as an indication of irrationality? If the term "rationality" is being used in a normative fashion,[3] then we do not find anything particularly irrational in the context-dependence of R in Examples 1(a), 1(b), 2, and 3. We do not, however, rule out the possibility that some other specific manifestations of context-dependence of R may be judged "irrational". Thus, if a person living in Los Angeles and ordering her lunch to be eaten in Los Angeles, prefers ice cream to a cup of chicken noodle soup when it is hot in Delhi and prefers a cup of chicken noodle soup to ice cream when it is cold in Delhi, the person's preferences can be modeled as context-dependent preference, the context being the specific weather prevailing in Delhi (cf. Example 3 above). But, in the absence of any known connection between the weather in Delhi and the state of affairs in Los Angeles, we believe that such context-dependence will be considered irrational by most people. Therefore, it seems to us that, in discussing the normative concept of rational choice, there is some advantage in using the context-inclusive framework, which permits context-dependence of R and allows us to discuss which forms of context-dependence may or may not be "reasonable". The question may arise whether one can lay down some general criteria to judge whether a particular manifestation of context-dependence is "unreasonable". Saying that certain manifestations of context-dependent preferences are reasonable or unreasonable involves a value judgment and value judgments are not beyond the boundaries of systematic reasoning.[4] But, though one can outline the formal structure of the type of reasoning that can be used to question or support value judgments, the intuitive content of such reasoning has to depend on the

[3] In the following section, we elaborate the distinction between normative and non-normative uses of the term "rationality".

[4] See, among others, Hare [11, 12], Sen [18], and Pattanaik [15, Chap. 2].

specific value judgment under consideration. Thus, it would seem difficult, if not impossible, to formulate *a priori* standards for judging which types of context-dependent preferences are unreasonable.

Remark 5 Note that, in the context-inclusive framework, if O has exactly one element and the agent's observed choices are rationalizable in terms of an ordering R over S, then the agent's choices are rational in the standard economic framework. This is because that, when $O = \{o\}$ and R is an ordering, R is trivially context-independent (I), (II), and (III). Therefore, by Proposition 2, the agent's observed choices are rational in the standard economic framework.

4 Discussion

In this paper, we have explored a component of the standard economic conception of a rational agent and rational choices. This component requires that for an agent to be rational, she should have a fixed weak-preference relation over the universal set of all possible options and she should make her choices from different sets of feasible options on the basis of this fixed weak preference relation over the universal set of options. This, of course, implies that the preferences of the agent should be independent of the context in which she makes her choice. Over the last few decades, the literature on choice and preference has highlighted many instances where an individual's preferences are context-dependent. We have discussed some of these examples. What are their implications for the standard theory of rational choice? To see this, it will be useful to distinguish between two distinct ways in which the theory of rational choice in economics can be viewed. First, it can be viewed as an attempt to articulate what we intuitively mean when we say that an agent is rational or is choosing options in a rational fashion. Viewed in this way, the concept of a rational agent has normative content: a person whose choices in different choice situations are not based on a fixed preference ordering defined over the universal set of options is then regarded as falling short of fulfilling our intuitive standards for judging rationality. One can, however, take a second, and alternative, view of the theory of rational choice in economics. From this perspective, the term "choosing rationally" is simply a technical shorthand expression for "making choices in different choice situations on the basis of a fixed preference ordering defined over the universal set of options", and it does not seek to articulate any deep intuition about rationality. Under this interpretation, economists' theory of rational choice can be viewed as an exploration of the implications of the empirically testable hypothesis that an agent always makes her choices in different choice situations on the basis of a fixed preference ordering defined over the universal set of options (this central hypothesis, however, is typically combined with other empirical hypotheses to construct a theoretical model).

Irrespective of which of the two interpretations of the economists' theory of rational choice we adopt, the concept of context-independent preferences constitutes

an integral part of the notion of rational choice in this theory. Earlier, we have discussed various examples of context-dependent preferences. Our reaction to these examples will depend on our specific interpretation of the economic theory of rational choice. Consider first the case where we interpret the definition of rational choice as an attempt to articulate our intuitive notion regarding the "rationality" of choices so that the definition is taken to have a normative significance. Given this interpretation of the definition of rational choice, it seems reasonable to maintain that there is nothing irrational about the choices of the agents in our Examples 1(a), 1(b), 2, and 3. Thus, the agent in Example 1(a) can be considered to be choosing rationally, despite his context-dependent preferences, since he has good reasons for choosing an orange over an apple in one context and choosing an apple over an orange in a different context. In the case of this example, we feel that: (1) if the options are specified simply as alternative fruits, then there are good reasons not to insist on treating context-independence of preferences as a condition for rational choice of options specified in this fashion; and (2) if, at all, we want to have context-independence as a condition for rational choice of options, we should redefine the options so that the redefined options are "picking up an apple in a dinner party when the apple happens to be the only apple in the available fruit basket" , "picking up the apple in a dinner party when there are several apples in the available fruit basket" , and so on. In contrast, while, in the example discussed in Remark 4 (where the choice between chicken noodle soup and ice cream in Los Angeles is contingent on the weather in Delhi), one can model the agent's preferences as being context-dependent, such context-dependence and the resultant choices would seem irrational, in a normative sense of the word "irrational", to most people. Now consider the second interpretation of the theory of rational choice which is devoid of normative content and under which saying that all agents make their choices rationally amounts to an empirical hypothesis that every agent makes her choices in different choice situations on the basis of a fixed preference ordering defined over a universal set of options. In this case, manifestations of context-dependent preferences, including those in the examples that we have considered in this paper, falsify that hypothesis given the specification of options that the theorist uses in his model. Faced with such falsification, the theorist may decide to take any one of the following routes:

(a) The theorist may re-specify the options so as to incorporate in the new specification of options certain features of what was called a context under the original specification of options (the phenomenon of context-dependent preferences that arose given the original specification of options, may disappear after such re-specification).
(b) The theorist may retain her original specification of options and continue to work with the falsified empirical hypothesis of rational choice by agents on the ground that, though, given her specification of the options, the hypothesis is contradicted by some empirical evidence, the hypothesis is still useful since it "works most of the time" or at least "works most of the time for the class of choice problems under consideration".

(c) The theorist may retain her specification of options, abandon altogether the empirical hypothesis of rational choice (at least for the type of choices under consideration), and proceed to explore the implications of alternative models which permit context-dependent preferences.

We believe that, when, despite some evidence of context-dependent preferences, most economists continue to retain the standard theory of rational choice, they are really opting for route (b) or (c) above.

Acknowledgements We are grateful to the referees and Miriam Teschl for helpful comments.

References

1. Baigent N (2007) Choice, norms and revealed preference. Analyse Kritik 29:139–145
2. Baigent N, Gaertner W (1996) Never choose the uniquely largest: a characterization. Econ Theory 8:239–249
3. Bandyopadhyay T, Dasgupta I, Pattanaik PK (1999) Stochastic revealed preference. J Econ Theory 84:95–110
4. Bandyopadhyay T, Dasgupta I, Pattanaik PK (2004) A general revealed preference theorem for stochastic demand behavior. Econ Theory 23:589–599
5. Bhattacharyya A, Pattanaik PK, Xu Y (2011) Choice, internal consistency and rationality. Econ Philos 27:123–149
6. Bossert W, Suzumura K (2009) External norms and rationality of choice. Econ Philos 25:139–152
7. Bossert W, Suzumura K (2011) Rationality, external norms and the epistemic value of menus. Soc Choice Welf 37:729–741
8. Gaertner W, Xu Y (1999) On the structure of choice under different external references. Econ Theory 14:609–620
9. Gaertner W, Xu Y (2004) Procedural choice. Econ Theory 24:335–349
10. Gaertner W, Xu Y (2009) Individual choices in a non-consequentialist framework: a procedural approach. In: Basu K, Kanbur R (eds) Arguments for a better world: essays in honor of Amartya Sen, vol I: ethics, welfare and measurement, Chap. 9. Oxford University Press, Oxford
11. Hare RM (1964) The language of morals. Oxford Paperback edition. Oxford University Press, Oxford
12. Hare RM (1965) Freedom and reason. Oxford Paperback edition. Oxford University Press, Oxford
13. Hausman D (2012) Preference, value, choice and welfare. Cambridge University Press, Cambridge
14. Luce RD, Raiffa H (1957) Games and decisions. Wiley, New York
15. Pattanaik PK (1971) Voting and collective choice. Cambridge University Press, Cambridge
16. Savage LJ (1972) The foundations of statistics, 2nd revised edn. Dover Publications, New York
17. Sen AK (1993) Internal consistency of choice. Econometrica 61:495–521
18. Sen AK (1967) The nature and classes of prescriptive judgments. Philol Q 17:46–62
19. Xu Y (2007) Norm-constrained choices. Analyses Kritik 29:329–339

A Primer on Economic Choice Automata

Mark R. Johnson

Abstract This paper presents a development of the transformation semigroup of economic choice automata as a subgroup of the semigroup (monoid) of partial functions defined over the states of a finite state machine. The classes of consistency behavior considered are those rationalized by linear orders, weak orders, quasi-transitive relations and non-rationalizable path independent choice functions. For each of these classes of choice behavior, a particular class of lattice is identified as the action semigroup that drives the automaton. Given these characterizations, several features of the choice behavior are considered. In particular, the simplifying interval property of path independent choice, the importance of the distributive property of quasi-transitive rational choice in reducing the complexity of dynamic choice is addressed. Based on the algebraic structure of semiautomata implementing path independent choice functions it is possible to rank these semiautomata by the mathematical power required to implement a particular class of choice functions. This provides a means for ranking these machines by their "implementation complexity". Dually, the computational complexity of constructing a semiautomaton that implements a particular class of choice functions is investigated. It is seen that these complexities are inversely related.

Keywords Automata • Choice functions • Computational complexity • Implementation complexity • Semiautomata

1 Introduction

Notions of complexity lie at the core of our thinking about economic decision making. Generally, there is a belief that differences in complexity will affect the behavior of individuals or the structure of economic institutions. For example, in an elementary, unchanging world, a simple rule of thumb may be an adequate and appropriate method for making decisions. In a more complicated, rapidly changing

M.R. Johnson (✉)
Department of Finance and Economics, A. B. Freeman School of Business, Tulane University, New Orleans, LA 70118, USA
e-mail: mjohnso@wave.tulane.edu

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*, Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_5

world, the same simple rule of thumb may be quickly over come and lead to mistakes. Similarly, an overly complicated corporate decision structure involving many people and elaborate departmental check-offs might be too expensive and reduce the firm profits. The discussion below identifies the limitations imposed on the complexity of choice by economic consistency axioms.

Simon was one of the first economists explicitly to incorporate constraints on the information processing capacities of individuals and firms. For Simon [56], there were several ways that information processing limitations could affect economic activity and he categorized models incorporating these limitations as "theories of bounded rationality". While Simon was one of the early articulators on the importance of information processing to economic behavior and structure, other economists have appealed to these arguments as well. Most have adhered to Simon's view that different economic structures might have different complexities, and their writings have reflected a shared belief that complexities can either be scaled or, at least, compared. Among other early examples of information processing considerations in the model of individual choice is Strotz's [57, 58] exchange with Gorman [25] on the separability property for utility functions. In that exchange, Strotz maintained that separability was desirable because it simplified the consumer budgeting problem. A more recent example of information processing impacts on economic models is Auman's [4] suggestion that players in a game be modeled as automata and that the complexity measure based on the number of states in the automaton required to implement each strategy be used to separate classes of strategies.[1] Rubinstein [54] was one of many to follow up on this suggestion. With respect to organizational structure, Radner [53] has suggested that information processing is a major part of a corporation's "management" activities and that the hierarchical structure of a firm arises, at least in part because of this structure's efficiency at processing decentralized information.[2] Other authors also have made links between collective decision-making structures and complexity issues.[3]

A number of authors have suggested that one way information processing costs affect economic behavior is through the consistency axioms either satisfied by or

[1] An early survey of the literature growing out of this suggestion is offered in Kalai [40] while Chatterjee and Sabourian [15] offer a more recent treatment.

[2] Radner [53] also suggested that "the costliness of information processing contributes to organizational economies or diseconomies of scale". This idea that there are information costs contributing to the operational costs of a decision making organizations is very similar to Hurwicz's [30] suggestion of a resource cost to allocation mechanisms.

[3] For example, Bartholdi and Orlin [7], Bartholdi et al. [8, 9] address matters of computational complexity and the complexity of strategic manipulation in voting schemes. Johnson [31] references the relationship between the distribution of power and notions of choice complexity in Arrow type choice procedures.

imposed on the choices made by the individual or institution being modeled.[4] In particular both Arrow [2] and Plott [52] envisioned choice as an algorithm or process in which the elements of the feasible set were examined and a choice set determined.[5] Plott specifically suggested that there was a "computational efficiency" aspect to consistency axioms in general and to his path independence axiom in particular. Especially relevant to the results presented here, Plott provided a link between computing machines and path independent choice functions by proving that path independent choice functions form a semigroup under a naturally defined operation. The significance of this result is that, every semigroup can be used to define a semiautomaton (the basic building block of computing machines) and every semiautomaton has an associated semigroup.[6] More recently, Johnson [32, 34] has suggested that the link between the computational efficiency of a choice function and the consistency axiom it satisfies is captured by the algebraic structure of a subsemigroup of the semigroup originally identified by Plott.

Because the consistency axioms impose limitations on the relationship between the choice made on one set and that made on other sets, Plott's suggestion is intuitively appealing. Especially in the case of the most commonly used consistency requirements (i.e., rationalizability by either linear orders or weak orders and quasi transitivity for his path independence axiom), Plott observed that consistency axioms removed the requirement of re-examining previously considered alternatives. For the most commonly used axioms, it is the case, also, that the classes of choice functions satisfying these axioms form a nested set. Thus, the expectation is that, if consistency axioms are related to information processing efficiencies, then the choice function complexity should grow as the consistency axioms are relaxed. This intuition provides the same complexity ranking as the ordering by algebraic structure suggested by Johnson [31].

In these early discussions there, often, was an ambiguity in the use of the term "complexity" with imprecision in distinguishing between what information

[4]Virtually all economic models of individual choice assume that the individual satisfies the Weak Axiom of Revealed Preference. Often, as in the case of theorems following out of Arrow's general Possibilities Theorem [3], group decision structures are assumed to satisfy some type of consistency axiom. Many of these structures are surveyed in Kelly [42].

[5]Several other authors have formalized the search/process aspects of choice, which Arrow and Plott left somewhat unspecified, as an algorithm and investigated the relationship between these algorithms and the act of choice. Of particular note are Campbell's [13, 14] work on the existence of desirable algorithms for computing choice functions, Kelly's [43] look at the computability of collective choice rules and Bandyopadhyay's [5] characterization of a class of choice functions by means of a specific algorithm. Also contributing to this line of research are Lewis's [49, 50] investigations into the ability of a Turing machine (see Turing [59]) to compute economic choice automata on non-finite sets.

[6]More precisely, the link is between transformation semigroups and semiautomata. The transformation semigroup representation of a semiautomaton is discussed in Sect. 2.3. The relationship between transformation semigroups and semiautomata is discussed in Holcombe [28, pp. 31–34]. Holcombe [28, p. 145] also has a discussion of the relationship between semiautomata and automata.

scientists call "implementation complexity" and "computational complexity". The difference between these two complexities is that implementation complexity is determined by the difficulty of making an already determined choice. That is, given a choice function defined on a finite set, how much "computing power" is required by a semiautomaton to make choices consistent with the already specified choice function. Johnson [31] presents an intuitive example of this aspect of choice "difficulty". There, the difficulty of implementing a choice function is scaled by the number of entries in the incidence matrix representing the binary relation that rationalizes the choice function that are required to be known in order to assure the correct choice is made as the feasible set is expanded. There, this difficulty was called the "bit cost" of making choice. This implementation complexity is contrasted with "computational complexity" which reflects the difficulty of making the choice function. Until recently this "computational complexity" has arisen most explicitly in game theory.

In game theoretic discussions, the complexity of computing the best response (e.g., [10, 24, 45, 51]) has been substantially discussed. In these papers, the computational complexity is treated in the classical manner; usually some tool such as a time or memory space bound on determining a solution. Typically each of these is a computational complexity measure satisfying what are known as the Blum axioms [12].[7] One attraction of a computational complexity measure satisfying the Blum axioms is that the resulting scaling is independent of the machine performing the calculations. Similarly, the computational complexity measure introduced in Sect. 4 also satisfies the Blum axioms.

In contrast, implementation complexity has been interpreted in a number of disparate manners. Most commonly the scaling of implementation complexity is by the number of states in the semiautomaton implementing the desired strategy. The number-of-states measure of implementation complexity has some appeal in that it is numerical, simple to use, and as a result, it is easy to obtain results. However, even Abreu and Rubinstein [1] note limitations to their use of the number-of-states measure;

> Various features of the model presented below, such as the complexity measure we use, are rather special. Our results are therefore regarded as suggestive, and we strongly emphasize the exploratory nature of the present paper.

Kalai and Stanford [41] amend the basic number-of-states measure to reflect the fact that a semiautomaton can have extraneous states and appeal to the minimum number-of-states machine implementing a strategy. Banks and Sundaram [6] tried to capture the fact that there is more to machine complexity than the number of states in the machine by using a criterion that depends on both the number of states as well as the number of edges in the directed graph representation of the machine. Johnson [33] pointed out that, on machines with two states and two transitions,

[7] The reader is referred directly to Blum's classic paper for a discussion of the issues and techniques. The paper is short and readable.

there are, up to isomorphism, four distinct semiautomata and that these machines are categorized as (1) a machine that can't count, and can't detect differences in order of play, (2) a machine that can count but can't detect differences in the order of play, (3) a machine that can't count but can detect differences in the order of play, and, finally (4) a machine that can count *and* detect differences in the order of play. Johnson obtained these observations by using algebraic techniques similar to those employed in the results presented here.

Within algebraic complexity, the Krohn-Rhodes [47, 48] measure is commonly used.[8] This measure is based on the minimum number of simple groups in a wreath product that forms a semigroup covering of the semigroup generated by the machine of interest. Gottinger [26] in particular has suggested using the Krohn-Rhodes measure for some economic applications. In applications, however, there is a general problem in that it is not known how to determine what is the minimum number of simple groups in the decomposition of a particular semigroup. For specific application to choice functions, the measure is not very useful because there are no groups in the semigroup associated with the machines implementing path independent choice functions and, as a result, the Krohn-Rhodes measure would give an implementation complexity of zero to all path independent choice function implementing semiautomata. In the following, the power required to implement a choice function is ordered by the algebraic class of the machine implementing the choice function. Because these algebras are nested, comparison of the power of the different systems is straightforward.

From an economic perspective, one interpretation of these different complexities is that the computational complexity can be viewed as a fixed cost. This is the cost of determining or constructing the choice machine that will be used to implement choices. The computational complexity cost is born only once when the machine is made. This interpretation is consistent with classical views on computational complexity and with the way computational complexity is used in game theory. In contrast, the implementation complexity is more like a marginal cost. In order to implement a particular choice function, you need to maintain a machine with the requisite power to implement the choice function. This intuition is similar to the justification Rubinstein [54] gave in support of the number of state based measure of implementation complexity he used in his introduction of complexity costs into repeated play games.[9]

The first results presented here identify the structure of semiautomata constrained to satisfy Path Independence. The key requirement is to demonstrate that a particular semigroup is a subsemigroup of the semigroup of partial functions defined on the feasible sets. The particular subsemigroup also is a the subsemigroup of the semigroup identified by Plott in his first demonstration of the link between path independent choice functions and semigroups. In the past, Johnson [34] has

[8] See Krohn-Rhodes [47, 48] directly or Eilenberg [20, 21] for a more modern treatment.

[9] While not investigated here, this intuition naturally raises the possibility of trade offs between fixed costs and marginal costs in the design of choice machines/institutions.

demonstrated that choice semiautomata can be constructed by means of Eilenberg's embedding theorem. Demonstrating that the required semigroup already is a subsemigroup of the semigroup of partial functions, both simplifies the presentation and tightens the link between path independent choice functions and automata. Given this, it can be seen that choice semiautomata have elementary structure. For example, no "memory" is required to implement path independent choice.

Subsequent results identify a complexity ordering of path independent choice functions defined on finite sets. The classes of choice functions considered are those satisfying the Strong Axiom of Preference (SAP), the Weak Axiom of Revealed Preference (WARP), Quasitransitive rational Path Independence and (not necessarily rational) Path Independence. Each of the classes of choice functions are ordered by the mathematical power of the system implied by the defining consistency axiom. These comparisons confirm Plott's conjecture that consistency axioms contain information processing implications as well as conforming to Johnson's [31] ranking of choice function complexity by algebraic structure.

In order to focus the presentation a number of simplifications are made. For example, it is assumed that the choice functions are complete. This simplification will imply that the resulting automata, also are complete. In fact, the semiautomata model is perfectly capable of, and, in many ways, ideally suited for dealing with situations where completeness is not present but the exposition of the choice function implementing automata is simpler in the complete case. In addition, in the complete case, a natural nesting arises that eases discussion of implementation complexity rankings. If completeness is not assumed, then, depending on the precise nature of the incompleteness, the nesting of systems seen in this presentation may or may not be visible.[10] Another example of a simplification is that, while it is necessary to implement choice both as the feasible set *expands* as well as when it *contracts*, this presentation details only those machines that implement choice as the feasible set expands.[11]

Following specification of choice semiautomata and identifying the implementation complexity for each class, the matter of "computational complexity" is addressed. In this treatment a very simple approach is adopted. First, a result is presented that allows construction of all path independent choice functions on a finite set by means of a series of contractions.[12] This result provides the intuition for a partial order of the computational complexity of different choice functions.

[10] As a particular example, one of the referees inquired about restricting choice to sets that have null intersection. For the Path Independent choice functions considered, the mathematical structure that arises is a particular class of lattice. In a finite lattice, the meet and join of any two elements in the lattice must be well defined. If a choice operation from sets that have a non-empty intersection were not allowed, the lattice structure exploited here may not obtain.

[11] Choice functions that implement choice as the feasible set both expands and contracts are addressed in Johnson [34].

[12] I thank my co-author Richard A. Dean for permission to use our previously unpublished, original proof of this result. The appeal of this proof is that it is based on traditional lattice theoretic tools and, as a result, some may find it more accessible that other proofs.

Loosely, the scaling for the computational complexity is related to the number of elements removed from a particular feasible set in determining its choice set. This interpretation arises because each contraction deletes a single alternative as a possible choice element from a particular feasible set. It is seen that it is possible to order the classes of choice functions by their computational complexity. And, as noted earlier, the method satisfies the Blum axioms.

Most intriguingly, in a very natural manner, it turns out that the implementation complexity and the computational complexity are dual in the sense that classes of choice functions that have higher implementation complexity have lower computational complexity. Because of the number and range of different choice functions in each class, there is some overlap in the computational complexities of specific choice functions but, for the most computationally intensive and least computationally intensive representatives of each class, the duality holds.[13]

Section 2 presents the basic definitions and tools. The semiautomaton model and the transition to path independent choice function implementing machines are summarized in Sect. 3. A principle focus is in demonstrating that the mappings for the choice semiautomaton form a subsemigroup of the semigroup of partial functions among the possible subsets. This section also reports on the results for implementation complexity. Section 4 presents the results on computational complexity. Conclusions are in Sect. 5 and the original proof of the contraction theorem is provided in the Appendix.

2 Definitions and Notation

The technical tools used in the results presented below are choice functions, the elementary algebra of sets, primarily semigroups and lattices, and the transformation semigroup representation of a semiautomaton. The definitions and prerequisites of choice functions and consistency requirements on choice functions are presented in Sect. 2.1. Algebras are covered in Sect. 2.2. Semiautomata and the links to algebra are covered in Sect. 2.3.

2.1 Notation, Choice Functions and Consistency Requirements

The universal set V is composed of a finite number of distinct alternatives, and 2^V is the power set of V. Subsets of V, denoted by v, are elements of 2^V. Unless otherwise stated, the cardinality of V, denoted by $|V|$, is t, and the cardinality of $v \in 2^V$ is n; note $n \leq t$. Distinct subsets of V are subscripted with an integer $i \in \{1, \ldots 2^t\}$; where $\{v_i\} = \{v_j\}$ if and only if $i = j$.

[13] Previous results existed only for the least computationally intensive representative of each class.

A *choice function* is a mapping $C : 2^V \to 2^V$, such that $C(v) \subseteq v$ and $C(v) = \emptyset$ if and only if $v = \emptyset$. A choice function C defined on V is *discriminating* if there is some $v \in 2^V$ for which $C(v) \neq v$. A choice function C is *rational* if and only if there exists a relation R such that, for every $v \in 2^V$, $C(v) = G(v; R)$ where, $G(v; R) = \{x \in v | xRy, \forall y \in v\}$. The function $G(v; R)$ selects the R-maximal elements.

The classes of choice functions considered are those satisfying the Strong Axiom of Preference, the Weak axiom of Revealed Preference, the conjunction of Path Independence and Extension, and Path Independence (alone). The consistency axioms are defined formally as follows.

Strong Axiom of Preference (SAP):

(i) $\forall x, y \in V, x \in C(\{x, y\}) \Rightarrow y \notin C(\{x, y\})$, and

(ii) $\forall v_1, v_2 \subseteq V, v_1 \subseteq v_2 \Rightarrow \{v_1 \cap C(v_2)\} = \begin{cases} \emptyset, \text{or} \\ C(v_1) \end{cases}$.

Weak Axiom of Revealed Preference (WARP)

$\forall v_1, v_2 \subseteq V, v_1 \subseteq v_2 \Rightarrow \{v_1 \cap C(v_2)\} = \begin{cases} \emptyset, \text{or} \\ C(v_1) \end{cases}$.

Rational Path Independence (RPI)

(i) $\forall v_1, v_2 \subseteq V, C(C(v_1) \cup C(v_2)) = C(v_1 \cup v_2)$, and

(ii) Extension (E): $\forall v \subseteq V, (x \in v \text{ and } (\forall y_{y \in v}, x \in C(\{x, y\})) \Rightarrow x \in C(v))$.

Path Independence (PI)

$\forall v_1, v_2 \subseteq V, C(C(v_1) \cup C(v_2)) = C(v_1 \cup v_2)$.

The first two of these classes always can be rationalized by a complete, reflexive and transitive binary relation [2]. Choice functions satisfying the Strong Axiom are always single-valued and rationalized by linear orders while choice functions meeting WARP need not be single-valued and are rationalized by weak orders. The two classes of path independent choice functions are distinguished by whether or not they are rationalizable; choice functions satisfying both PI and E are rationalizable by a quasitransitive relation while choice functions satisfying PI need not be rationalizable [52].[14]

2.2 Algebras

The definitions of binary systems and system properties are provided in terms of an arbitrary finite non-empty set N, which is used as both the domain and the range,

[14] A complete, reflexive relation R, where the strict preference part is denoted by P, is *quasitransitive* if for all $x, y, z \in V, xPy$ and $yPz \to xPz$. Thus strict preference is transitive while indifference need not be.

and a binary operation denoted by $(\cdot)$. Thus, $\cdot : N \times N \rightarrow N$, and the binary system for N under the operation $(\cdot)$ is $\langle N; \cdot \rangle$. Algebraic properties defined for all $v_i, v_j, v_k \in N$ are

(B-1) Closure: $v_i \cdot v_j \in N$,
(B-2) Associativity: $v_i \cdot (v_j \cdot v_k) = (v_i \cdot v_j) \cdot v_k$,
(B-3) Commutativity: $v_i \cdot v_j = v_j \cdot v_i$, and
(B-4) Idempotence: $v_i \cdot v_i = v_i$.

A binary system satisfying (B-1) and (B-2) is called a *semigroup*, and a semigroup satisfying (B-3) is called a *commutative semigroup*. A semigroup for which every element satisfies (B-4) is called an idempotent semigroup. A commutative idempotent semigroup is a *semilattice*. Conceptually, some may find it easier to consider these semilattices diagrammatically under the natural partial ordering of the semigroup where the *natural partial ordering* is defined as follows: $a \cdot b = b \Leftrightarrow a \leq b$.[15]

For application to choice functions, the power set of the universal set V is used as both the domain and the range, and the binary operation $(\bullet)$ is adopted from Plott [52].

Definition 1 Given a path independent choice function C, the *Plott Product* $(\bullet)$ is defined as follows, $\bullet : 2^V \times 2^V \rightarrow 2^V$, where $\forall v_1, v_2 \in 2^V, v_1 \bullet v_2 = C(C(v_1) \cup C(v_2))$.

Formally, the operation $(\bullet)$ should be subscripted by the choice function used in its definition, however, to avoid excess notation, this subscripting is omitted where the choice function can be inferred from the context. The binary system for V under the operation $(\bullet)$ is denoted by $\langle 2^V; \bullet \rangle$. Plott [52] proved that this system is a commutative semigroup.

In addition to the properties of the operation, it is useful to identify two special members of binary systems.

Definition 2 Given a binary system $T = \langle N; \cdot \rangle$ an element z such that $x \cdot z = z \cdot x = z, \forall x \in T$ is called a *zero*, and an element e such that $t \cdot e = e \cdot t = t, \forall t \in T$ is called an *identity*.

A semigroup that has an identity is a *monoid*. An idempotent commutative monoid with a zero is a *lattice*. Johnson [32] identified a subsemigroup of Plott's semigroup that has precisely these properties. Further, Johnson conjectured that this subsemigroup might be relevant to economic applications of automata theory. The results below validate that conjecture. While initially identified by means of Plott's single operation $(\bullet)$, lattices actually have two operations, typically called the *join* denoted by $\vee$ and the *meet* denoted by $\wedge$. A lattice L is denoted by $\langle L; \vee, \wedge \rangle$. A lattice L has a *dual* denoted by D obtained by interchanging the roles of the meet and join operations so that if $a \vee b = a$ in L then $a \vee b = b$ in D. Given a

[15] The natural partial ordering is adopted from Clifford and Preston [16, 17].

lattice $L = \langle L; \wedge, \vee \rangle$ for which there is a set K such that $\emptyset \neq K \subseteq L$ and $a, b \in K$ implies $a \wedge b \in K$ and $a \vee b \in K$ then $K = \langle K; \wedge, \vee \rangle$ is a *sublattice* of L. An element x in a lattice is called *join-irreducible* if $a \vee b = x$ implies $x = a$ or $x = b$. By convention, bottom elements of a lattice are not called join-irreducible. Dually, an element y in a lattice is called *meet-irreducible* if $a \wedge b = y$ implies $y = a$ or $y = b$. For both the join irreducibles and the meet irreducibles, the partially ordered set of irreducibles $\langle P; \leqslant \rangle$ may be important. In a partially ordered set P, x *covers* y if $x > y$ and for no $a \in P$, $x > a > y$. Lattices are well covered in such classics as Birkhoff [11] and Davey and Priestly [18]; however, a few especially useful properties are summarized here. One important property of some lattices is the distributive law. A lattice $\langle L; \vee, \wedge \rangle$ is a *distributive lattice* if it satisfies the *distributive law*:

$$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c) \text{ for all } (a, b, c \in L).$$

Given a lattice L with a zero 0 and an identity 1, and some element $a \in L$, for which there is an element $b \in L$ such that $a \wedge b = 0$ and $a \vee b = 1$, then a is said to have a *compliment*. If a has a unique compliment, then the compliment is denoted by a'. Taking the compliment is a *unary* operation. A *Boolean algebra* is a system $\langle B; \wedge, \vee, ', 0, 1 \rangle$ such that (1) $\langle B, \wedge, \vee \rangle$ is a distributive lattice, (2) $a \wedge 1 = a$ and $a \vee 0 = a$ for all $a \in B$, and (3) $a \wedge a' = 0$ and $a \vee a' = 1$ for all $a \in B$. The finite Boolean algebras considered here are isomorphic to 2^V under set union, intersection and complementation.

Within the Boolean algebra 2^V for sets $V \supseteq T \supseteq B$, the collection of sets K such that $T \supseteq K \supseteq B$ is called an *interval*, and the interval is denoted by T/B. T is the *top* of the interval, and B is the *bottom* of the interval. An interval T/B is called *proper* if $T \neq B$. If $T = B \cup \{x\}$, then T *covers* B and the interval T/B is a *prime* interval.[16] It will turn out that intervals are significant in path independent choice functions. In fact, little else is required in addition to the interval property in order to define a path independent choice function.[17] Two examples of the presence of intervals in the domain of the choice function being mapped into a particular choice element are presented in Fig. 1.

A particular class of lattices initially identified by Dilworth [19] and now known as *lower locally distributive lattices* (LLDs) is relevant for choice functions. A lattice is an LLD if every element in the lattice has a unique irredundant representation as the join of join irreducibles. Here a representation of an element a in a lattice as the *irredundant* join of join irreducibles means that if $a = x_1 \vee x_2 \vee \ldots \vee x_k$ then a is not the join of any proper subset of $\{x_1, \ldots, x_k\}$. This representation also is *unique* in the sense that if $a = y_1 \vee y_2 \vee \ldots \vee y_h$ as well as

[16] Birkhoff attributes this use of the term prime interval to Morgan Ward.

[17] See Johnson and Dean [39] for the characterization of path independent choice functions result using partitions of the domain satisfying the interval property and little else.

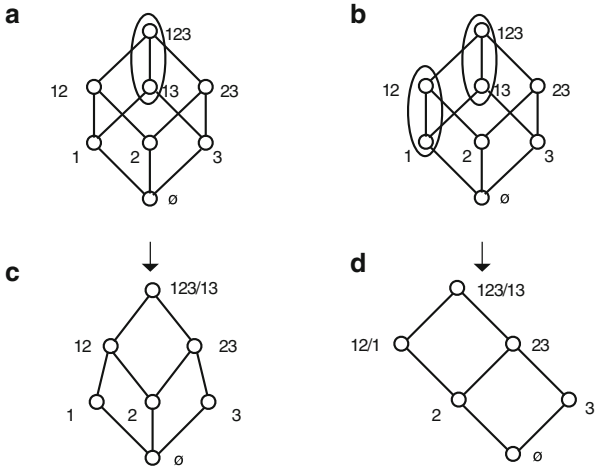


Fig. 1 Demonstration of relationship between identified intervals and lattice representations of Path Independent choice functions; (**a**) Boolean algebra with one interval identified, (**b**) Boolean algebra with two intervals identified, (**c**) image lattice of (**a**) with interval 123/13 identified, and (**d**) image lattice of (**b**) with intervals 123/13 and 12/1 identified

$a = x_1 \vee x_2 \vee \ldots \vee x_k$ then $h = k$ and $\{x_1, \ldots, x_k\} = \{y_1, \ldots, y_h\}$.[18] Johnson and Dean [35, 36] and, independently, Koshevoy [46] demonstrated a direct link between path independent choice functions and LLDs in that every PI choice function has a representation as an LLD lattice and for every LLD lattice, there is an associated PI choice function.[19] Further, Johnson and Dean demonstrated characterization results between the predominant classes of PI choice functions and subclasses of LLDs. Significantly, not all of these LLDs are distributive.

2.3 *Semiautomata, Transformation Semigroups and Action*

Although employed here only as a link to other literature, a common means for representing a semiautomaton, or finite state machine, in economics is through the directed graph. A directed graph representation of a semiautomaton $\mathscr{M} = (Q, \Sigma, F)$ consists of a finite number of states Q, an alphabet Σ and partial functions F. The partial functions F are the transitions so that $F : Q \times \Sigma \to Q$. If the partial functions F are functions then the semiautomaton is a *complete*

[18]The partition lattice on five elements is an example of a lattice that fails to meet this requirement.

[19]Koshevoy [46] used convex geometries to obtain results related to a subset of the Johnson and Dean [35, 36] results. Here the full range of the Johnson and Dean characterizations is used.

semiautomaton. In the directed graph, the states become the vertices, the partial functions F are the edges, and the alphabet labels the edges.[20]

Within information science and mathematics, a standard alternative to the directed graph representation is the transformation semigroup. This fundamental semigroup of algebraic automata theory consists of an underlying set and an action semigroup (see Eilenberg [20, 21] or Holcombe [28]). The transformation semigroup and its component parts are defined as follows.[21]

Let Q be a finite set, and let $PF(Q)$ be the monoid of partial functions $Q \to Q$ with composition of partial functions as multiplication. The identity partial function is the unit denoted by 1_Q. (In this paper, the situation is simplified because the mappings considered are functions.) A *transformation semigroup* $X = (Q, S)$ is a finite set Q and S is a subsemigroup of $PF(Q)$. The set Q is called the *underlying set* of X, and the members of Q are called *states*. The semigroup S is called the *action semigroup* of X, and the elements of S are called the *transformations* of X.

The transformation semigroup X is *complete* if the following two conditions are met.

(a) $Q \neq \emptyset$
(b) $qs \neq \emptyset$ for all $q \in Q, s \in S$.

Condition (a) assures that the underlying set is not empty and condition (b) requires that each transformation be defined at every state. Thus, as with the directed graph representation, the transformation semigroup representation is complete if the transformations of X are functions.

Both the transformation semigroup and the action semigroup are important items in the study of automata theory. However, while the transformation semigroups characterize semiautomata, all of the mathematical power or algebraic complexity is contained in the action semigroup [21]. For this reason, much of the remaining analysis is focused on the action semigroup.

3 From Generic Semiautomata to Choice Semiautomata

As described above, semiautomata can be represented either as a directed graph or as a transformation semigroup. Two small examples of the directed graph representations are depicted in Fig. 2. In the left most example (Fig. 2a), there are three states and two letters in the alphabet labeling transitions among the states. The transitions labeled "a" form a cycle among the three states while the transitions labeled "b" flip the triangle about the state 1. The algebra associated with the

[20] For perspective, the more commonly employed automaton is a semiautomaton that has been augmented by identification of an *initial* state i and a collection of *terminal* states T. Thus an automaton $\mathscr{A} = (\mathscr{M}, i, T)$.

[21] This summary borrows from Eilenberg [20, 21].

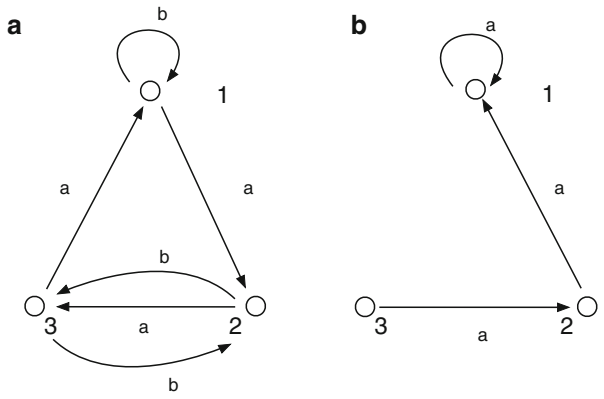


Fig. 2 Example of two three-state semiautomata, (**a**) on the *left* with two-letter alphabet labeling transitions, and (**b**) on the *right* with a single-letter alphabet

semiautomaton is the dihedral group. This system is one of the most "powerful" systems on three states with two transitions. To information scientists, what makes this example "powerful" is that it is a group.[22] The group structure endows this system with the ability to count repeatedly. Moreover, this particular group is not commutative, thus, the machine associated with it has the ability to detect differences in the sequence of action. Standard algebraic complexity holds that any system without a group structure (one or more groups in the algebra associated with the machine) has *zero* algebraic complexity (also called implementation complexity). Significantly, none of the structures implementing path independent choice posses a group structure. All of the "choice semiautomata" considered rely only on the lattices we will see as we progress. To make this semiautomaton an automaton what needs to be done is to identify one of the states as the "initial" state and some subset of the states as potential "terminal states". On the right side of Fig. 2, (Fig. 2b) is another three-state semiautomaton with a single letter alphabet. In this case the algebra associated with this machine is a chain. This semiautomaton is one of the "simplest" on three states. It is simple, in part, because, lacking a group structure, it can count only once and, being a commutative system, it does not have the ability to detect differences in sequence. An automaton is achieved once again by specifying an initial state and possible terminal states.

There are several ways to determine the algebra associated with a particular directed graph representation of a semiautomaton. One useful technique is demonstrated here using the Fig. 2a directed graph representation as a start. This method works by representing the transformations by transformation matrices and then working out the operation table for the semigroup obtained as the multiplicative

[22] A *group* is a semigroup in which there is an identity e and for which every element a in the semigroup has an *inverse* a^{-1} such that $a \cdot a^{-1} = a^{-1} \cdot a = e$.

closure of all possible products of the transformation matrices. For example, the transformation a takes state 1 into state 2 and state 2 into state 3 and state 3 into state 1. From there the cycle repeats. If we represent the initial state as the row and the destination state as the column, this transformation is represented by the following matrix,

$$a = \begin{bmatrix} 0\,1\,0 \\ 0\,0\,1 \\ 1\,0\,0 \end{bmatrix}.$$

Similarly, the transformation b is represented by this matrix,

$$b = \begin{bmatrix} 1\,0\,0 \\ 0\,0\,1 \\ 0\,1\,0 \end{bmatrix}.$$

The letters a and b are the letters of the alphabet for this Fig. 2a semiautomaton. Taking the multiplicative closure of these two matrices generates the four words of this machine. These words are presented below.

$$a^2 = \begin{bmatrix} 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \end{bmatrix}, a^3 = b^2 = I = \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix},$$

$$ba = \begin{bmatrix} 0\,1\,0 \\ 1\,0\,0 \\ 0\,0\,1 \end{bmatrix}, ab = \begin{bmatrix} 0\,0\,1 \\ 0\,1\,0 \\ 1\,0\,0 \end{bmatrix}$$

This information is compactly summarized by the operation table presented in Fig. 3.

To complete our description of this machine as a transformation semigroup, note that the underlying set $Q_3 = \{1, 2, 3\}$ while the action semigroup is the dihedral group on three elements S_{D3} with the operation table as depicted in Fig. 3. Thus, the transformation semigroup $X_D = (Q_3, S_{D3})$. Evident in the operation table is the

Fig. 3 Operation table for S_{D3} dihedral group associated with the directed graph in Fig. 2a

	I	a	a^2	b	ab	ba
I	I	a	a^2	b	ab	ba
a	a	a^2	I	ab	ba	b
a^2	a^2	I	a	ba	b	ab
b	b	ba	ab	I	a^2	a
ab	ab	b	ba	a	I	a^2
ba	ba	ab	b	a^2	a	I

Fig. 4 Operation table for S_{C3} the chain associated with the directed graph of Fig. 2b

	a	a^2
a	a^2	a^2
a^2	a^2	a^2

cyclic counting from the interaction of the "a" mapping while interaction between a and b or between a and products of a and b provide the sequence detecting ability. For example, the sequence ba followed by ab leads to a different result from the sequence ab followed by the sequence ba.

Another technique for constructing a machine algebra is exposited using the Fig. 2b semiautomaton.[23] The semiautomaton depicted in Fig. 2b is less rich. By examination, it can be seen that the single mapping a has the property that once it is applied to any state twice ($aa = a^2$) the result is that, independent of where the semiautomaton is started, it will be in state 1. This situation is represented formally in the state transition table below. On the top is a listing of the states; 1, 2,and 3. Subsequent rows specify what happens when the transition a is applied once (first row) and twice (second row).

	1 2 3
a	1 1 2
a^2	1 1 1

The corresponding operation table is presented in Fig. 4. Notably, because this semiautomaton does not have an identity mapping and, as a consequence, neither does it's algebra the operation table is not especially informative. These mappings do endow the algebraic system with a zero (a^2) and this is evident in the operation table. And, as before, we can identify the transformation semigroup as follows. The underlying set is the same $Q_3 = \{1, 2, 3\}$ but the action semigroup is different S_{C3} with the transitions as described in the state transition table and operation table as in Fig. 4. Given this, we see the transformation semigroup is $X_C = (Q_3, S_{C3})$.

Of course, these are just two of the many semiautomata that can be defined on three states and having one or two transitions. In fact, these two are both subsemigroups of the semigroup of partial functions on these three states. There are many more subsemigroups including the null machine and the identity machine and all other possibilities of partial functions mapping a state to itself or some other state or to no state at all (often called the null map). Each of these semiautomata has an associated semigroup and every semigroup has at least one semiautomaton that is associated with it. The number of semigroups varies on the precise class meant (e.g., mononids, commutative, only non-isomorphic semigroups, etc.). One nice result provided by Kleitman et al. [44] offers an asymptotic approximation for

[23] Yet another technique using permutation notation for representing the partial functions can be seen in Howie [29].

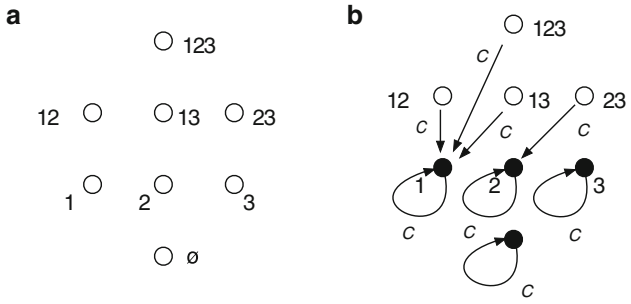


Fig. 5 Start at defining a "choice semiautomaton". (**a**) Depicts the eight states labeled as elements of the Boolean algebra on three alternatives. (**b**) Depicts the same eight states and one particular set of transformations C

the number of semigroups $S(n)$ on n elements as $S(n) = \left[\sum_{i=1}^{n} f(t)\right](1 + o(1))$ where $f(t) = \binom{n}{t} t^{1+(n-t)^2}$. Other estimates under different assumptions range from the very early Forsythe [22] to the more recent Grillet [27].

Of all the possible semiautomata on n states, we are interested only in those that implement Path Independent choice. From Johnson [31] we know that not all semiautomata meet the requirements for Path Independent choice. In Fig. 5a, b we see both the initial layout of the eight states representing the possible choice sets on three alternatives. Of course, three alternatives means there will be *eight* different sets (including the empty set from which the choice only can be the empty set) from which we might have to choose and thus, eight states in the semiautomaton.[24] Each of the states is labeled as a subset of a feasible set $\{1, 2, 3\}$. In Fig. 5a there are no transitions while in Fig. 5b there are a number of transitions. The intent is to build a semiautomaton that implements the choice function rationalized by the linear order where 1 is preferred to 2 and both 1 and 2 are preferred to 3. Clearly visible is the interval property of Path Independent choice functions; for example, every state between 1 and 123 in the Boolean algebra on $\{1, 2, 3, \}$ is mapped into 1.[25] What the interval property reflects is an equivalence class of the input signals to the semiautomaton. Specifically, because any of the input signals (here, the signal are new sets of available alternatives) $\{1, 2, 3\}$, $\{1, 2\}$, $\{1, 3\}$, and $\{1\}$, has the same effect as the choice element from each of these sets, viz. $\{1\}$. After a little further development, this choice function is used as a example in specifying the choice semiautomaton.

[24] Satoh et al. [55] calculate the number of non-equivalent semigroups of order 8 at around 1.85 billion.

[25] As in Fig. 1, the soft brackets are omitted to simplify notation.

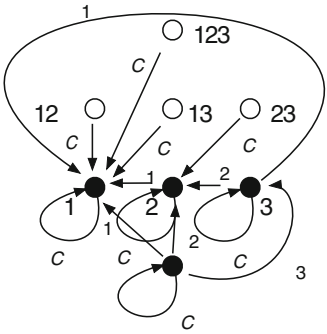


Fig. 6 Figure 5 mappings under C with additional mappings added as required for meeting the implications of Plott's product (•)

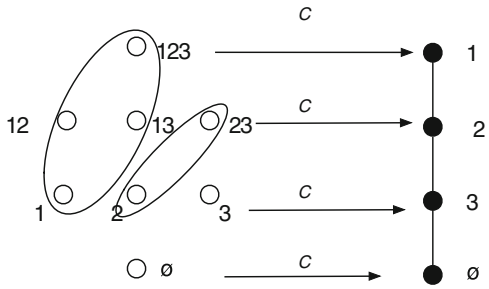


Fig. 7 On the *left* is the Boolean algebra on three alternatives and on the *right* is the image of the mapping C combined with the information of Plott's "•" to produce the chain I_{C3} on the *right*

Now, there are a number of possible labelings for the partial functions in Fig. 5 but one useful way of labeling is to use the notation C, for *choice* for each of the transitions. In Fig. 6, the additional requirements imposed by the ranking of 1 over 2 and both of these over 3 is incorporated. At this point, the diagram is getting fairly confusing. A, perhaps, more easily read representation is presented in Fig. 7. On the left in Fig. 7a is the same diagram as in Fig. 5b with the labeling and in Fig. 7b the presentation I find most useful. In the presentation, the underlying set Q is on the left and the subset of the underlying set to which all states are mapped is on the right. Here, these elements are ordered in the manner required by the Plott operation (•) while the elements $\{1, 2, 3, \emptyset\}$ in Fig. 5 use only the information of the mapping C. The ordering in Fig. 7 is a chain which is labeled I_{C3} for future use.

This representation is very useful in understanding the operation of a "choice semiautomaton". In particular, the C mapping incorporates the interval property present in all path independent choice functions and the operation (•) helps provide the ordering. In fact, for all path independent choice functions it is useful to recognize that both the mappings C and the impact of the interactions resulting from aggregating choice sets leads to a representation as a lattice.

The following remarks summarize salient points of this discussion.

Remark 1 Note that the mappings in Fig. 6 are a subset of all possible partial functions among the elements of the underlying set. While demonstrated only on this small example, the observation extends to all finite sets.

Remark 2 As a result of Remark 1, the semigroup obtained from the transitions will be a subsemigroup of the semigroup of all partial functions on the underlying set. Again, while this result is demonstrated only on a small example, the result extends to all finite sets.

In addition, because the choice functions are *functions* this is a complete semiautomaton. Analytically, it is useful to think of choice functions and Plott's (•) operation in the Boolean algebra domain and the meet and join operations in the lattice domain. And, conveniently, the lattices that form the range of the mapping have been characterized.[26] These results are summarized in the following proposition.

Proposition 1 *Let C be a discriminating choice function that satisfies PI on V, let* (•) *be the Plott product, and let* $T = (2^V; J)$ *be the complete transformation semigroup derived from C. Then*

1. *C satisfies PI if and only if J is an LLD lattice,*
2. *C satisfies PI and E if and only if J is a distributive lattice,*
3. *C satisfies WARP if and only if J is a chain of Boolean algebras, and*
4. *C satisfies SAP if and only if J is a chain.*

Examples of each of these systems on a three alternative domain are represented in Fig. 8. Figure 8a is the canonical seven-element LLD lattice that is a sublattice of every LLD lattice, Fig. 8b is a six-element distributive lattice, Fig. 8c is a chain

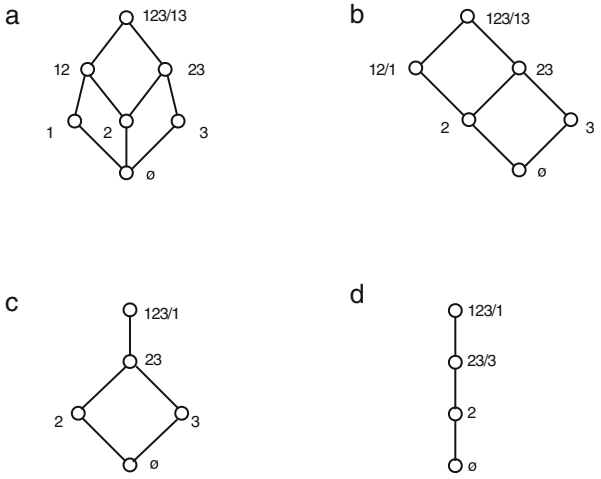


Fig. 8 (**a**) is an LLD lattice, (**b**) is a distributive lattice, (**c**) is chain of Boolean algebras and (**d**) is a chain

[26] See Johnson and Dean [35, 36] for the complete set of characterizations and Koshevoy [46] for a subset of these characterizations.

of Boolean algebras and Fig. 8d is a chain. For these lattices derived from choice functions defined on three alternatives, it turns out that the representatives of each class of lattice are differentiated by the number of elements in the image lattice. When the domain has four or more alternatives, this is no longer true.[27] On four or more alternatives, there is substantial overlap in the number of elements of the representatives of each class. Most important, independent of the number of alternatives in V, the class of distributive lattices is a strictly contained in the class of LLDS lattices and the class of chains of Boolean algebras is strictly contained in the class of distributive lattices, while the class of chains is strictly contained in the class of chains of Boolean algebras. Thus, the relevant algebras for the action semigroup are nested.

Notably the systems identified are nested so that comparison of the mathematical powers is direct. A chain is capable only of ordering a set; it does not have the ability to allow for indifference among alternatives. Similarly, a chain of Boolean algebras has the ability to permit indifference but cannot handle the case where strict preference is transitive but indifference need not be transitive. The distributive lattices have the power to allow for intransitive indifference but can not handle the case where the final choice can depend on the *sequence of expansions and contractions*. Finally, an LLD lattice has sufficient power to handle choice situations where there may not be an underlying relation rationalizing choice and where the final choice may depend of the sequence of expansions and contractions.[28] Given this nesting, it is seen that non-rationalizable path independent choice functions have the highest requirement for implementation while choice functions rationalized by linear orders have the lowest mathematical requirement.

Example Building on the choice function rationalized by the linear order of 1 preferred to 2 and 2 preferred to 3 the transformation semigroup X of choice semiautomaton can be specified. First, the underlying set, Q, is the Boolean algebra on $\{1, 2, 3\}$. As specified in Proposition 1, the action semigroup of the lattice of idempotents that, in this case, is a chain. For this choice function the idempotents are $I = \{\emptyset, 1, 2, 3\}$. Thus, the transformation semigroup is $X = (2^V, I)$. In this representation, the underlying set is not the "minimum number of states" that will implement the relevant choice. In the minimum number of state machine, only the "representative" states are required; these "representative" states are the same elements as the idempotents, I. Thus, the transformation semigroup of the minimum-number-of-state semiautomaton, $\bar{X}$, is $\bar{X} = (I, I)$. To see the operation of this choice function, consider the following choice problem

$$\{2, 3\} \bullet \{1, 3\} = ?$$

[27] See Johnson and Dean [37] for an atlas of unique LLDs on four alternatives.

[28] The key issue here is that LLD lattices are not distributive so that the final choice can depend on the interactions of the meet and join operations. Most critically, in the non-rational path independent choice functions, the choice from the intersection of two sets depends on the largest sets from which the relevant choice sets are drawn. See Johnson [34] for a concrete example of this "path dependence".

In this case, the representative state for $\{2, 3\} = 2$ (with the brackets left off the representative element). Similarly, for $\{1, 3\} = 1$ and in the lattice domain, the join operation yields $2 \vee 1 = 1$. $\diamond$

Similar examples can be constructed for the other choice functions presented in this section.

In the next section it will be seen that this ordering by implementation complexity is reversed when computational complexity is considered.

4 Computational Complexity of Path Independent Choice Functions

In general, the number of steps required to solve a problem of a particular type can depend on the nature of the computer applied to the problem. For this reason many approaches to computational complexity look for some more fundamental aspect of what is required to solve the problem. Frequently, the goal that is sought is some scaling that is independent of the particular machine or algorithm applied to the problem.[29] The measure that is used here fits within that framework. For path independent choice functions the effort that is expended to create a choice implementing semiautomaton must identify the collection of sets from which the same choice will be made. For path independent choice functions, an attractive computational complexity measure of a particular choice implementing semiautomaton is the number of prime intervals defining the collection of sets from which the same choice will be made. As noted earlier, each of these collections of sets will be an interval in the domain of the choice function being implemented.

In addition to being independent of a particular algorithm or machine, using the prime intervals has two attractive economic intuitions. First, identifying a prime interval selects two subsets of the feasible set from which the same choice will be made. Effectively, identifying a prime interval answers the question, "Do you want to make the same choice from these two sets?". This idea is firmly rooted in the view that consistency axioms are concerned with the relationship between choice made on one set and choices made on other sets. Second, the size of the two sets related by a prime interval differ by a single alternative with the superset being larger than the subset. Thus, identifying a prime interval also specifies some particular alternative that will not be in the choice from the larger of these sets.

The foundation for this approach is the following lemma from Johnson and Dean [36] which provides conditions under which a contraction of an interval will result in a new path independent choice function. Combined with a related result that assures

[29] As noted earlier computational complexity measures that have this property satisfy the Blum [12] axioms.

that the contractions are reversible, this lemma provides a means of constructing all path independent choice functions on a finite set.[30]

Lemma 1 *Let C be a path independent choice function on a set V. Let B be a meet irreducible element in the lattice of idempotents of C that is not equal to $C(V)$ or $\emptyset$. Let A be the unique element covering B in this lattice. Suppose that $A = B \cup \{x\}$. Let the function C^* defined as:*

$$C^*(S) = C(S) \text{ if } C(S) \neq A$$

$$C^*(B) = B \text{ if } C(S) = A$$

The function C^* is a path independent choice function on V. We say that C^* is obtained from C by contracting the quotient A/B in the lattice of idempotents under C.

Proof See Appendix.

While the measure of computational complexity is independent of any particular algorithm, it is useful to examine a concrete example of how the sequence of contractions described in the lemma above actually works. In particular, reviewing the operation of the contractions provides a clear understanding of what really happens to the prime intervals in the process of identifying a particular action semigroup.

Example Construction of the five discriminating choice lattices on three elements is diagramed in Fig. 9. Application of the contraction procedure is direct. In most cases, so is identifying the computational complexity. The exceptions are the sequences of contractions resulting in the two chains of Boolean algebras. Special note will be made of features of those contractions and how that relates to the computational complexity of the choice function. Consider PI choice functions defined on $V = \{1, 2, 3\}$. To help in the exposition of the process as well as to identify the prime intervals that will be the scale for the computational complexity Fig. 9 presents the domain of the choice function 2^3 on the left side, the five discriminating choice functions in the center and the intermediate lattices on the far right side. The Boolean algebras presented on the left are used to identify and keep track of the total number of prime intervals that are identified in the domain. The intermediate lattices on the right are used to identify the single prime intervals that are contracted in each application of Lemma 1.

The algorithm begins with the Boolean algebra on three elements presented at the top left of Fig. 9. The first discriminating lattice, labeled L7 at the center top of the diagram, is constructed by contracting one of the three intervals 123/12, 123/13, 123/23. Each of these contractions will result in a lattice isomorphic to the lattice L7. In L7, the interval 123/13 has been contracted. Note that *one* prime interval has

[30] See Johnson and Dean [36].

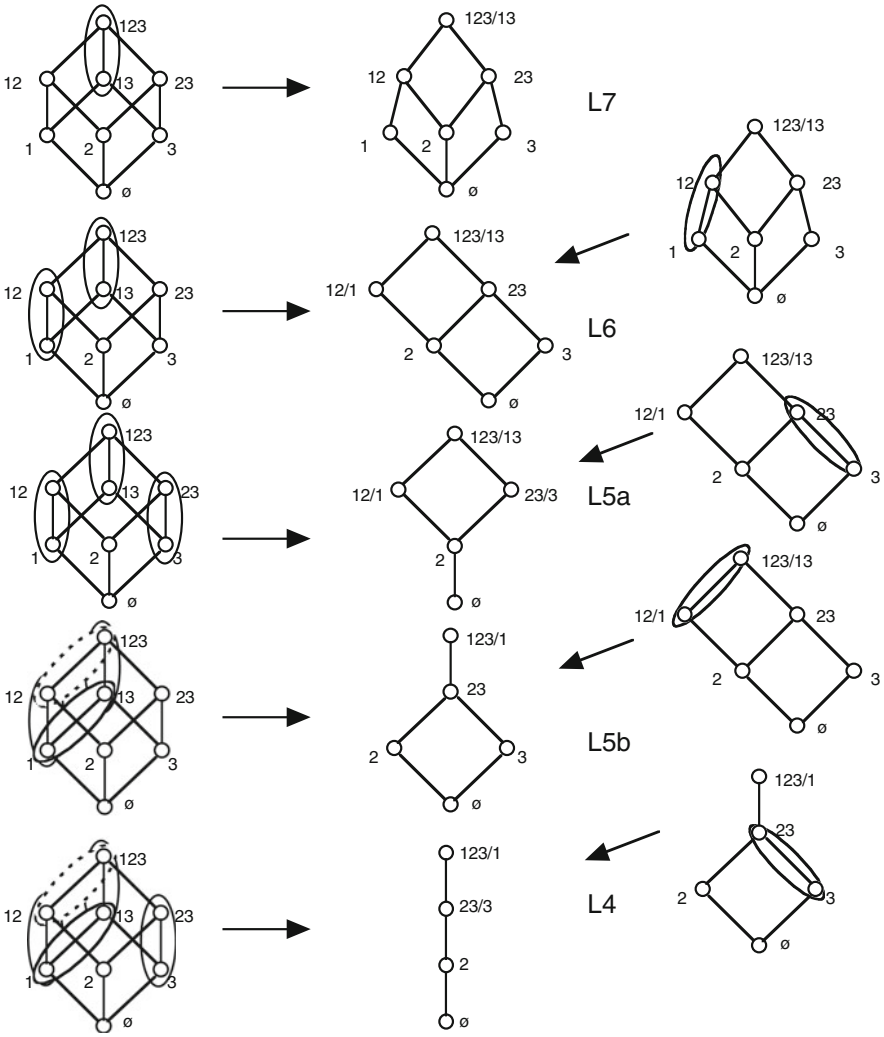


Fig. 9 Sequence of contractions leading to construction of all non-isomorphic LLD lattices that are possible action semigroups of semiautomata implementing Path Independent Choice Functions on a three alternative domain using the contraction process

been contracted and it has been determined that the alternative 2 will not be chosen from the set $\{1, 2, 3\}$.

The second lattice constructed, L6 is obtained by contracting one of two intervals in L7 meeting the conditions of Lemma 1. In this case the two intervals that could have been contracted: 12/1 or 23/3. Either contraction will result in a lattice that is isomorphic to the six element distributive lattice second from the top of the middle column labeled L6. In this case the interval 12/1 has been contracted as indicated

in the seven element lattice to the upper right of L6. It has been determined that the alternative 2 will not be selected from the set $\{1, 2\}$. As can be seen in the Boolean algebra to the left of L6, two prime intervals in the domain have been identified.

In L6, there are again two intervals that can be contracted. This time however, the contractions will not result in isomorphic lattices (in fact, they will be dual lattices). One of possible contractions, 23/3, is easy to see because it is just like the earlier contractions. Contracting this interval leads to the lattice L5a in the middle of the center column. This lattice has a two-element Boolean algebra on top of a singleton. As can be seen in the Boolean algebra to the left of L5a, in this lattice, *three* prime intervals have been contracted.

The other interval in L6 that can be contracted is an interval that involves two previously contracted intervals; this interval consists of the lattice point labeled 12/1 and the lattice point labeled 123/13 (see the copy of L6 to the upper left of L5b). The result of this contraction is the lattice with a two-element Boolean algebra on the bottom, labeled L5b. Looking to the left of this lattice, it can be seen that, although only three contraction operations have been made, a total of *four* prime intervals have been contracted. This is because the interval being contracted consisted of previously contracted intervals. Even though only three contraction operations have been made, the fourth prime interval has been contracted because of the implication of the two contractions made previously. The induced contraction is depicted by the dashed loop in the Boolean algebra to the left of L5b. Notice that this mapping is the first case of an interval that is not just a prime interval. Here, the interval consists of all the sets between $\{1\}$ and $\{1, 2, 3\}$ in the Boolean algebra for a total of four prime intervals. Equally important, the cumulative impact of the individual prime interval contractions is that alternative 1 is the only alternative chosen from any of the sets in the interval $1/123$. Observe that identifying the first prime interval determined that alternative 2 would not be chosen from the set $\{1, 2, 3\}$, identifying the second prime interval determined that alternative 2 would not be chosen from the set $\{1, 2\}$ and the final prime interval, determined that alternative 3 would not be chosen from the set $\{1, 3\}$ with the implication that only alternative 1 will be chosen from the set $\{1, 2, 3\}$.

The final contraction is performed on L5b and results in the chain labeled L4. In this case the interval $23/3$ is contracted in lattice L5b. As can be seen in the Boolean algebra to the left of L4 this lattice has *five* prime intervals that have been contracted. If instead of working with L5b we had stayed with L5a where the Boolean algebra is on top, then there are two intervals that can be contracted, and both of them involve previously contracted intervals. Contracting either of them results in a lattice isomorphic to the chain in L4, and that chain must have *five* prime intervals that have been contracted. $\diamond$

Now, the computational complexity measures can be formalized. For a standard problem it is common to consider three different computational complexities; (1) the best case for determining the answer, (2) the average case for determining the answer, and (3) the worst case for determining the answer. Currently, not enough is known about the distribution of the numbers of each type of path independent

choice function to be able to determine the average number of prime intervals that must be identified for the class. It is, however, possible to determine the best and worst cases and the results below accomplish that task.

First to simplify notation, let P denote the class of LLD lattices, D the class of distributive lattices, W the class of chains of Boolean algebras and S the class of chains. Then let P^* be the class of LLD lattices that are not distributive, D^* the class of distributive lattices that are not chains of Boolean algebras or chains and W^* the class of chains of Boolean algebras that are not chains. Note that $S = S^*$. Thus, the starred classes are in some sense, "pure" representatives of their class.

Definition 3 Let C be a path independent choice function defined on V and let J be the associated idempotent action semigroup. The *computational complexity* of J, $k(J)$, is the number of prime intervals in the Boolean algebra 2^V that are contracted in J.

Here we see that the computational complexity of a particular choice function is measured by the number of prime intervals that must be contracted in the Boolean algebra in order to construct the action semigroup for the choice implementing semiautomaton. This measure simply reflects the effort required to identify the collections of sets from which the same choice will be made.

In economic applications, a major focus is on the computational complexity of the classes of choice functions and their associated semiautomata rather than the complexity of a specific semiautomaton. This is a common event in computational complexity. Where the complexity of the a class of problems is considered, the standard approach is to identify separate complexities for the minimum complexity of the class, the average complexity of the class and the maximum complexity of the class. For choice functions, it is possible to define each of these computational complexities. For the minimum computational complexity and the maximum computational complexity of a class, the class computational complexity measure can be identified. At this stage, however, not enough is know about the number of members of each class to be able to calculate the average computational complexity.

Definition 4 For path independent choice functions defined on V with cardinality t satisfying a consistency axiom A and action semigroups J with t join irreducibles belonging to the class of *LLD* lattices B, let the *minimum computational complexity* K_{min} *of a class B* of *LLD* lattices be defined as follows:

$$K_{min}(B) = (r|r = min_{J \in B}(k(J))).$$

Definition 5 For path independent choice functions defined on V with cardinality t satisfying a consistency axiom A and action semigroups J with t join irreducibles belonging to the class of *LLD* lattices B, let the *maximum computational complexity* K_{max} *of a class B* of *LLD* lattices be defined as follows:

$$K_{max}(B) = (r|r = max_{J \in B}(k(J))).$$

Definition 6 For path independent choice functions defined on V with cardinality t satisfying a consistency axiom A and action semigroups J with t join irreducibles belonging to the class of *LLD* lattices B, let the *average computational complexity* K_{max} *of a class B* of *LLD* lattices be defined as follows:

$$K_{ave}(B) = (r|r = ave_{J \in B}(k(J))).$$

This sequence of definitions identifies: (1) a computational complexity measure for a semiautomaton implementing a particular choice function based on the number of prime intervals that must be identified in order to construct the action semigroup, (2) for a class of LLD action semigroup lattices, the computational complexity of the class by either of the minimum number of prime intervals that must be identified in order to construct a member of the class, the maximum number of prime intervals identified for a member of the class and the average number of intervals identified for the non-isomorphic members of the class. Note $K_{min}(B) < K_{ave}(B) < K_{max}(B)$ so that $K_{min}(B)$ and $K_{max}(B)$ bound $K_{ave}(B)$.[31]

Remark 3 Let V be a collection of $t \geq 3$ join irreducibles and let 2^V be the Boolean algebra on V. For discriminating choice functions defined on V, the minimum computational complexities of the following classes of action semigroups P^*, D^*, W^*, and S^* are ordered as follows:

$$K_{min}(P^*) < K_{min}(D^*) < K_{min}(W^*) < K_{min}(S^*).$$

Remark 4 Let V be a collection of $t \geq 3$ join irreducibles and let 2^V be the Boolean algebra on V. For discriminating choice functions defined on V, the maximum computational complexities of the following classes of action semigroups P^*, D^*, W^*, and S^* are ordered as follows:

$$K_{max}(P^*) < K_{max}(D^*) < K_{max}(W^*) < K_{max}(S^*).$$

Formal proofs of these results are presented in Johnson [34]. The primary technique used in the proofs is to provide general examples on t join irreducibles for each of the distinguishing cases. The remarks above demonstrate that both $K_{min}(B)$ and $K_{max}(B)$ provide the same ordering of the computational complexities of the classes of choice functions. Finally, while it is not yet possible to determine $K_{ave}(B)$ for arbitrary sized domains, direct computation on three and four element domains confirms that $K_{ave}(B)$ orders choice function classes the same as $K_{min}(B)$ and $K_{max}(B)$ for those domains.

Referring back to the example of this section, in these sample construction there is only one member of P^* which has a $K_{min}(P^*) = K_{max}(P^*) = 1$. Similarly,

[31] A joke I heard frequently from information scientists working on computational complexity problems is that "the average case is almost always the worst case".

there is only one member of D^* and it has a $K_{min}(D^*) = K_{max}(D^*) = 2$. The class W^* has two members and we see, for the first time, that $K_{min}(W^*) = 3 \neq 4 = K_{max}(W^*)$. Finally, for S^*, $K_{min}(S^*) = K_{max}(S^*) = 5$. Of course, these numbers are only for the particular choice functions in the example. A complete treatment would have to include all possible choice functions on three elements, however, in this case, up to isomorphism, all members are represented.[32]

5 Conclusions

While differing from previous approaches to addressing the structure of economic choice automata (for example, see Futia [23] or Gottinger [26]), the results presented here characterize the structure of choice implementing semiautomata when constrained to satisfy standard economic consistency axioms. The approach is to combine previous algebraic results on choice functions of Plott [52], Johnson [31, 32] and Johnson and Dean [35–38] with work on classic algebraic automata theory results from Eilenberg [20, 21] and Holcombe [28]. Notably, the characterization results are tight in that each class of PI choice function is associated with a specific class of action semigroup.

For these choice implementing semiautomata, two different complexities are identified. The first, deriving directly from the characterization results, is implementation or algebraic complexity, which reflects the mathematical power required of the semiautomaton in order to correctly implement the choice rule being effected. When ranked by algebraic complexity, the broadest class of choice functions is identified as requiring the highest power in order to be correctly implemented. As the class of choice functions becomes increasingly restricted, the power required correctly to implement the choice rule is reduced. When viewed as choice implementing semiautomata, one intuition is that as the class of choice functions becomes more restrictive, the environments in which the semiautomata operate becomes "simpler" .

In contrast, the computational complexity which is determined by the effort required to make the action semigroups of the choice implementing semiautomaton is demonstrated to be lowest for the broadest class of choice functions and increasingly higher for the more restrictive classes. The class of choice functions with the highest computational complexity is the class of choice functions rationalized by linear orders.

Perhaps most intriguingly, the two complexities are dual with algebraic complexity being highest when the computational complexity is lowest.

[32]Note here that both of the choice machines in the class W^* have the same number of lattice points and, yet, have different computational complexities.

Appendix

Proof of Lemma 1 [33]

Proof We omit the verification that C^* is a choice function on V. We shall verify that C satisfies properties Q and CA. As a preliminary recall from Lemma 11 that if, under C, $arc(B) = B^\wedge/B$ and $arc(A) = A^\vee/A$ then $A^\vee \supseteq B^\vee$. Then it follows because B is meet irreducible in the lattice of idempotents under C, that if $K \in A^\wedge/B$ then $C(K) = A$ or $C(K) = B$.

To see this, the computation: $C(K) \bullet B = C(C(K) \cup C(B)) = C(K \cup B) = C(K)$ means that $C(K) \geq B$ and so either $C(K) = B$ or $C(K) \geq A$ in the lattice. In the latter case, $C(K) = C(K) \bullet A$ while the computation $C(K) \bullet A = C(C(K) \cup C(A)) = C(K \cup A) = C(A)$ since $K \cup A \in A^\wedge/A$. This means that $K \in A^\wedge/B$ implies $K \in A^\wedge/A$ or $K \in B^\wedge/B$ and hence that $B^\wedge$ is a relative compliment of A in the quotient $A^\wedge/B$ in 2^V. It also means that $K \in A^\wedge/B$ implies that $C^*(K) = B$.

First we verify that the inverse images under C^* are quotients in 2^V. Now the inverse images of C are unchanged under C^* unless $C^*(S) = B$. So it must be verified that the inverse image of B under C^* is an interval. We prove that $\{S : C^*(S) = B\} = A^\wedge/B$. We have just shown that for $K \in A^\wedge/B, C^*(K) = B$. Conversely, as we have shown $A^\wedge \supseteq B^\wedge$ [34] so if $C(X) = B$ then $X \in A^\wedge/B$. Now $C^*(S) = B$ if and only if $C(S) \neq A$ and $C(S) = B$, in which case $S \in B^\wedge/B$ or $C(S) = A$, in which case $S \in A^\wedge/A$. So the inverse image of B under C^* is contained in $A^\wedge/B$.

Second we verify the condition: $D \supseteq E$ implies $C^*(E) \supseteq C^*(D) \cap E$. Because C is path independent, $C(E) \supseteq C(D) \cap E$. There are four cases to check:

Case 1. Neither D nor E belong to $A^\wedge/A$.

In this case there is no change from C to C^* and so the condition holds.

Case 2. Both D and E belong to $A^\wedge/A$.

Then $C^*(D) = C^*(E) = B$ and the condition holds.

Case 3. $D \in A^\wedge/A$ and $E \notin A^\wedge/A$.

In this case $C^*(E) = C(E), C(D) = A$ and $C^*(D) = B$, so the condition to be verified is $C(E) \supseteq B \cap E$. but this is true because $C(E) \supseteq C(D) \cap E = A \cap E \supset B \cap E$.

Case 4. $D \notin A^\wedge/A$ and $E \in A^\wedge/A$.

[33] The proof reproduced here is the original proof of Johnson and Dean [35]. This proof is offered here because it is more algorithmic and instructive for this application to computational complexity. Additionally, this proof is founded on basic principles instead of relying on additional constructs as in Johnson and Dean [36].

[34] The claim is from Lemma 11 of Johnson and Dean [35] restated at the end of this proof.

The condition $D \supseteq E$ entails $C(E) \supseteq C(D) \cap E$ since C is path independent. $C(E) = A$ in this case so $A \supseteq C(D) \cap E$. We are to verify that $C^*(E) \supseteq C^*(D) \cap E$, or in this case, that $B \supseteq C(D) \cap E$. Since $A = B \cup \{x\}$ this condition will hold if $x \notin C(D)$, so we suppose for the remainder of the discussion of Case 4 that $x \in C(D)$ and derive a contraction. We prove that $D \in A^\wedge/A$ contrary to the Case 4 hypothesis.

In any even we have $D \supseteq E \supseteq A$. Let $y \in D, y \notin A$, in particular $y \neq x$. Consider $B \cup \{y\}$. The computation $B \bullet C(B \cup \{y\}) = C(B \cup C(B \cup \{y\})) = C(B \cup B \cup \{y\}) = C(B \cup \{y\})$ shows that $C(B \cup \{y\}) \geq B$ in the lattice of idempotents under C. Because B is meet irreducible, either $C(B \cup \{y\}) = B$ or $C(B \cup \{y\}) \geq A$ in the lattice.

The second alternative cannot hold as we now argue. If it did $C(B \cup \{y\}) = C(B \cup \{y\}) \bullet A = C(A \cup C(B \cup \{y\})) = C(A \cup B \cup \{y\}) = C(A \cup \{y\})$. Now $D \supseteq A \cup \{y\}$ so $C(A \cup \{y\}) \supseteq C(D) \cap (A \cup \{y\})$ and since $x \in C(D) \cap A$, it follows that $x \in C(A \cup \{y\}) = C(B \cup \{y\})$; but $x \notin B \cup \{y\}$, a contradiction.

Thus for all $y \in D, y \notin A, C(B \cup \{y\}) = B$, or $B \cup \{y\} \in B^\wedge/B$, hence for these y's, $A^\wedge \supseteq B \cup \{y\}$. But if $y \in A$, then $A^\wedge \supseteq B \cup \{y\}$ anyway, so for all $y \in D, A^\wedge \supset \{y\}$; i.e. $A^\wedge \supseteq D$, but that means $D \in A^\wedge/A$, contrary to Case 4. □

Lemma 11 [35] *If C is PI and $A > B$ in the lattice of idempotents then $A^\wedge \supset B^\wedge$.*

References

1. Abreu D, Rubinstein A (1988) The structure of Nash equilibrium in repeated games with finite automata. Econometrica 56:1259–1281
2. Arrow KJ (1959) Rational choice functions and orderings. Econometrica 26:121–127
3. Arrow KJ (1963) Social choice and individual values, 2nd edn. Yale University Press, New Haven
4. Auman RJ (1981) Survey of repeated games. In: Aumann et al. (eds) Essays in game theory and mathematical economics in Honor of Oskar Morgenstern, vol 4. of Gesellschaft, Recht, Wirtschaft, Wissenschaftsverlag Bibliographisches Institute, Mannheim, pp 11–42
5. Bandyopadhyay T (1988) Revealed preference theory, ordering and the axiom of sequential path independence. Rev Econ Stud 55:343–351
6. Banks JS, Sundaram RK (1990) Repeated games, finite automata and complexity. Games Econ Behav 2:97–117
7. Bartholdi III JJ, Orlin JB (1991) Single transferable vote resists strategic voting. Soc Choice Welf 8:341–354
8. Bartholdi III JJ, Tovey CA, Trick MA (1989) Voting schemes for which it can be difficult to tell who won the election. Soc Choice Welf 6:157–165
9. Bartholdi III JJ, Tovey CA, Trick MA (1989) The computational difficulty of manipulating an election. Soc Choice Welf 6:227–241
10. Ben-Porath E (1990) The complexity of computing a best response automaton in repeated games with mixed strategies. Games Econ Behav 2:1–12
11. Birkhoff G (1979) Lattice theory, 3rd edn. American Mathematical Society, Providence

[35] See Johnson and Dean [35].

12. Blum M (1967) A machine-independent theory of complexity of recursive functions. J Assoc Comput Mach 14:322–336
13. Campbell DE (1978) Realization of choice functions. Econometrica 46:171–180
14. Campbell DE (1978) Rationality from a computational standpoint. Theor Decis 9:255–266
15. Chatterjee K, Sabourian H (2009) Game theory and strategic complexity. Encyclopedia of complexity and systems science. Springer, New York, pp 4098–4114
16. Clifford AH, Preston GB (1961) The algebraic theory of semigroups, vol 1. American Mathematical Society, Providence
17. Clifford AH, Preston GB (1961) The algebraic theory of semigroups, vol 2. American Mathematical Society, Providence
18. Davey BA, Priestly HA (1990) Introduction to lattices and order. Cambridge University Press, Cambridge
19. Dilworth RP (1950) A decomposition theorem for partially ordered sets. Ann Math 51:161–166
20. Eilenberg S (1974) Automata, languages and machines, vol A. Academic, New York
21. Eilenberg S (1976) Automata, languages and machines, vol B. Academic, New York
22. Forsythe GE (1955) SWAC computes 126 distinct semigroups of order 4. Proc Am Math Soc 6:443–447
23. Futia C (1977) The complexity of economic decision rules. J Math Econ 4:289–299
24. Gilboa I (1988) The complexity of computing best response automata in repeated games. J Econ Theory 45:342–352
25. Gorman WM (1959) Separable utility and aggregation. Econometrica 27:469–481
26. Gottinger HW (1978) Complexity in social decision rules. In: Gottenger HW, Leinfellner W (eds) Decision theory and social ethics, pp 251–269. D. Reidel, Dordecht
27. Grillet PA (1995) The number of commutative semigroups of order N. Semigroup Forum 50:317–326
28. Holcombe WML (1982) Algebraic automata theory. Cambridge University Press, Cambridge
29. Howie JM (1991) Automata and languages. Oxford University Press, New York
30. Hurwicz L (1986) On informational decentralized systems. In: McGuire CB, Radner R (eds) Decision and organization. North-Holland, Amsterdam
31. Johnson MR (1990) Information, associativity and choice. J Econ Theory 52:440–452
32. Johnson MR (1995) Ideal structures of path independent choice functions. J Econ Theory 65:468–504
33. Johnson MR (1996) Algebraic complexity of strategy-implementing semiautomata for repeated-play games. Mimeo presented southeast economic theory and international trade meetings 1–3 Nov 1996, Florida International University, Miami, FL
34. Johnson MR (2005) Economic choice semiautomata; structure, complexities and aggregations. Presented at Econometric Society 2005 World Congress, London
35. Johnson MR, Dean RA (1996) An algebraic characterization of path independent choice functions. Presented at the third international meeting of the Society for Social Choice and Welfare, Maastricht, 1996. Available at http://markrjohnson.net/read-me/
36. Johnson MR, Dean RA (2001) Locally complete path independent choice functions and their lattices. Math Soc Sci 42:53–87
37. Johnson MR, Dean RA (2001) The construction of all lower locally distributive lattices on 4 elements. Available at http://markrjohnson.net/read-me/
38. Johnson MR, RA Dean (2002) Construction of finite lower locally distributive lattices. Presented at American Mathematics Society Meetings, San Diego, CA. Available at http://markrjohnson.net/read-me/
39. Johnson MR, Dean RA (2005) Designer path independent choice functions. Econ Theory 26:729–740
40. Kalai E (1990) Bounded rationality and strategic complexity in repeated games. In: Ichiishi T, Neyman A, Tauman Y (eds) Game theory and applications, pp 131–157. Academic, San Diego
41. Kalai E, Stanford W (1988) Finite rationality and interpersonal complexity in repeated games. Econometrica 56:397–410
42. Kelly JS (1978) Arrow impossibility theorems. Academic, New York

43. Kelly JS (1988) Social choice and computational complexity. J Math Econ 17:1–8
44. Kleitman JK, Rothschild BR, Spencer JH (1976) The number of semigroups of order n. Proc Am Math Soc 55:227–232
45. Knoblauch V (1994) Computable strategies for repeated prisoner's dilemma. Games Econ Behav 7:381–389
46. Koshevoy G (1999) Choice functions and abstract convex geometries. Math Soc Sci 38:35–44
47. Krohn KB, Rhodes JL (1962) Algebraic theory of machines. In: Proceedings of symposium on the mathematical theory of automata, pp 341–378. Polytechnic Institute of Brooklyn, New York
48. Krohn KB, Rhodes JL (1965) Algebraic theory of machines, I Prime decomposition theorem for finite semigroups and machines. Trans Am Math Soc 116:L450–L464
49. Lewis A (1985) On effectively computable realizations of choice functions. Math Soc Sci 10:43–80
50. Lewis A (1985) The minimum degree of recursively representable choice functions. Math Soc Sci 10:179–188
51. Papadimitriou CH (1992) On players with a bounded number of states. Games Econ Behav 4:122–131
52. Plott CR (1973) Path independence, rationality and social choice. Econometrica 41:1075–1091
53. Radner R (1993) The organization of decentralized information processing. Econometrica 61:1109–1146
54. Rubinstein A (1991) Comments on the interpretation of game theory. Econometrica 59:909–924
55. Satoh S, Yama K, Tokizawa M (1994) Semigroups of order 8. Semigroup Forum 49:7–29
56. Simon HA (1972) Theories of bounded rationality. In: McGuire CB, Radner R (eds) Decision and organization. North-Holland, Amsterdam
57. Strotz RH (1957) The empirical implications of the utility tree. Econometrica 25:269–280
58. Strotz RH (1959) The utility tree a correction and further appraisal. Econometrica 27:482–488
59. Turing AM (1937) On computable numbers, with an application to the entscheidungs problem. Proc Lond Math Soc 42:230–265

Moral Responsibility and Individual Choice

Constanze Binder and Martin van Hees

Abstract In this paper we analyze within the framework of individual choice theory assignments of moral responsibility. For this purpose we introduce a so-called responsibility function that describes for any choice situation the alternatives for which the agent would be deemed responsible if she were to choose one of them in that situation. We show under which conditions a responsibility function can be rationalized by information about which courses of action constitute reasonable alternatives to other courses of action. After thus having characterized one way of assigning responsibility, we show that it leads to what we call the agency paradox: a rational person will in many cases not be responsible for her actions. It is argued that a decision rule that is formally the same as the 'never choose the uniquely largest'-rule characterized by Baigent and Gaertner (1996) circumvents the paradox. Turning to a possible counterargument to the analysis presented, we conclude by suggesting that moral responsibility should be seen as a criterion for the assessment of the quality of our choice sets rather than as a consideration that is relevant when making our choices.

Keywords Choice • Context-dependency • Rationality • Responsibility

1 Introduction

How does the nature of the choice situation we face affect our responsibility for the choices that we make in that situation? The objective of this paper is to address this question by using the methods of individual choice theory.

C. Binder (✉)
Faculty of Philosophy, Erasmus University Rotterdam, Campus Woudestein, Postbus 1738, 3000 DR Rotterdam, The Netherlands
e-mail: binder@fwb.eur.nl

M. van Hees
Department of Philosophy, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
e-mail: m.van.hees@vu.nl

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_6

Moral responsibility is a complex concept and a topic of intense debate in the philosophical literature. A particularly contested question concerns the conditions under which a person can be deemed responsible for the outcomes resulting from her actions.[1] The philosophical debate can roughly be divided into discussions about three different kinds of conditions for such moral responsibility [3]. First, in order to deem a person responsible, the way she makes her decisions has to satisfy certain attributes of *agency*. This is usually taken to mean that the person is an autonomous agent who intentionally acts on the basis of reasons and can distinguish between right and wrong. People who were manipulated or brain-washed, for instance, are excluded from bearing (full) moral responsibility for their actions. Second, a person's act and the outcome resulting from it should be *causally connected* in the right way. This means that there should not only be a causal path between the action and the outcome, but the path should be strong or relevant enough to constitute responsibility. If my sun-glasses dazzle a passing car driver who hits another car as a result, I will not be morally responsible for the outcome, i.e. the car accident. Third, *reasonable alternatives* to the action leading to the outcome should have been available. If a mountaineer cannot come to the rescue of his comrade because he is stuck in a crevice himself, he is not morally responsible for not having helped his friend. Moreover, the alternative action should be a reasonable one. If the mountaineer can only help his friend by crossing an avalanche-imperilled slope, and thus by seriously risking his own life, we may not deem him responsible if he refrains from attempting to do so.

Each of the three conditions has been contested but the third is perhaps the most controversial one. Indeed, Frankfurt famously argued that having had the option (reasonable or not) to do differently than one did is not a necessary requirement for being responsible [5]. Yet Frankfurt's arguments have also been contested. In particular, it has been argued that, contrary to what Frankfurt claims, his counter-example does presuppose the existence of alternative courses of action (see e.g. [3, 15]). We shall not enter this debate here and simply assume that having had an alternative opportunity to do otherwise is indeed a necessary requirement. In fact, we shall assume throughout the analysis that the first two conditions—the agency condition and the causality condition—are met, and focus only on the analysis of the third condition.

We do so by drawing on the formal apparatus employed in individual choice theory. Section 2 presents the central elements of our framework. It describes a particular kind of choice function, which we call a *responsibility function*, and which assigns to each choice situation those outcomes for which a person can be held responsible if she were to choose them in that situation. Furthermore, it introduces a binary relation describing which actions are reasonable alternatives to each other as well as a definition of how such a relation can be said to rationalize responsibility functions. It is however not the usual notion of rationalization that we

[1] We shall not distinguish between claims that an agent *is* responsible and claims that she can justifiably *be held* responsible.

employ but a version of it which we call 'negative rationalization'; it expresses the idea that one should have a reasonable alternative to what one is doing in order to be responsible. Section 3 contains the sufficient and necessary conditions under which negative rationalizations can be obtained. The result is used, in Sect. 4, to point out that a paradox emerges if the standard of value underlying a responsibility relation coincides with a person's actual preferences: in such cases a person cannot be held responsible for choosing her uniquely most preferred option. In Sect. 5 we show how the rule characterized by Baigent and Gaertner [1] allows an escape from the highlighted paradox. The paper is concluded in Sect. 6 where, through a discussion of a possible counterargument to our approach, we suggest that moral responsibility is relevant for assessing the quality of choice sets we have but not for determining how we make our choices from those sets.

2 Responsibility and the Context of Choice

Let X denote the universal finite set of alternatives, each element to be interpreted as a combination of acts (that brings about some particular outcome). The set of all non-empty subsets of X will be denoted by Z. A choice situation $A \in Z$ describes the various combinations of actions the person can adopt and by assumption she can choose one, and only one, of those combinations. We assume complete information, which here means that each element of X can be associated with one particular outcome. For this reason, we can refer to the elements of some X both as actions and as outcomes.

Given some $A \in Z$, we let $V(A)$ denote the set of all elements of A for which the agent will be deemed responsible if she decided to choose them. Thus, $x \in V(A)$ the agent would be responsible for choosing x if she indeed were to choose x from A. Since we take responsibility to be a key feature of human agency, we shall refer to the elements of $V(A)$ as the 'agentive' elements of A.

Definition 1 A responsibility function is a mapping V assigning to each $A \in Z$ a (possibly empty) subset of A.

Whether a person can be held responsible for a particular choice will—among other things—depend on the choice set the alternative was chosen from. As mentioned above, we take the possibility of doing otherwise, i.e. the existence of other options in one's choice set, to be necessary for deeming a person responsible for her actions. In particular, the availability of reasonable alternatives is required. The local grocer who is ordered at gun point to hand over the cash register has an alternative option, namely to ignore the command. Yet given the circumstances it is not a reasonable alternative and, assuming she had no other options, we will therefore not say she is responsible for having handed over the money.

To capture this formally, we assume the existence of a binary relation M over X, where xMy is to be interpreted as 'x forms a reasonable alternative to y'. In most cases, the relation M will have a symmetric part as well as an a-symmetric one.

That is, two alternatives can be reasonable *vis-à-vis* each other, as is the case with two very attractive job offers for instance, whereas in other cases only one of the two alternatives is a reasonable alternative to the other one, such as in the robbery example.

Given such a reasonableness relation M, consider the following definition of 'negative rationalization' of a responsibility function V:

Definition 2 A relation M negatively rationalizes a responsibility function V if, and only if, for all non-singleton subsets A, $V(A) = \{x \in A \mid yMx \text{ for some } y \in A\ (y \neq x)\}$.

A responsibility function is (negatively) rationalized if it always selects those alternatives for which there is a reasonable alternative. Thus it expresses the idea that, assuming all other responsibility conditions are met, to be responsible for one's choice is to have had reasonable alternatives to it. Since an alternative cannot be a reasonable alternative to itself, the definition of rationalization is applied to choices from non-singleton sets only.

3 Rationalizing Responsibility

In this paper we shall only consider rationalizations through orderings, i.e., M is transitive (if x is a reasonable alternative to y, and y is to z, then x is so to z) and complete (for every pair of alternatives at least one of the two options is a reasonable alternative to the other one). We do so to keep the analysis relatively simple, but it should be pointed out that both transitivity and completeness are strong assumptions. First, instances of the Sorites paradox often constitute violations of transitivity and they may do so here as well. Consider the mechanism of a weakness of will scenario in which, for any positive integer k, consuming k glasses of wine (cigarettes, pieces of candy) may well be a reasonable alternative to consuming $k - 1$ of them. Yet, to consume an abundance of glasses of wine is not a reasonable alternative to drinking no wine.

Completeness will be hard to defend when an agent has to make a choice between two equally gruesome alternatives, alternatives that would never be chosen voluntarily by an agent.[2] To illustrate, take a very extreme case of a hard choice: the situation the protagonist of the novel *Sophie's Choice* [13] is in. She has to decide which of her two children will be killed and which will survive. Saying that she had a reasonable alternative to whatever choice she makes, and thus to say that she

[2]Completeness entails reflexivity and reflexivity can also be questioned: how can an option be a reasonable alternative to itself? Yet, reflexivity is of no further relevance for the analysis and could in fact also be dropped. That is, the results can be reformulated in such a way that they hold for connectedness (i.e. for all *distinct* x and y, x is a reasonable alternative to y or y is so to x) rather than for completeness of M.

bears responsibility for choosing one child rather than the other, not only seems counterintuitive but plainly immoral. In defence of completeness, it could be argued that we do in fact assign some responsibility for her action. That is, one explanation for the horrible nature of the choice that is inflicted upon her, and which in the novel eventually leads to her suicide, is that whatever she does she will bear *some* responsibility for having chosen as she did (although of course not for having to make *a* choice). Yet such a defence of completeness is controversial, to say the least.[3]

Negative rationalizability is characterized by three axioms. The first is the inverse of the well-known consistency condition α and states that if a person is responsible for a choice of x in a set A, then he is also so if he were to choose x from some superset of A. Stated differently, adding elements to a set does not undermine the responsibility for one's choice.

Axiom 1 *For all singleton sets A and all $B \in Z$: $V(A) \subseteq V(A \cup B)$.*

It could be objected that having more freedom of choice can in fact decrease our responsibility. There is ample evidence that the anxiety caused by having too much choice can undermine our decision-making skills. If this in turn undermines our responsibility then, so the argument would run, having more choice decreases our responsibility. However, the argument misinterprets the causal relations between choice, stress and responsibility. The correct causal view which is compatible with the axiom without denying the burdens of choice is what we may call the 'Sartrean' one: more choice leads to more responsibility, and more responsibility leads to more anxiety.[4]

The next axiom states that only in the case of singleton choice sets can a person not be deemed responsible for any of the choices available to her.

Axiom 2 *For all A, if $V(A) = \emptyset$, then A is a singleton set.*

It is easily seen that the axiom entails completeness of any relation M that would negatively rationalize the choice function. A critique of completeness thus can also be formulated as a critique of the axiom. Indeed, the axiom states that one can always make at least one choice for which one is responsible. It is precisely this assumption that is contested when we discuss an agent confronted with equally unattractive options, as in the example of Sophie's choice discussed above.

Axiom 3 *For all non-singleton sets A: if $V(A) = A$ then there is some non-singleton set $B \subseteq A$ such that (a) for all distinct $x, y \in B$, $V(\{x, y\}) = \{x, y\}$, and (b) if $A - B \neq \emptyset$, then for all $x \in B$ and all $y \in A - B$: $x \notin V(\{x, y\})$.*

[3] Another complication is that it can only work as an argument if one presupposes that there are different degrees of responsibility. That is, one can only claim that Sophie bears at least *some* responsibility for her choice if one takes the degree of her responsibility for the death of her child to be positive but minute compared to the responsibility of the Nazi who puts her in the position. Yet the framework that we develop here does not allow for such comparative judgments.

[4] Note that Barry Schwartz [9] takes this line in his analysis of choice stress.

The axiom can best be explained through an example. Suppose a person entering the job market has received various very attractive as well as various rather unattractive job offers. Letting A denote the choice situation, each of the elements of A is what we have called an agentive one—if the agent were to make a choice from A, he would be responsible for his choice (there was a reasonable alternative to each possible choice). Call such a set containing only agentive choices a 'fully agentive' set. The axiom now states that any fully agentive set containing two or more elements has some fully agentive B also containing two or more elements such that each element of B is *not* agentive when considered in combination with exactly one of the elements not in B. In the example, the set B is the set of all attractive job offers. Indeed, if the agent's choice set contains exactly one element of B (i.e. one very attractive option) and one element of $A - B$ (i.e. one of the very unattractive job offers), then the person is not responsible if he chooses the attractive option. The underlying intuition of course is that he had no real choice.

Proposition 1 *A responsibility function can be negatively rationalized by an ordering M of X if, and only if, it satisfies Axioms 1–3.*

Proof

$\Leftarrow$: Let V satisfy the axioms and define M as (a) xMx for all $x \in X$, (b) for all distinct x, y, xMy iff $y \in V(\{x, y\})$. For all distinct x, y, we have xMy or yMx by Axiom 2. Since we have xMx for all x by stipulation, M is complete. To prove transitivity, assume to the contrary that for some distinct x, y, z with $y \in V(\{x, y\})$ and $z \in V(\{y, z\})$, we have $z \notin V(\{x, z\})$. We then have $V(\{x, z\}) = \{x\}$ by Axiom 2. By Axiom 1 therefore $V(\{x, y, z\}) = \{x, y, z\}$. Since $V(\{x, z\}) = \{x\}$, the set B to which Axiom 3 refers must be $\{y, z\}$ or $\{x, y\}$. It cannot be $\{y, z\}$ since Axiom 3 then implies $y \notin V(\{x, y\})$, which would be a contradiction. It cannot be $\{x, y\}$ either, since then Axiom 3 would imply $x \notin \{x, z\}$. But then $V(\{x, z\}) = \emptyset$, which is a contradiction.
Next we show that M negatively rationalizes $V(.)$. Take arbitrary non-singleton set A. First, we show that $V(A) \subseteq \{x \in A \mid yMx \text{ for some } y \in A\ (y \neq x)\}$. Let $x \in V(A)$. By definition of M, we need to prove that $x \in V(\{x, y\})$ for some $y \in A\ (y \neq x)$. Suppose there is no such y. Then for all $y \in A$, $x \neq y$, we have $V(\{x, y\}) = \{y\}$ by Axiom 2. Axiom 1 and $x \in V(A)$ subsequently implies that $V(A) = A$. Applying Axiom 3, there is some non-singleton set $B \subseteq A$ such that (a) for all distinct $v, w \in B$, $V(\{v, w\}) = \{v, w\}$, and (b) if $A - B \neq \emptyset$, then for all $v \in B$ and all $w \in A - B$: $v \notin V(\{v, w\})$. If $x \in A - B$, $y \notin V(\{x, y\})$ for all $y \in B$, contradicting $V(\{x, y\}) = \{y\}$ for all $y \neq x$. Hence, $x \in B$. Since B is a non-singleton set, for some $y \in B$ $(x \neq y)$, $V(\{x, y\}) = \{x, y\}$, which is a contradiction. Second, we show that $\{x \in A \mid yMx \text{ for some } y \in A\ (y \neq x)\} \subseteq V(A)$. Assume x is an element such that $x \in A$ and yMx for some $y \in A\ (y \neq x)$. By definition of M, $x \in V(\{x, y\})$. By Axiom 1, therefore $x \in V(A)$, which entails $\{x \in A \mid yMx \text{ for some } y \in A\ (y \neq x)\} \subseteq V(A)$.

$\Rightarrow$: Let M be an ordering that negatively rationalizes V. We have to show that it satisfies Axioms 1–3. Consider any non-singleton subset A and $x \in V(A)$. By

definition of negative rationalization, there is some $y \in A$ such that yMx. Since $y \in A \cup B$ for any B, x is also in $V(A \cup B)$, which shows that Axiom 1 is satisfied. Axiom 2 follows directly from M being an ordering and the definition of rationalization. To show Axiom 3 is satisfied, assume $A = V(A)$, and let B denote the set of M-best elements of A. Because M is an ordering, B is non-empty. If B is a singleton set, say $B = \{x\}$, not-yMx for all $y \in A - \{x\}$. By definition of negative rationalization, $x \notin V(A)$, which is a contradiction. Hence, B is not a singleton set, and for all distinct $x, y \in B$, xMy and yMx; and for all $z \in A - B$, xMz but not zMx. Hence, we then have for all $x, y \in B$, $V(\{x, y\}) = \{x, y\}$ and, if $A - B \neq \emptyset$, for all $x \in B$ and all $z \in A - B$, $x \notin V(\{x, z\})$. Hence, Axiom 3 is satisfied as well. □

One question raised by Proposition 1 is the nature of the relation M. What makes an option a reasonable alternative to another one? In our examples we appealed to the intuition that an alternative that is utterly unattractive compared with some other alternative, cannot be considered to be a reasonable alternative to the latter. That is, an option must be of sufficient quality relative to the other option in order to be considered a reasonable alternative to it. This immediately raises the question which standard of value is to be invoked when assessing the quality of an alternative. In the next section we consider one possible but rather controversial answer to this question.

4 The Agency Paradox

It is sometimes said that the picture of a decision maker optimizing a given preference ordering is at odds with the idea of the person being free or autonomous. Suppose that preferences can be distinguished from our actions in the sense that they precede them (rather than that they describe them, as in a revealed preference approach). If we now say that the rational agent always chooses his most preferred outcome, and if the persons preference relation coincides with the reasonableness relation, then there seems to be no scope for actions that are sub-optimal.[5] By virtue of his rationality, the rational agent will never choose to perform such actions—so it is argued. We may call this the 'agency paradox' of rational choice: rationality and autonomy are both seen as essential features of what it means to be an agent yet the presuppositions of rational choice theory are taken to imply that an individual can never be both rational and autonomous.

[5]Note that, though they coincide, the preference relation and the reasonableness relation are still conceptually different.

In terms of our framework, the reasoning leading to the supposed clash between autonomy and rationality can be reconstructed as follows:

Premise 1. A rational agent will always choose one of his most preferred outcomes.

Premise 2. For a rational agent an outcome x is a reasonable alternative to y if, and only if, he weakly prefers x to y.

Conclusion. Whenever the opportunity set of a rational agent contains a uniquely best outcome, the agent is not responsible for his choice.

To illustrate, suppose a manager having to make a decision about whom to hire for some position finds candidate x to be strictly better than any of the other candidates. By Premise 1 and by the assumption of individual rationality, she will indeed hire x. Premise 2 and Proposition 1 entail that x is not an element of $V(A)$. By definition of V the manager cannot be said to be responsible for appointing the person—she did not have a reasonable alternative.

It follows on this view that a rational agent can only be responsible for her actions if she has more than one best outcome. For instance, if there had been another candidate y that she found equally good as x, then the manager would have had a reasonable alternative to x. But it is unsatisfactory to say that only in cases of indifference can responsibility for one's choices emerge. When someone is indifferent between two alternatives then one has no reason to choose one alternative rather than the other. The indifferent person is picking rather than choosing [14] and it is not very attractive to say that the locus of our responsibility is to be found in such random acts.[6] However, if we assume that the alternatives between which an individual is indifferent cannot be reasonable alternatives to each other, we derive an even stronger version of the agency paradox: a rational individual then *never* is responsible for her choices.

5 Never Choose The Uniquely Largest Element

One possible escape route from the agency paradox can be drawn from the analysis of Baigent and Gaertner [1] of a decision rule suggested by Sen. Sen [10] famously illustrated the possibility of norms constraining individual choices with an example of someone who is a guest somewhere and is being invited to take a slice of cake. The slices are of different size and, in order not to be impolite, the person does not choose the single largest piece on the plate but the next to largest piece, despite his preference for larger slices of cake. Reasonable as such behaviour is, it cannot be rationalized by the standard consistency conditions of individual choice.

[6]The argument parallels discussions in the freedom of will literature. There the luck principle, as Kane [6] has labelled the argument, states that indeterminism entails chance or luck which is incompatible with moral responsibility.

Baigent and Gaertner present a generalization of such choice behavior as well as a characterization of it in terms of non-standard conditions of consistency. The generalization makes use of a weak ordering, which represents the relevant quality description of the various choices. In the context of Sen's original example it would represent the size of the pieces of cake but other interpretations can be given in other choice situations. The choice function C characterized by Baigent and Gaertner can be interpreted as a two-step procedure.[7] In the first step, the ordering describing the relevant quality features is used to filter out the elements of A that are inadmissible. In the cake example the filter deems the unique largest element to be inadmissible. The second step consists of choosing the most preferred elements of the remaining alternatives.

To define the two-step procedure in more precise terms, let for any $A \in Z$ and any ordering R of A, $B^*(A, R)$ denote the uniquely best R-element of A if there is one and the empty set otherwise, and let $B(A, R)$ be the set of R-best elements of A.[8] Let R' denote the preferences of the agent.

Definition 3 ((Self-)Constrained Choice Functions) Given some orderings R and R', we say that a choice function C is *constrained* if for all non-singleton sets $A \in Z$, $C(A) = B(A - B^*(A, R), R')$. If $R = R'$, C is said to be *self-constrained*.

Baigent's and Gaertner's choice function is self-constrained: the relation that is used in the first step coincides with the individual preferences of the agent. In the cake example, for instance, the agent always prefers the larger slice of cake. As a result, the inadmissible element $B^*(A, R)$ (the uniquely largest slice of cake) is also the utility maximizing element of A (the most preferred slice of cake).

We can now see that if $R = M = R'$, then we arrive at a type of self-constrained rationality that is formally identical to Baigent's and Gaertner's choice function but interpreted in terms of responsibility. Indeed, Premise 1 of the inference presented in the previous section can then be reformulated. It now states that a rational agent will never choose the uniquely largest element, which, in terms of responsibility, comes down to:

> **Premise 1'.** A rational agent will always choose one of her most preferred outcomes from the set of outcomes for which she would be responsible if chosen by her.

Given this premise, a rational agent will always be responsible for her actions.[9]

The amendment of the first premise can be seen as a compromise between rationality and morality which forms a way out of the agency paradox. If the preferences do not coincide with the responsibility relation, we arrive at a constrained rather

[7] For a more general account of norm-constraint choices, see Bossert and Suzumura [2].

[8] Note that with respect to $B^*(A, R)$ we use the term 'best' in a purely formal sense of the word: the uniquely best element is in fact an inadmissible alternative.

[9] As one of the referees pointed out to us, if one uses a theory of moral responsibility in which responsibility judgements *only* make sense if there is a conflict between individual rationality and some other motivation, then responsibility always is about the possibility of constrained choice; in other contexts the agency paradox does not arise.

than self-constrained choice function: in a first step all admissible alternatives are selected on the basis of the responsibility relation; in a second step a person chooses her most preferred element from the set of admissible alternatives on the basis of her preference relation.[10]

6 Conclusion

The objective of this paper was to explore the role of the context of choice for the assignment of judgments of moral responsibility. We did so by drawing on the formal framework of rational choice theory and examined the conditions under which a responsibility function can be negatively rationalized by a responsibility relation. Our analysis revealed a paradox if the standard of value underlying a person's responsibility relation coincides with a person's preferences: a person cannot be deemed responsible for choosing her uniquely most preferred option. We also showed that the self-constrained behavior that forms a way out of the paradox is formally the same as a decision rule characterized by Baigent and Gaertner [1]. Given the many different positions in the responsibility literature, it should not come as a surprise that our analysis can be contested. We conclude with a brief discussion of two possible objections to our approach.

First, it could be doubted whether choosing a unique best M-element—where M may or may not coincide with one's preferences—can never be an act for which one is responsible in the sense of being praiseworthy. Yet this follows from our account.[11] Take Peter Singer's famous example of the person who is passing a pond and sees a child drowning [12]. Her only reasonable option is to jump in and rescue the child. Should we claim that she is not responsible for her act when she rescues the child? This is indeed what we claim. In our opinion, the counterintuitive feel of this only arises because of an implicit and unjustified equivocation with heroic acts. When a rescuer makes a real sacrifice or takes grave risks, then not interfering *is* a reasonable alternative (indeed, that fact is what makes the act heroic) and thus she can be held responsible, i.e. can be praised, for saving the child. It is this intuition that is carried over incorrectly, so we believe, to situations in which there is no reasonable alternative.[12]

[10] See also Sen [11].

[11] Dennett [4] has taken such cases—labelled by Moya [7] as 'Luther cases' ('Here I stand, I can do no other')—as reasons for rejecting the necessity of having alternative possibilities when assigning moral responsibility.

[12] A different response, but not compatible with our assumptions, is to argue for the existence of a praise-blame asymmetry. On that view, assignments of praise do not require alternative possibilities but assignments of blame do. For a defence of this view, see Moya [8].

A second critique, and in our view the more important one, agrees with the conclusion that we would not be responsible for rescuing the child. If rescuing the child was the only reasonable option we had, we should indeed not be praised for it. However, rather than seeing the lack of responsibility as a problem, this view denies that autonomy is about choosing an agentive outcome, that is, an outcome for which one can be deemed responsible. Instead, autonomy is about doing the most *reasonable* thing. Stated differently, when we say that responsibility is important or essential for our agency, we mean that having a choice set that is sufficiently rich to yield responsibility assignments is important. It does *not* mean that we should be motivated to make a choice for which we can be deemed responsible. Moral responsibility is important for assessing the quality of our choice situations, not for determining how to make our choices.

Acknowledgements We would like to thank an anonymous referee for very helpful comments on an earlier version of this paper.

References

1. Baigent N, Gaertner W (1996) Never choose the uniquely largest: a characterization. Econ Theory 8(2):239–249
2. Bossert W, Suzumura K (2009) External norms and rationality of choice. Econ Philos 25:139–152
3. Braham M, van Hees M (2012) An anatomy of moral responsibility. Mind 121:601–634
4. Dennett DC (1984) Elbow room: the varieties of free will worth wanting. Clarendon, Oxford
5. Frankfurt H (1969) Alternate possibilities and moral responsibility. J Philos 66:829–839
6. Kane R (1999) Responsibility, luck, and chance: reflections on free will and indeterminism. J Philos 96:217–240
7. Moya CJ (2006) Moral responsibility. The ways of scepticism. Routledge, Abingdon
8. Moya CJ (2010) Alternatives and responsibility: an asymmetrical approach. In: Reboul A (ed) Philosophical papers dedicated to Kevin Mulligan, Genève. http://www.philosophie.ch/kevin/festschrift/
9. Schwartz B (2004) Paradox of choice: why more is less. Harper Collins, New York
10. Sen A (1993) Internal consistency of choice. Econometrica 61:495–521
11. Sen A (1997) Maximization and the act of choice. Econometrica 65:745–779
12. Singer P (1972) Famine, affluence, and morality. Philos Public Aff 1(3):229–243
13. Styron W (1979) Sophie's choice. Random House, New York
14. Ullmann-Margalit E (1977) Picking and choosing. Soc Res 44:757–785
15. Widerker D, McKennan M (eds) (2003) Moral responsibility and alternative possibilities—essays on the importance of alternative possibilities. Ashgate, Aldershot

Part II
Collective Choice and Collective Rationality

Multi-Profile Intertemporal Social Choice: A Survey

Walter Bossert and Kotaro Suzumura

Abstract We provide a brief survey of some literature on intertemporal social choice theory in a multi-profile setting. As is well-known, Arrow's impossibility result hinges on the assumption that the population is finite. For infinite populations, there exist non-dictatorial social welfare functions satisfying Arrow's axioms and they can be described by their corresponding collections of decisive coalitions. We review contributions that explore whether this possibility in the infinite-population context allows for a richer class of social welfare functions in an intergenerational model. Different notions of stationarity formulated for individual and for social preferences are examined.

Keywords Decisiveness • Infinite-population social choice • Intergenerational choice • Multi-profile social choice

1 Introduction

The conclusion of Arrow's [1, 1963 (2nd edn.); 2012 (3rd edn.)] dictatorship theorem depends on the assumption that the population under consideration is finite. This observation goes back to Fishburn [9]. However, Hansson [10, p. 89] points out (quoting correspondence with Peter Fishburn) that Julian Blau was aware of the existence of non-dictatorial social welfare functions in the infinite-population case as early as 1960 without publishing this observation. Sen [14] and Suzumura [16] highlight the role played by the finiteness assumption in their respective methods of proving Arrow's theorem. Kirman and Sondermann [11] and Hansson [10] cast a new light on the structure of an Arrovian social welfare function with an infinite population, showing that the set of decisive coalitions for a social welfare function

W. Bossert (✉)
Department of Economics and CIREQ, University of Montreal, Montreal, QC, Canada H3C 3J7
e-mail: walter.bossert@umontreal.ca

K. Suzumura
School of Political Science and Economics, Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan
e-mail: ktr.suzumura@gmail.com

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*, Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_7

that satisfies Arrow's axioms of unlimited domain, weak Pareto and independence of irrelevant alternatives forms an ultrafilter. Their analyses apply to any (finite or infinite) population without any further structural assumptions. In an important contribution, Ferejohn and Page [8] enriched the infinite population social choice model by adding an intertemporal component. Time flows only unidirectionally, and any two distinct members of the society (or generations) are such that one generation appears in the society after the other. As a result of introducing this time structure of infinite population, Ferejohn and Page [8] provide a new link between Arrow's multi-profile approach to social choice and the theory of evaluating infinite intergenerational utility streams as initiated by Koopmans [12] and Diamond [7]. In the Koopmans-Diamond framework, the focus is on resource allocations among different generations with a fixed utility function for each generation. Thus, multi-profile considerations do not arise in this traditional setting.

Starting out with Hansson's [10] result on the ultrafilter structure of the set of decisive coalitions, Ferejohn and Page [8] propose a classical stationarity condition in an infinite-horizon multi-profile social choice model and show that if a social welfare function that satisfies their stationarity property in addition to Arrow's conditions exists, then generation one must be a dictator. Stationarity as defined by Ferejohn and Page demands that if a common first-period alternative is eliminated from two infinite streams of per-period alternatives, then the resulting continuation streams must be ranked in the same way as the original streams according to the social ranking obtained for the original profile. The reason why generation one is the only candidate for a dictator is the conjunction of the unidirectional nature of the flow of time and the resulting bias in favor of the first generation embodied in the stationarity property. Dictatorships of later generations cannot be stationary because we can only move forward but not backward in time.

As Ferejohn and Page [8] note themselves, the question whether such a social welfare function exists is left open by their analysis; they show that, conditional on its existence, a stationary social welfare function satisfying Arrow's [1, 1963 (2nd edn.); 2012 (3rd edn.)] axioms must be dictatorial with generation one being the dictator. Packel [13] resolved the existence issue by establishing a strong impossibility result: no collective choice rule that generates complete social rankings can satisfy unlimited domain, weak Pareto and stationarity. Neither transitivity of the social preference relations nor independence of irrelevant alternatives are needed for this result. Bossert and Suzumura [3] prove a slightly stronger version of Packel's [13] impossibility theorem by dropping completeness of the social relations from the list of requirements. It is possible to obtain further generalizations of this impossibility result. As is clear from its proof (which is provided later in the paper), only a single preference profile is required and, as a consequence, any domain that includes such a profile will produce the impossibility. For instance, Bossert and Suzumura [3] point out that the same conclusion holds if individual preferences are restricted to those that are history-independent.

In the face of this rather strong impossibility result, a question arises naturally: what modifications to the domain assumption or to the classical stationarity condition allow us to obtain possibility results? Packel [13] and Bossert and Suzumura

[3] choose two different paths in order to resolve the impossibility; see also Bossert and Suzumura [2, Chapter 10] for a discussion.

Packel's [13] approach consists of restricting the domain of a social welfare function to profiles where the individual preferences (or generation one's preferences) are themselves stationary. This domain assumption, which is plausible if social preferences are required to be stationary in Ferejohn and Page's [8] sense, allows for the existence of social welfare functions that satisfy the remaining axioms.

Bossert and Suzumura [3], on the other hand, consider an alternative domain assumption—namely, the assumption that each generation's preference relation is selfish in the sense that it depends on the per-period outcome for this generation only. This selfish domain also allows for the existence of social welfare functions that satisfy weak Pareto and classical stationarity. However, requiring independence of irrelevant alternatives or Pareto indifference in addition again generates impossibilities. The impossibility result involving independence does, as far as we are aware, not appear in the earlier literature. In order to circumvent these new impossibilities, the selfish domain assumption is supplemented by a modification of the stationarity axiom. Especially in the context of selfish preferences, it seems natural to consider a suitable multi-profile version of stationarity. Multi-profile stationarity requires that, for any two streams of per-period alternatives and for any preference profile, if the first-period alternatives are the same in the two streams, then the social ranking of the two streams according to this profile is the same as the social ranking that results if the common first-period alternative is removed along with the preference ordering of generation one. When combined, multi-profile stationarity and selfish domain allow for social welfare functions that also satisfy weak Pareto, independence of irrelevant alternatives and Pareto indifference. Moreover, these properties can be used to characterize the lexicographic dictatorship in which the generations are taken into consideration in chronological order.

Both approaches—employing the classical stationary domain or the selfish domain—allow for possibilities. However, the existence of a dictator (generation one) is implied under either of the two domain assumptions. Thus, although the infinite-population version of Arrow's social choice problem permits, in principle, non-dictatorial rules, these additional possibilities vanish in an intergenerational setting if the above-described notions of stationarity are imposed.

In this paper, we provide a brief survey of multi-profile intergenerational social choice as outlined above. After introducing the basic definitions, we provide a statement of Hansson's [10] ultrafilter theorem which is used in several of the subsequent results. This is followed by a discussion of the fundamental Ferejohn-Page [8] theorem and the impossibility established by Packel [13] and generalized by Bossert and Suzumura [3]. Then the two methods of modifying the domain or the stationarity axiom and their consequences are reviewed and further possibility, impossibility and characterization results are stated. This includes the new impossibility theorem involving selfish domains and independence of irrelevant alternatives alluded to above.

We provide full proofs whenever they are based on elementary methods. Hansson's [10] theorem is stated without a proof and we refer the reader to the

original article instead. Full proofs of results that rely on variants of the ultrafilter theorem are not given; this is the case because these variants would themselves require proofs due to the different domain assumptions employed. In these cases (to be precise, Packel's [13] possibility result and Bossert and Suzumura's [3] characterization), however, the proof ideas are outlined in some detail to explain the intuition underlying the respective result.

2 Intergenerational Social Choice and Decisiveness

Suppose there is a set of per-period alternatives X with $|X| \geq 3$. Let X^∞ be the set of all infinite streams of per-period alternatives $\mathbf{x} = (x_1, x_2, \ldots)$ where, for each generation $t \in \mathbb{N}$, $x_t \in X$ is the period-t alternative experienced by generation t.

The set of all binary relations on X^∞ is denoted by $\mathcal{B}$. An ordering is a reflexive, complete and transitive relation and the set of all orderings on X^∞ is denoted by $\mathcal{R}$. The asymmetric part and the symmetric part of a relation R are denoted by $P(R)$ and $I(R)$, respectively. Furthermore, for all $\mathbf{x}, \mathbf{y} \in X^\infty$ and for all $R \in \mathcal{B}$, $R|_{\{\mathbf{x},\mathbf{y}\}}$ is the restriction of R to the set $\{\mathbf{x}, \mathbf{y}\}$.

The preference ordering of generation $t \in \mathbb{N}$ is $R_t \in \mathcal{R}$. A (preference) profile is a stream $\mathbf{R} = (R_1, R_2, \ldots)$ of orderings on X^∞. The set of all such profiles is denoted by $\mathcal{R}^\infty$. Throughout the paper, we assume that individual preferences are orderings.

In the infinite-horizon context studied in this paper, a collective choice rule is a mapping $f : \mathcal{D} \to \mathcal{B}$, where $\mathcal{D} \subseteq \mathcal{R}^\infty$ with $\mathcal{D} \neq \emptyset$ is the domain of f. The interpretation is that, for a profile $\mathbf{R} \in \mathcal{D}$, $f(\mathbf{R})$ is the social ranking of streams in X^∞. If $f(\mathbf{R})$ is an ordering for all $\mathbf{R} \in \mathcal{D}$, f is a social welfare function.

Arrow's [1, 1963 (2nd edn.); 2012 (3rd edn.)] fundamental properties are an unlimited domain assumption, the weak Pareto principle and independence of irrelevant alternatives. These axioms are well-established and require no further discussion.

Unlimited Domain. $\mathcal{D} = \mathcal{R}^\infty$.

Weak Pareto. For all $\mathbf{x}, \mathbf{y} \in X^\infty$ and for all $\mathbf{R} \in \mathcal{D}$,

$$\mathbf{x} P(R_t) \mathbf{y} \text{ for all } t \in \mathbb{N} \Rightarrow \mathbf{x} P(f(\mathbf{R})) \mathbf{y}.$$

Independence of Irrelevant Alternatives. For all $\mathbf{x}, \mathbf{y} \in X^\infty$ and for all $\mathbf{R}, \mathbf{R}' \in \mathcal{D}$,

$$R_t|_{\{\mathbf{x},\mathbf{y}\}} = R'_t\big|_{\{\mathbf{x},\mathbf{y}\}} \text{ for all } t \in \mathbb{N} \Rightarrow f(\mathbf{R})|_{\{\mathbf{x},\mathbf{y}\}} = f(\mathbf{R}')\big|_{\{\mathbf{x},\mathbf{y}\}}.$$

A set $T \subseteq \mathbb{N}$ is decisive for a social welfare function f if and only if, for all $\mathbf{x}, \mathbf{y} \in X^{\infty}$ and for all $\mathbf{R} \in \mathcal{D}$,

$$\mathbf{x}P(R_t)\mathbf{y} \text{ for all } t \in T \Rightarrow \mathbf{x}P(f(\mathbf{R}))\mathbf{y}.$$

Clearly, $\mathbb{N}$ is decisive for any social welfare function f that satisfies weak Pareto. If there is a generation $t \in \mathbb{N}$ such that $\{t\}$ is decisive for f, generation t is a dictator for f and the social welfare function f is said to be dictatorial.

A filter on $\mathbb{N}$ is a collection $\mathcal{F}$ of subsets of $\mathbb{N}$ such that

1. $\emptyset \notin \mathcal{F}$;
2. $N \in \mathcal{F}$;
3. for all $T, T' \in \mathcal{F}$, $T \cap T' \in \mathcal{F}$;
4. for all $T, T' \subseteq \mathbb{N}$, $[[T \in \mathcal{F} \text{ and } T \subseteq T'] \Rightarrow T' \in \mathcal{F}]$.

An ultrafilter on $\mathbb{N}$ is a collection $\mathcal{U}$ of subsets of $\mathbb{N}$ such that

1. $\emptyset \notin \mathcal{U}$;
2. for all $T \subseteq \mathbb{N}$, $[T \in \mathcal{U} \text{ or } \mathbb{N} \setminus T \in \mathcal{U}]$;
3. for all $T, T' \in \mathcal{U}$, $T \cap T' \in \mathcal{U}$.

The conjunction of properties 1 and 2 in the definition of an ultrafilter implies that $\mathbb{N} \in \mathcal{U}$ and, furthermore, the conjunction of properties 1 and 3 implies that the disjunction in property 2 is exclusive—that is, T and $\mathbb{N} \setminus T$ cannot both be in $\mathcal{U}$.

An ultrafilter $\mathcal{U}$ is principal if and only if there exists a $t \in \mathbb{N}$ such that, for all $T \subseteq \mathbb{N}$, $T \in \mathcal{U}$ if and only if $t \in T$. Otherwise, $\mathcal{U}$ is a free ultrafilter. It can be verified easily that if $\mathbb{N}$ is replaced with a finite set, then the only ultrafilters are principal and, therefore, Hansson's theorem reformulated for finite populations reduces to Arrow's [1, 1963 (2nd edn.); 2012 (3rd edn.)] theorem—that is, there exists an individual (or a generation) t which is a dictator. In the infinite-population case, a set of decisive coalitions that is a principal ultrafilter corresponds to a dictatorship just as in the finite case. Unlike in the finite case, there also exist free ultrafilters but they cannot be defined explicitly; the proof of their existence relies on non-constructive methods in the sense of using variants of the axiom of choice. These free ultrafilters are non-dictatorial. However, social preferences associated with sets of decisive coalitions that form free ultrafilters fail to be continuous with respect to most standard topologies; see, for instance, Campbell [4–6].

Hansson [10] shows that if a social welfare function f satisfies unlimited domain, weak Pareto and independence of irrelevant alternatives, then the set of all decisive coalitions for f must be an ultrafilter. For future reference, we provide a statement of Hansson's [10] theorem formulated in the intertemporal context and refer the reader to the original paper for the proof of the more general result that applies to any population with at least two members.

Theorem 1 (Hansson [10]) *If a social welfare function f satisfies unlimited domain, weak Pareto and independence of irrelevant alternatives, then the set of all decisive coalitions for f is an ultrafilter.*

Hansson [10] also shows that, for any ultrafilter $\mathcal{U}$, there exists a social welfare function f that satisfies unlimited domain, weak Pareto and independence of irrelevant alternatives such that the set of decisive coalitions for f is equal to $\mathcal{U}$. Moreover, he provides a parallel analysis for situations where the transitivity requirement on social rankings is weakened to quasi-transitivity. In this case, the resulting sets of decisive coalitions are filters rather than ultrafilters; see Hansson [10] for details.

3 Classical Stationarity

Arrow's [1, 1963 (2nd edn.); 2012 (3rd edn.)] axioms introduced in the previous section are well-known from the relevant literature and require no further discussion. None of them, however, invoke the intertemporal structure of our model. Classical stationarity introduced by Ferejohn and Page [8], in contrast, is based on the unidirectional nature of time. The underlying idea is due to Koopmans [12] in a related but distinct context: if two streams of per-period alternatives agree in the first period, their relative social ranking is the same as that of their respective subsequences from period two onward. To formulate a property of this nature in a multi-profile setting, the profile under consideration for each of the two comparisons must be specified.

First, we introduce the definition of a stationary binary relation on X^∞. Let $t \in \mathbb{N}$. For $\mathbf{x} \in X^\infty$, the period-t continuation of $\mathbf{x}$ is

$$\mathbf{x}_{\geq t} = (x_t, x_{t+1}, \ldots),$$

that is, $(\mathbf{x}_{\geq t})_\tau = x_{\tau+t-1}$ for all $\tau \in \mathbb{N}$. Analogously, for $\mathbf{R} \in \mathcal{R}^\infty$, the period-$t$ continuation of $\mathbf{R}$ is

$$\mathbf{R}_{\geq t} = (R_t, R_{t+1}, \ldots).$$

A relation R on X^∞ is stationary if and only if, for all $\mathbf{x}, \mathbf{y} \in X^\infty$, if $x_1 = y_1$, then

$$\mathbf{x} R \mathbf{y} \Leftrightarrow \mathbf{x}_{\geq 2} R \mathbf{y}_{\geq 2}.$$

In Ferejohn and Page's [8] and Packel's [13] definitions of stationarity, the same profile is employed before and after the common first-period alternative is removed. This leads to the following axiom.

Classical Stationarity. For all $\mathbf{R} \in \mathcal{D}$, $f(\mathbf{R})$ is stationary.

Ferejohn and Page's [8] fundamental result establishes that if there exists a social welfare function f that satisfies unlimited domain, weak Pareto, independence of irrelevant alternatives and classical stationarity, then generation one must be a

dictator for f. As they clearly acknowledge, the existence issue itself remains unresolved by their theorem; however, this question of existence has been resolved in the meantime (see Bossert and Suzumura [3] and Packel [13]). To be very clear from the outset, let us state that there exists no collective choice rule (and, thus, no social welfare function) that satisfies unlimited domain, weak Pareto and classical stationarity; see Theorem 3 below. Thus, considerably more is known now about the issue at hand than at the time when Ferejohn and Page wrote their path-breaking paper. Nevertheless, the major purpose of the present contribution is to provide a survey of the relevant literature and, since the Ferejohn-Page theorem plays such a fundamental role, we consider it imperative to present their result in detail, in spite of the observation that stronger results are available nowadays. In our opinion, a proper account of the developments in this area would be incomplete without giving the seminal contribution of Ferejohn and Page [8] the credit and the appreciation that it deserves.

Theorem 2 (Ferejohn and Page [8]) *If a social welfare function f satisfies unlimited domain, weak Pareto, independence of irrelevant alternatives and classical stationarity, then generation one is a dictator for f.*

Proof The proof proceeds by showing that, given the axioms in the theorem statement, $\{1\}$ is decisive and, thus, generation one is a dictator for f.

By Theorem 1, the set of decisive coalitions is an ultrafilter. Let $x, y \in X$ and let $\succeq$ be an ordering on X such that

$$xP(\succeq)y.$$

Define the profile $\mathbf{R}$ as follows. Let, for all $\mathbf{x}, \mathbf{y} \in X^{\infty}$ and for all $t \in \mathbb{N}$,

$$\mathbf{x}R_t\mathbf{y} \Leftrightarrow x_t \succeq y_t.$$

Now consider the streams

$$\begin{aligned}
\mathbf{x} &= (x, x, y, x, y, x, \ldots);\\
\mathbf{y} &= (y, y, x, y, x, y, \ldots);\\
\mathbf{z} &= (x, y, x, y, x, y, \ldots) = \mathbf{x}_{\geq 2} = \mathbf{w}_{\geq 2};\\
\mathbf{w} &= (y, x, y, x, y, x, \ldots) = \mathbf{z}_{\geq 2} = \mathbf{y}_{\geq 2}.
\end{aligned}$$

We have $\mathbf{w}P(R_t)\mathbf{z}$ for all even $t \in \mathbb{N}$ and $\mathbf{z}P(R_t)\mathbf{w}$ for all odd $t \in \mathbb{N}$. By definition of an ultrafilter, either

$$\{t \in \mathbb{N} \mid t \text{ is even}\} \text{ is decisive} \tag{1}$$

or

$$\{t \in \mathbb{N} \mid t \text{ is odd}\} \text{ is decisive.} \tag{2}$$

If (1) is true, it follows that

$$\mathbf{w}P(f(\mathbf{R}))\mathbf{z}. \tag{3}$$

Because

$$\mathbf{w} = \mathbf{z}_{\geq 2} \text{ and } \mathbf{z} = \mathbf{x}_{\geq 2} \text{ and } (\mathbf{z}_{\geq 2})_1 = (\mathbf{x}_{\geq 2})_1 = x,$$

classical stationarity implies

$$\mathbf{z}P(f(\mathbf{R}))\mathbf{x}. \tag{4}$$

Analogously, because

$$\mathbf{w} = \mathbf{y}_{\geq 2} \text{ and } \mathbf{z} = \mathbf{w}_{\geq 2} \text{ and } (\mathbf{y}_{\geq 2})_1 = (\mathbf{w}_{\geq 2})_1 = y,$$

classical stationarity implies

$$\mathbf{y}P(f(\mathbf{R}))\mathbf{w}. \tag{5}$$

Using (5), (3) and (4), transitivity implies $\mathbf{y}P(f(\mathbf{R}))\mathbf{x}$. But $\mathbf{x}P(R_t)\mathbf{y}$ for all even $t \in \mathbb{N}$ and, thus, we obtain a contradiction to the decisiveness of $\{t \in \mathbb{N} \mid t \text{ is even}\}$. Therefore, (1) is false and (2) must apply.

By (2),

$$\mathbf{z}P(f(\mathbf{R}))\mathbf{w}. \tag{6}$$

Because

$$\mathbf{z} = \mathbf{x}_{\geq 2} \text{ and } \mathbf{w} = \mathbf{z}_{\geq 2} \text{ and } (\mathbf{x}_{\geq 2})_1 = (\mathbf{z}_{\geq 2})_1 = x,$$

classical stationarity implies

$$\mathbf{x}P(f(\mathbf{R}))\mathbf{z}. \tag{7}$$

Analogously, because

$$\mathbf{z} = \mathbf{w}_{\geq 2} \text{ and } \mathbf{w} = \mathbf{y}_{\geq 2} \text{ and } (\mathbf{w}_{\geq 2})_1 = (\mathbf{y}_{\geq 2})_1 = y,$$

classical stationarity implies

$$\mathbf{w}P(f(\mathbf{R}))\mathbf{y}. \tag{8}$$

Using (7), (6) and (8), transitivity implies $\mathbf{x}P(f(\mathbf{R}))\mathbf{y}$. We have

$$\{t \in \mathbb{N} \mid \mathbf{x}P(R_t)\mathbf{y}\} = \{1\} \cup \{t \in \mathbb{N} \mid t \text{ is even}\}$$

and, thus, the complement of this set cannot be decisive. Therefore,

$$\{1\} \cup \{t \in \mathbb{N} \mid t \text{ is even}\}$$

is decisive and, by property 3 of an ultrafilter, it follows that

$$\{1\} = \{t \in \mathbb{N} \mid t \text{ is odd}\} \cap (\{1\} \cup \{t \in \mathbb{N} \mid t \text{ is even}\})$$

is decisive, and the proof is complete. ■

Packel's [13] answers the existence question left open by Ferejohn and Page [8] in the negative by showing that there does not exist any collective choice rule that generates complete social rankings and satisfies unlimited domain, weak Pareto and classical stationarity. Bossert and Suzumura [3] slightly strengthen Packel's [13] impossibility result by dropping the completeness assumption.

Theorem 3 (Bossert and Suzumura [3], Packel [13]) *There exists no collective choice rule f that satisfies unlimited domain, weak Pareto and classical stationarity.*

Proof Suppose f is a collective choice rule that satisfies the axioms of the theorem statement. Let $x, y \in X$.

For each odd $t \in \mathbb{N}$, let $\succeq_t$ be an antisymmetric ordering on X such that

$$yP(\succeq_t)x.$$

For each even $t \in \mathbb{N}$, let $\succeq_t$ be an antisymmetric ordering on X such that

$$xP(\succeq_t)y.$$

Define a profile $\mathbf{R}$ as follows. For all $\mathbf{x}, \mathbf{y} \in X^\infty$, let

$$\mathbf{x}P(R_1)\mathbf{y} \;\Leftrightarrow\; x_1P(\succeq_1)y_1 \;\text{ or }\; [x_1 = y_1 \;\text{ and }\; x_3P(\succeq_1)y_3].$$

Now let, for all $\mathbf{x}, \mathbf{y} \in X^\infty$,

$$\mathbf{x}R_1\mathbf{y} \;\Leftrightarrow\; \neg\,\mathbf{y}P(R_1)\mathbf{x}.$$

For all $t \in \mathbb{N} \setminus \{1\}$ and for all $\mathbf{x}, \mathbf{y} \in X^\infty$, let

$$\mathbf{x}R_t\mathbf{y} \;\Leftrightarrow\; x_t \succeq_t y_t.$$

Now consider the streams

$$\mathbf{x} = (x, x, y, x, y, x, \ldots);$$
$$\mathbf{y} = (x, y, x, y, x, y, \ldots) = \mathbf{x}_{\geq 2};$$
$$\mathbf{z} = (y, x, y, x, y, x, \ldots) = \mathbf{y}_{\geq 2}.$$

We have $\mathbf{x}P(R_t)\mathbf{y}$ for all $t \in \mathbb{N}$ and, by weak Pareto, $\mathbf{x}P(f(\mathbf{R}))\mathbf{y}$. Stationarity implies $\mathbf{y}P(f(\mathbf{R}))\mathbf{z}$. But $\mathbf{z}P(R_t)\mathbf{y}$ for all $t \in \mathbb{N}$, and we obtain a contradiction to weak Pareto. ■

The result established in the above theorem can be strengthened by restricting the domain. Note that only a single profile is used in its proof and, thus, the impossibility remains valid if the domain is restricted to any subset containing this specific profile. For instance, Bossert and Suzumura [3] phrase the result by using a forward-looking domain such that each generation t compares any two streams exclusively on the basis of their period-t continuations.

Based on the observations of this section, there appear to be two natural ways to proceed in order to arrive at possibility results.

The first of these, due to Packel [13], is discussed in the following section. Packel [13] retains the classical stationarity assumption throughout his analysis. To resolve the impossibility, he employs domains that only contain stationary individual preferences (or domains that only include profiles in which generation one's preference ordering must be stationary and the orderings of all other generations may be arbitrary).

The second approach involves replacing classical stationarity with a multi-profile variant of stationarity and an alternative domain restriction due to Bossert and Suzumura [3]. Multi-profile stationarity differs from classical stationarity in that not only common first-period outcomes are removed from two streams but also the preference ordering of the first generation. Bossert and Suzumura also employ selfish domains, that is, domains such that each generation's preference ordering depends on this generation's per-period outcomes only. We illustrate the consequences of using multi-profile stationarity in a setting with selfish domains in the final section of the paper.

4 Classical Stationary Domains

The classical stationary domain $\mathcal{R}_C^\infty$ is composed of all profiles $\mathbf{R} \in \mathcal{R}^\infty$ such that R_t is stationary for each $t \in \mathbb{N}$. The resulting domain assumption is used by Packel [13].

Classical Stationary Domain. $\mathcal{D} = \mathcal{R}_C^\infty$.

Packel's [13] possibility result establishes that classical stationarity is compatible with weak Pareto and independence of irrelevant alternatives on classical stationary domains even if the collective choice rule is required to produce social orderings.

Theorem 4 (Packel [13]) *There exists a social welfare function f that satisfies classical stationary domain, weak Pareto, independence of irrelevant alternatives and classical stationarity.*

Proof An example is sufficient to prove the theorem. Let, for all $\mathbf{R} \in \mathcal{R}^{\infty}$, $f(\mathbf{R}) = R_1$. Because R_1 is stationary by the domain assumption, it follows immediately that all the axioms of the theorem statement are satisfied. ■

As is evident from Packel's [13, p. 223] formulation of the result, the possibility survives if the domain is expanded by allowing the preference orderings of all generations other than generation one to be arbitrary; what matters is that generation one's preferences are stationary.

Although the above theorem provides a possibility result, the example invoked is not very promising—it involves a dictatorship of generation one. Indeed, this is no coincidence; as Packel [13] shows, even if the domain is restricted so as to allow no preferences other than stationary ones for all generations, generation-one dictatorships are the only social welfare functions that satisfy weak Pareto, independence of irrelevant alternatives and classical stationarity. This observation is parallel to that of Ferejohn and Page [8] but, in the case of the following theorem, existence is not an issue as Theorem 4 illustrates.

Theorem 5 (Packel [13]) *If a social welfare function f satisfies classical stationary domain, weak Pareto, independence of irrelevant alternatives and classical stationarity, then generation one is a dictator for f.*

Sketch of Proof The proof of this theorem is similar to that of Theorem 2. However, it is necessary to prove a variant of Hansson's [10] result (Theorem 1) that applies to the classical stationary domain as opposed to the unlimited domain. Once this is accomplished, the remaining steps parallel those employed in the proof of Theorem 2. ■

5 Selfish Domains and Multi-Profile Stationarity

The selfish domain $\mathcal{R}_S^{\infty}$ is obtained by letting, for all $\mathbf{R} \in \mathcal{R}^{\infty}$, $\mathbf{R} \in \mathcal{R}_S^{\infty}$ if and only if, for each $t \in \mathbb{N}$, there exists an ordering $\succeq_t$ on X such that, for all $\mathbf{x}, \mathbf{y} \in X^{\infty}$,

$$\mathbf{x} R_t \mathbf{y} \Leftrightarrow x_t \succeq_t y_t.$$

Selfish Domain. $\mathcal{D} = \mathcal{R}_S^{\infty}$.

As an alternative to Theorem 4, we can replace unlimited domain with selfish domain instead of classical stationary domain in order to obtain a possibility result.

Theorem 6 (Bossert and Suzumura [3]) *There exists a social welfare function f that satisfies selfish domain, weak Pareto and classical stationarity.*

Proof Again, an example is sufficient to prove this theorem. Suppose $(\succeq_1, \succeq_2, \ldots)$ is the profile of orderings on X associated with the selfish profile $\mathbf{R} \in \mathcal{R}_S^\infty$ of orderings on X^∞. Define a social welfare function f by letting, for all $\mathbf{x}, \mathbf{y} \in X^\infty$ and for all $\mathbf{R} \in \mathcal{R}_S^\infty$, $\mathbf{x} f(\mathbf{R}) \mathbf{y}$ if and only if

$$[x_\tau I(\succeq_1) y_\tau \text{ for all } \tau \in \mathbb{N}] \text{ or}$$
$$[\text{there exists } t \in \mathbb{N} \text{ such that } x_\tau I(\succeq_1) y_\tau \text{ for all } \tau < t \text{ and } x_t P(\succeq_1) y_t].$$

That f satisfies selfish domain follows immediately by definition.

To see that weak Pareto is satisfied, note first that, according to f as defined above,

$$\mathbf{x} P(R_t) \mathbf{y}$$

is equivalent to

$$x_t P(\succeq_t) y_t$$

for all $t \in \mathbb{N}$. Thus, $x_\tau I(\succeq_1) y_\tau$ for all $\tau < 1$ and $x_1 P(\succeq_1) y_1$ (the indifference relations are vacuously true because the set of periods τ such that $\tau < 1$ is empty). By definition of f, it follows that $\mathbf{x} P(f(\mathbf{R})) \mathbf{y}$.

It remains to establish that f satisfies classical stationarity. Suppose the alternatives $\mathbf{x}, \mathbf{y} \in X^\infty$ and the stationary profile $\mathbf{R} \in \mathcal{R}_S^\infty$ are such that $x_1 = y_1$. Thus, because $\succeq_1$ is an ordering (and, thus, reflexive), we obtain $x_1 I(\succeq_1) y_1$. It follows that

$$\begin{aligned}
\mathbf{x}_{\geq 2} f(\mathbf{R}) \mathbf{y}_{\geq 2} \Leftrightarrow\ & [x_\tau I(\succeq_1) y_\tau \text{ for all } \tau \in \mathbb{N} \setminus \{1\} \text{ and } x_1 I(\succeq_1) y_1] \text{ or} \\
& [\text{there exists } t \in \mathbb{N} \setminus \{1\} \text{ such that } x_\tau I(\succeq_1) y_\tau \text{ for all } \tau < t \\
& \text{and } x_t P(\succeq_1) y_t] \\
\Leftrightarrow\ & [x_\tau I(\succeq_1) y_\tau \text{ for all } \tau \in \mathbb{N}] \text{ or} \\
& [\text{there exists } t \in \mathbb{N} \text{ such that } x_\tau I(\succeq_1) y_\tau \text{ for all } \tau < t \\
& \text{and } x_t P(\succeq_1) y_t] \\
\Leftrightarrow\ & \mathbf{x} f(\mathbf{R}) \mathbf{y}.
\end{aligned}$$

■

Unlike Theorem 4, the statement of Theorem 6 does not include independence of irrelevant alternatives as one of the axioms that can be satisfied under the alternative domain assumption. In fact, adding the independence property to the list of axioms leads to another impossibility. This is a new observation that, to the best of our knowledge, does not appear in the previous literature.

Theorem 7 *There exists no collective choice rule f that satisfies selfish domain, weak Pareto, independence of irrelevant alternatives and classical stationarity.*

Proof Suppose f is a collective choice rule that satisfies the axioms of the theorem statement. Let $x, y, z \in X$.

For each $t \in \mathbb{N}$, let $\succeq_t$ be an ordering on X such that

$$yP(\succeq_t)x \text{ and } xP(\succeq_t)z.$$

Furthermore, for each odd $t \in \mathbb{N}$, let $\succeq'_t$ be an ordering on X such that

$$yP(\succeq'_t)z \text{ and } zP(\succeq'_t)x.$$

Finally, for each even $t \in \mathbb{N}$, let $\succeq'_t$ be an ordering on X such that

$$xP(\succeq'_t)y \text{ and } yP(\succeq'_t)z.$$

Define two profiles $\mathbf{R}$ and $\mathbf{R}'$ as follows. For all $\mathbf{x}, \mathbf{y} \in X^\infty$ and for all $t \in \mathbb{N}$, let

$$\mathbf{x}R_t\mathbf{y} \;\Leftrightarrow\; x_t \succeq_t y_t$$

and

$$\mathbf{x}R'_t\mathbf{y} \;\Leftrightarrow\; x_t \succeq'_t y_t.$$

Clearly, the profiles thus defined are in $\mathcal{R}_S^\infty$.

Now consider the streams

$$\begin{aligned}
\mathbf{x} &= (z, x, y, x, y, x, y, \ldots);\\
\mathbf{y} &= (z, z, x, z, x, z, x, \ldots);\\
\mathbf{z} &= (x, y, x, y, x, y, x, \ldots) = \mathbf{x}_{\geq 2};\\
\mathbf{w} &= (z, x, z, x, z, x, z, \ldots) = \mathbf{y}_{\geq 2}.
\end{aligned}$$

We have $\mathbf{z}P(R_t)\mathbf{w}$ for all $t \in \mathbb{N}$ and, by weak Pareto, $\mathbf{z}P(f(\mathbf{R}))\mathbf{w}$. Classical stationarity implies

$$\mathbf{x}P(f(\mathbf{R}))\mathbf{y}. \tag{9}$$

Furthermore, $\mathbf{w}P(R'_t)\mathbf{z}$ for all $t \in \mathbb{N}$ and, using weak Pareto again, $\mathbf{w}P(f(\mathbf{R}'))\mathbf{z}$. Classical stationarity implies

$$\mathbf{y}P(f(\mathbf{R}'))\mathbf{x}. \tag{10}$$

But we also have

$$\mathbf{x}I_1\mathbf{y} \text{ and } \mathbf{x}I'_1\mathbf{y}$$

and

$$\mathbf{x}P_t\mathbf{y} \text{ and } \mathbf{x}P'_t\mathbf{y}$$

for all $t \in \mathbb{N} \setminus \{1\}$ which, by independence of irrelevant alternatives, requires that

$$\mathbf{x}f(\mathbf{R})\mathbf{y} \Leftrightarrow \mathbf{x}f(\mathbf{R}')\mathbf{y},$$

contradicting the conjunction of (9) and (10). ■

Yet another impossibility emerges if Pareto indifference is used instead of independence of irrelevant alternatives. Pareto indifference is the analogue of weak Pareto that is obtained by replacing each occurrence of a strict preference with an indifference.

Pareto Indifference. For all $\mathbf{x}, \mathbf{y} \in X^\infty$ and for all $\mathbf{R} \in \mathcal{D}$,

$$\mathbf{x}I(R_t)\mathbf{y} \text{ for all } t \in \mathbb{N} \Rightarrow \mathbf{x}I(f(\mathbf{R}))\mathbf{y}.$$

Replacing independence of irrelevant alternatives with Pareto indifference leaves the incompatibility stated in the previous theorem intact, as shown by Bossert and Suzumura [3].

Theorem 8 (Bossert and Suzumura [3]) *There exists no collective choice rule f that satisfies selfish domain, weak Pareto, Pareto indifference and classical stationarity.*

Proof Suppose f is a collective choice rule that satisfies the axioms of the theorem statement. Let $x, y, z \in X$.

For each odd $t \in \mathbb{N}$, let $\succeq_t$ be an ordering on X such that

$$zP(\succeq_t)x \text{ and } xI(\succeq_t)y.$$

For each even $t \in \mathbb{N}$, let $\succeq_t$ be an ordering on X such that

$$zI(\succeq_t)x \text{ and } xP(\succeq_t)y.$$

Define a profile $\mathbf{R}$ as follows. For all $\mathbf{x}, \mathbf{y} \in X^{\infty}$ and for all $t \in \mathbb{N}$, let

$$\mathbf{x} R_t \mathbf{y} \;\Leftrightarrow\; x_t \succeq_t y_t.$$

Clearly, the profile thus defined is in $\mathcal{R}_S^{\infty}$.

Now consider the streams

$$\mathbf{x} = (z, z, x, z, x, z, x, \ldots);$$
$$\mathbf{y} = (z, x, y, x, y, x, y, \ldots);$$
$$\mathbf{z} = (z, x, z, x, z, x, \ldots) = \mathbf{x}_{\geq 2};$$
$$\mathbf{w} = (x, y, x, y, x, y, \ldots) = \mathbf{y}_{\geq 2}.$$

We have $\mathbf{x} I(R_t) \mathbf{y}$ for all $t \in \mathbb{N}$ and, by Pareto indifference, $\mathbf{x} I(f(\mathbf{R})) \mathbf{y}$. Classical stationarity implies $\mathbf{z} I(f(\mathbf{R})) \mathbf{w}$. But $\mathbf{z} P(R_t) \mathbf{w}$ for all $t \in \mathbb{N}$, and we obtain a contradiction to weak Pareto. ■

The two theorems just established suggest that the selfish domain assumption can only yield satisfactory possibilities if the classical stationary assumption is amended as well. There is a plausible alternative version according to which the (common) first-period component is eliminated not only from the streams to be compared but also from the profile for which the social ranking is to be determined. The resulting axiom, which appears to be suitable in conjunction with the path chosen by focusing on selfish preferences, is due to Bossert and Suzumura [3].

Multi-Profile Stationarity. For all $\mathbf{x}, \mathbf{y} \in X^{\infty}$ and for all $\mathbf{R} \in \mathcal{D}$, if $x_1 = y_1$, then

$$\mathbf{x} f(\mathbf{R}) \mathbf{y} \;\Leftrightarrow\; \mathbf{x}_{\geq 2} f(\mathbf{R}_{\geq 2}) \mathbf{y}_{\geq 2}.$$

If multi-profile stationarity is used instead of classical stationarity, both independence of irrelevant alternatives and Pareto indifference can be accommodated in addition to selfish domain and weak Pareto. A social welfare function that satisfies multi-profile stationarity in conjunction with selfish domain, weak Pareto, independence of irrelevant alternatives and Pareto indifference is the chronological dictatorship, to be introduced shortly. In fact, the chronological dictatorship is the only social welfare function satisfying this list of axioms and, thus, it can be characterized by means of this set of properties.

The chronological dictatorship is the social welfare function f defined by letting, for all $\mathbf{x}, \mathbf{y} \in X^{\infty}$ and for all $\mathbf{R} \in \mathcal{R}_S^{\infty}$, $\mathbf{x} f(\mathbf{R}) \mathbf{y}$ if and only if

$$[x_\tau I(\succeq_\tau) y_\tau \;\text{ for all } \tau \in \mathbb{N}] \;\text{ or}$$
$$[\text{there exists } t \in \mathbb{N} \text{ such that } x_\tau I(\succeq_\tau) y_\tau \;\text{ for all } \tau < t \;\text{ and }\; x_t P(\succeq_t) y_t].$$

The chronological dictatorship is, evidently, a special case of a dictatorial social welfare function and, thus, it turns out that the two alternative paths towards a resolution of Ferejohn and Page's [8] impossibility lead to similar results. Both Packel's [13] approach based on stationary individual preferences and Bossert and Suzumura's [3] attempt to use selfish individual preferences in conjunction with a new version of stationarity allow for the existence of social welfare functions with the desired properties. But, due to the bias in favor of generation one that is imposed by either form of stationarity (in conjunction with the unidirectional nature of the flow of time), the resulting rules must be dictatorial with generation one being the dictator. We conclude this survey with a statement and proof sketch of Bossert and Suzumura's [3] characterization.

Theorem 9 (Bossert and Suzumura [3]) *A social welfare function f satisfies selfish domain, weak Pareto, independence of irrelevant alternatives, Pareto indifference and multi-profile stationarity if and only if f is the chronological dictatorship.*

Sketch of Proof That the chronological dictatorship satisfies the axioms of the theorem statement is straightforward to verify.

In order to prove the reverse implication, a version of Hansson's [10] ultrafilter theorem (the theorem stated as Theorem 1 in Sect. 2 of the present paper) that applies to the selfish domain needs to be established, as is the case for Theorem 5. However, in the selfish case, Pareto indifference is required as an additional axiom. A modification of this nature is called for because the selfish domain is not sufficiently rich to generate arbitrary rankings of all streams. For example, whenever we have two streams $\mathbf{x}$ and $\mathbf{y}$ such that $x_t = y_t$ for some selfish generation $t \in \mathbb{N}$, this selfish generation must declare $\mathbf{x}$ and $\mathbf{y}$ indifferent; this is an immediate consequence of the conjunction of selfish domain and reflexivity. This addition of Pareto indifference to the list of axioms is necessitated by the observation that a fundamental preliminary result—an adaptation of Sen's [15, p. 4] field expansion lemma to our selfish domain setting—fails to be true if merely selfish domain, weak Pareto and independence of irrelevant alternatives are imposed. Loosely speaking, the field expansion lemma establishes that a decisiveness property over a given pair of alternatives can be expanded to all pairs of alternatives, thus producing full decisiveness from a weaker version that is restricted to a pair.

The proof of Theorem 9 consists of the following steps.

First, a version of the field expansion lemma for the selfish domain is proven, provided that f satisfies weak Pareto, independence of irrelevant alternatives and Pareto indifference.

In a second step, this result is used to establish a version of Hansson's [10] ultrafilter theorem that applies to the selfish domain. Again, Pareto indifference is required in order to invoke the above-described variant of the field expansion lemma.

The third step consists of showing that the axioms imply that generation one must be a dictator for f. This step parallels the corresponding step in the proofs of Theorems 2 and 5, except that multi-profile stationarity is used instead of classical stationarity.

Finally, the observation that generation one is a dictator is used to show that only the chronological dictatorship satisfies the required axioms. Because the proof method employed in this step does not appear in any of the proofs outlined earlier, we provide the details.

Because f is assumed to be a social welfare function (and, thus, produces social orderings for all profiles in its domain), it is sufficient to show that, for all $\mathbf{x}, \mathbf{y} \in X^\infty$ and for all $\mathbf{R} \in \mathcal{R}_S^\infty$, $\mathbf{x} P(f(\mathbf{R}))\mathbf{y}$ whenever $\mathbf{x}$ is strictly preferred to $\mathbf{y}$ according to the chronological dictatorship (the corresponding implication involving indifference is trivially satisfied because of Pareto indifference).

Suppose $t \in \mathbb{N}$, $\mathbf{x}, \mathbf{y} \in X^\infty$ and $\mathbf{R} \in \mathcal{R}_S^\infty$ are such that

$$x_\tau I(\succeq_\tau) y_\tau \text{ for all } \tau < t \text{ and } x_t P(\succeq_t) y_t.$$

If $t = 1$, let $\mathbf{z} = \mathbf{y}$; if $t \geq 2$, let $\mathbf{z} = (x_1, \ldots, x_{t-1}, \mathbf{y}_{\geq t})$. By Pareto indifference, $\mathbf{y} I(f(\mathbf{R}))\mathbf{z}$. Transitivity implies

$$\mathbf{x} f(\mathbf{R})\mathbf{y} \Leftrightarrow \mathbf{x} f(\mathbf{R})\mathbf{z}.$$

Together with the application of multi-profile stationarity $t-1$ times and noting that $\mathbf{z}_{\geq t} = \mathbf{y}_{\geq t}$, we obtain

$$\mathbf{x} f(\mathbf{R})\mathbf{y} \Leftrightarrow \mathbf{x} f(\mathbf{R})\mathbf{z} \Leftrightarrow \mathbf{x}_{\geq t} f(\mathbf{R}_{\geq t})\mathbf{z}_{\geq t} \Leftrightarrow \mathbf{x}_{\geq t} f(\mathbf{R}_{\geq t})\mathbf{y}_{\geq t}. \tag{11}$$

Because generation one is a dictator for f as established in the previous step, the relative ranking of $\mathbf{x}_{\geq t}$ and $\mathbf{y}_{\geq t}$ according to $\mathbf{R}_{\geq t}$ is determined by the strict preference for $\mathbf{x}$ over $\mathbf{y}$ according to the first generation in the profile $\mathbf{R}_{\geq t}$ (which is generation t in $\mathbf{R}$), so that $\mathbf{x}_{\geq t} P(f(\mathbf{R}_{\geq t}))\mathbf{y}_{\geq t}$ and, by (11), $\mathbf{x} P(f(\mathbf{R}))\mathbf{y}$. ■

Acknowledgements This paper is dedicated to Nick Baigent in appreciation of his invaluable contribution to the academic community. We thank Yongsheng Xu and two referees for their thoughtful comments and suggestions for improvements. Financial support from a Grant-in-Aid for Specially Promoted Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan for the *Project on Economic Analysis of Intergenerational Issues* (grant number 22000001), the Fonds de Recherche sur la Société et la Culture of Québec and the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

References

1. Arrow KJ (1951) Social choice and individual values. Wiley, New York; 2nd edn 1963; 3rd edn, Yale University Press, New Haven, 2012
2. Bossert W, Suzumura K (2010) Consistency, choice, and rationality. Harvard University Press, Cambridge, MA
3. Bossert W, Suzumura K (2011) Multi-profile intergenerational social choice. Soc Choice Welf 37:493–509

4. Campbell DE (1990) Intergenerational social choice without the Pareto principle. J Econ Theory 50:414–423
5. Campbell DE (1992) Quasitransitive intergenerational social choice for economic environments. J Math Econ 21:229–247
6. Campbell DE (1992) Equity, efficiency, and social choice. Clarendon Press, Oxford
7. Diamond P (1965) The evaluation of infinite utility streams. Econometrica 33:170–177
8. Ferejohn J, Page T (1978) On the foundations of intertemporal choice. Am J Agric Econ 60:269–275
9. Fishburn PC (1970) Arrow's impossibility theorem: concise proof and infinite voters. J Econ Theory 2:103–106
10. Hansson B (1976) The existence of group preference functions. Public Choice 38:89–98
11. Kirman AP, Sondermann D (1972) Arrow's theorem, many agents, and invisible dictators. J Econ Theory 5:267–277
12. Koopmans TC (1960) Stationary ordinal utility and impatience. Econometrica 28:287–309
13. Packel E (1980) Impossibility results in the axiomatic theory of intertemporal choice. Public Choice 35:219–227
14. Sen AK (1979) Personal utilities and public judgements: or what's wrong with welfare economics? Econ J 89:537–558. Reprinted in: Sen AK (1982) Choice, welfare and measurement, pp 327–352. Basil Blackwell, Oxford
15. Sen AK (1995) Rationality and social choice. Am Econ Rev 85:1–24. Reprinted in: Sen AK (2002) Rationality and freedom, pp 261–299. The Belknap Press/Harvard University Press, Cambridge, MA
16. Suzumura K (2000) Welfare economics beyond welfarist-consequentialism. Jpn Econ Rev 51:1–32

Minimal Maskin Monotonic Extensions of Tournament Solutions

İpek Özkal-Sanver, Pelin Pasin, and M. Remzi Sanver

Abstract In this paper we give a general characterization of the minimal Maskin monotonic extensions of Condorcet consistent tournament solutions. We then compute the minimal Maskin monotonic extensions for the following rules: The top-cycle, the uncovered set, the iterated uncovered set, the minimal covering set and the Copeland rule. Moreover, we characterize the minimal Maskin monotonic extensions of the social choice rules that are generated by the top-cycle, the uncovered set, the iterated uncovered set, and the minimal covering set via the majority rule. We also give results establishing the relation between the minimal Maskin monotonic extensions in the tournament environment and the social choice environment.

Keywords Condorcet consistency • Minimal Maskin monotoic extensions • Tournament solutions

1 Introduction

Maskin [7] in his seminal papers shows that a certain type of monotonicity is necessary for social choice rules to be Nash implementable.[1] However, Maskin monotonicity is a fairly strong condition which many social choice rules fail to

[1]It is worth mentioning two other notions of monotonicity that were introduced in the social choice framework and then adopted to tournaments. The first monotonicity condition is introduced by Moulin [10] (often called as Moulin monotonicity) and starting from a preference profile, it considers an improvement of an alternative while the rest kept unchanged. The second one, cover monotonicity, is introduced by Özkal-Sanver and Sanver [12] and considers an improvement of an alternative while the lower contour sets of other selected alternatives remain unchanged.

İ. Özkal-Sanver (✉) • M. Remzi Sanver
Department of Economics, Murat Sertel Center for Advanced Economic Studies,
İstanbul Bilgi University, İstanbul, Turkey
e-mail: isanver@bilgi.edu.tr; sanver@bilgi.edu.tr

P. Pasin
Department of Economics, İzmir Katip Çelebi University, Izmir, Turkey
e-mail: pelin.pasin@ikc.edu.tr

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_8

satisfy. In particular, Muller and Satterthwaite [11] shows that Maskin monotonicity is equivalent to dictatoriality when the social choice rule is citizen sovereign and singleton-valued.

Sen [14] proposes a way of evaluating the extent of non-monotonicity of social choice functions, by extending them minimally to social choice correspondences which are Maskin monotonic. Since then, the concept has been applied to a variety of frameworks, such as Erdem and Sanver [4] who characterizes the minimal Maskin monotonic extension of scoring rules, Kara and Sönmez [5] who apply it to matching problems and Thomson [16] who applies it to allocation problems.

We carry the concept to the framework of tournaments and compute the minimal Maskin monotonic extensions of four well-known tournament solutions.[2] Our findings are directly related to social choice rules, as Arrovian social welfare functions cannot avoid cyclic social preferences, hence requiring one to choose from a tournament. To establish this relation, we make a twofold analysis: The direct analysis conceives a tournament solution as a mapping from tournaments to alternatives while the indirect analysis conceives it as a social choice rule which maps a preference profile into alternatives via choosing from the tournament which is the majority relation induced by that preference profile. We interrelate the two analysis by establishing the equivalence between the Maskin monotonicity of a tournament solution and the Maskin monotonicity of the social choice rule induced by that tournament solution.

We know from Özkal-Sanver and Sanver [13] that no social choice rule which is generated by Condorcet consistent tournament solutions is Maskin monotonic. In this paper, we show that Condorcet consistent tournament solutions themselves fail Maskin monotonicity. We characterize the minimal Maskin monotonic extensions of neutral Condorcet consistent social choice rules. In our characterization, whether an alternative is in the extension or not depends on the number of alternatives beaten by it. We compute the minimal Maskin monotonic extensions of the top-cycle and four of its well-known refinements, namely the uncovered set, the iterated uncovered set, the minimal covering set and the Copeland rule [2]. Interestingly, the equivalence for Maskin monotonicity under the direct and indirect analysis does not carry to minimal Maskin monotonic extensions. Under the direct analysis, the Maskin monotonic extensions of top-cycle, uncovered set, the iterated uncovered set and the minimal covering set coincide in a rather coarse tournament solution which selects all alternatives but the Condorcet looser while for the Copeland rule we have a less coarse solution. For social choice rules that are generated by Condorcet consistent solutions the minimal Maskin monotonic extensions coincide for uncovered set, the iterated uncovered set and the minimal covering set which turn out to be less coarse than the top-cycle. The minimal Maskin monotonic extension of the social choice rule which is generated by the Copeland rule is not identified in this paper.

[2]For a thorough account of the literature see Laslier [6] which contains a variety of tournament solutions tested against a variety of properties.

Our analysis and findings are inspired by the literature which we owe to Professor Nicholas Baigent. Especially, Baigent and Klamler [1] addresses similar intransitivity problems and comparisons arising from simple majority rule in an environment with weak orders while we consider strict orders.

2 Basic Notions

Let X be a set with $\#X \geq 3$. A *tournament* T on X is a complete and asymmetric binary relation on X.[3] The set of all tournaments on X is denoted by $\mathscr{T}(X)$. Let $\underline{X}$ stand for all the nonempty subsets of X, $\underline{X} = 2^X \backslash \{\emptyset\}$. A *tournament solution* is a mapping $C : \mathscr{T}(X) \to \underline{X}$ which assigns each tournament a nonempty subset of X. A nonempty subset Y of X constitutes a *cycle* with respect to T if $Y = \{x_1, \ldots, x_{\#Y}\}$ is such that for all $i \in \{1, \ldots, \#Y - 1\}$, $x_i T x_{i+1}$ and $x_{\#Y} T x_1$. The *top-cycle* of a tournament T is a cycle Y such that for all $y \in Y$ and for all $x \in X \backslash Y$, yTx. The *Condorcet winner* of a tournament T is $x \in X$ such that xTy for all $y \in X \backslash \{x\}$. Each tournament has either a unique top-cycle or a unique Condorcet winner. Let $\gamma : \mathscr{T}(X) \to \underline{X}$ be the solution which assigns to each T its Condorcet winner or the top-cycle. A solution $C : \mathscr{T}(X) \to \underline{X}$ is Condorcet consistent if for all $T \in \mathscr{T}(X)$, $C(T) \subseteq \gamma(T)$. In this paper we will consider Condorcet consistent solutions.

The *lower contour set* of $x \in X$ at tournament T is $L(x, T) = \{y \in X : xTy\}$. A tournament solution C satisfies *Maskin monotonicity* if for any $T, T' \in \mathscr{T}(X)$ and $x \in C(T)$, $L(x, T) \subseteq L(x, T')$ implies $x \in C(T')$.[4]

3 Minimal Monotonic Extensions of Tournament Solutions

Maskin monotonicity is a necessary condition for Nash implementation which many social choice rules fail to satisfy. Özkal-Sanver and Sanver [13] shows that no social choice rule which is generated by a Condorcet consistent tournament solution is

[3] We use completeness in the weak sense; for all $x, y \in X$ either xTy or yTx holds and xTx for no $x \in X$, i.e., T is irreflexive.

[4] Adaptation of Moulin monotonicity for tournament solutions is defined in Laslier [6]. While the iterated uncovered set fail to satisfy it, many well known tournament solutions satisfy Moulin monotonicity. Cover monotonicity for tournament solutions is defined in Özkal-Sanver and Sanver [13] and successfully discriminates among the main tournament solutions.

Maskin monotonic and in the following theorem we establish that no Condorcet consistent tournament solution is Maskin monotonic.[5]

Theorem 1 *No Condorcet consistent tournament solution is Maskin monotonic.*

Proof Let C be a Condorcet consistent solution, $C \subseteq \gamma$, and $T \in \mathscr{T}(X)$ be a tournament such that the top-cycle of T is equal to X. As C is nonempty valued there exists $z \in X$ such that $z \in C(T)$. Moreover, there exists $y \in X \backslash \{z\}$ such that yTz as otherwise z would be the Condorcet winner of T. Let $T' \in \mathscr{T}(X)$ be such that $L(z,T) \subseteq L(z,T')$ and $yT'x$ for all $x \in X \backslash \{y\}$, i.e., y is the Condorcet winner of T'. As C is nonempty valued and Condorcet consistent $C(T') = \{y\}$. So we have $z \in C(T)$, $L(z,T) \subseteq L(z,T')$ but $z \notin C(T')$. Hence C is not Maskin monotonic. □

In this paper we will determine the minimal Maskin monotonic extensions of Condorcet consistent tournament solutions. The minimal Maskin monotonic extension of a tournament solution determines the alternatives that should be chosen at each tournament so that the extended solution is Maskin monotonic. Formally, let $C'(T) = C(T) \cup h(T)$ be Maskin monotonic with $h : \mathscr{T}(X) \to 2^X$. We call C' *a Maskin monotonic extension of* C and denote the set of all Maskin monotonic extensions of C by $ME(C) = \{C' \mid C' \text{ is a Maskin monotonic extension of } C\}$. The *minimal Maskin monotonic extension* $\bar{C}$ of a solution C is defined as $\bar{C} = \bigcap_{C' \in ME(C)} C'$.[6]

Before stating our main result in this section we will introduce some more notation. Given a tournament solution C, let $W_C : X \to \mathscr{T}(X)$ be a correspondence such that $W_C(x) = \{T \in \mathscr{T}(X) : x \in C(T)\}$. $W_C(x)$ is the set of all tournaments where x is a winner. Note that W_C is nonempty valued for Condorcet consistent solutions. We denote the minimum number of alternatives that $x \in X$ has to beat to be a winner of a tournament with respect to a solution C by $n_C(x)$: $n_C(x) = \min_{T \in W_C(x)} \#\{y : xTy\}$. For Condorcet consistent solutions $n_C(x) \geq 1$ for all $x \in X$ as W_C is nonempty valued and x can not be in the top-cycle or a Condorcet winner if it is beaten by all the other alternatives. Next we will define a neutrality condition for tournament solutions. Let $\pi : X \to X$ be a permutation of X. By abuse of notation, let $\pi(Y) = \bigcup_{y \in Y} \{\pi(y)\}$ for all $Y \subseteq X$. A tournament solution C is neutral if for any permutation π and any tournament $T \in \mathscr{T}(X)$, $C o \pi = \pi o C$.

Theorem 2 *Let C be a neutral Condorcet consistent solution and $\bar{C}$ be the minimal Maskin monotonic extension of C. For all $T \in \mathscr{T}(X)$, $x \in \bar{C}(T)$ if and only if $\#\{y \in X : xTy\} \geq n_c(x)$.*

[5] In Sect. 4 we are indeed proving the equivalence between the Maskin monotonicity of a Condorcet consistent tournament solution and social choice rules that are generated by Condorcet consistent tournament solutions. For the sake of completeness we are giving the direct proof of Theorem 1.

[6] Note that the minimal Maskin monotonic extension of C is unique.

Proof Let $x \in \bar{C}(T)$. If $x \in C(T)$, $\#\{y \in X : xTy\} \geq \underline{n}_c(x)$ as $T \in T_x$ and we are done. Suppose $x \notin C(T)$, i.e., $x \in h(T)$. As $x \in \bar{C}(T) \backslash C(T)$, there exists $T' \in \mathscr{T}(X)$ such that $x \in C(T')$ and $L(x, T') \subseteq L(x, T)$ but $x \notin C(T)$. By $x \in C(T')$ we have $\#\{y : xT'y\} \geq n_c(x)$ and by $L(x, T') \subseteq L(x, T)$ we have $\#\{y : xTy\} \geq \#\{y : xT'y\}$ which implies $\#\{y : xTy\} \geq n_c(x)$ as desired.

Conversely, let $\#\{y : xTy\} \geq n_c(x)$. Let $T' \in T_x$ be such that $\{y : xT'y\} = \{x_1, \ldots, x_k\}$, i.e., $\#\{y : xT'y\} = n_c(x) = k$. Let $\{y : xTy\} = \{y_1, \ldots, y_s\}$ with $s \geq k$. Let π be a permutation such that; $x \to x$, $x_1 \to y_1, \ldots, x_k \to y_k$ and all the alternatives in $X \backslash \{x, x_1, \ldots, x_k\}$ are mapped to alternatives in $X \backslash \{x, y_1, \ldots, y_k\}$ randomly. Let T'' be isomorphic to T' under π which implies $\{y : xT''y\} = \{y_1, \ldots, y_k\}$ and $L(x, T'') \subseteq L(x, T)$. Moreover, $x \in C(T'')$ by neutrality of C. Then by Maskin-monotonicity we conclude that $x \in \bar{C}(T)$. □

Next, we will characterize the minimal Maskin monotonic extensions of the following well-known Condorcet consistent solutions: The top-cycle, the uncovered set, the iterated uncovered set, the minimal covering set and the Copeland rule. Note that all these solutions are neutral and so $n_C(x) = n_C$ for all $x \in X$. We will determine the n_C for each of these solutions and the results will follow by Theorem 2. First we will give the formal definitions of the mentioned solutions.

Following Miller [9], given a tournament $T \in \mathscr{T}(X)$ and any distinct $x, y \in X$, we say that x covers y if xTy and for all $z \in X$, yTz implies xTz. We write $U(T) = \{x \in X : \nexists y \in X$ which covers $x\}$ for uncovered set of T. As $U(T) \neq \emptyset$ for each tournament T, there exists a solution $U : \mathscr{T}(X) \to \underline{X}$ called the *uncovered set*. Note that $U(T)$ is the set of alternatives which beat any other alternative by a path of length one or two which was called "Two steps principle" by Shepsle and Weingast [15].

One natural way to refine uncovered set is to consider its iterations. Let $T \in \mathscr{T}(X)$ and $Y \subseteq X$. $T' \in \mathscr{T}(Y)$ is *the restriction of* T *to* Y if $T' \subseteq T$ and is denoted by $T' = T|_Y$. Given $T \in \mathscr{T}(X)$, let $U^0(T) = X$ and define $U^{t+1}(T) = U(T|_{U^t(T)})$ for any non-negative integer t. Let t^* be the smallest integer for which $U^{t^*+1}(T) = U^{t^*}(T)$. We define the iterated uncovered set as $IU : \mathscr{T}(X) \to \underline{X}$ where $IU(T) = U^{t^*}(T)$ for all $T \in \mathscr{T}(X)$.

Another refinement of the uncovered set that we will consider is the minimal covering set which is introduced by Dutta [3]. First we need to define the notion of a "covering set." Let $T \in \mathscr{T}(X)$ and $Y \subseteq X$. We say that Y is *a covering set* for T if $U(T|_Y) = Y$ and for any $x \in X \backslash Y$, $x \notin U(T|_{Y \cup \{x\}})$. Let $COV(T) = \{Y \subseteq X : Y$ is a covering set for $T\}$. As [3] shows $COV(T) \neq \emptyset$ and there exists a *minimal covering set* $MC(T) \in COV(T)$ such that $MC(T) \subseteq Y$ for all $Y \in COV(T)$. We define the solution $MC : \mathscr{T}(X) \to \underline{X}$ which assigns to each $T \in \mathscr{T}(X)$ its corresponding minimal covering set $MC(T)$. Note that $MC(T) \subseteq IU(T)$ for all $T \in \mathscr{T}(X)$.

We now show that the n_C for the top-cycle, the uncovered set, the iterated uncovered set and the minimal covering set turn out to be the same.

Lemma 1 $n_\gamma = n_U = n_{IU} = n_{MC} = 1$.

Proof $n_\gamma = 1$: $x \in \gamma(T)$ if either x is the Condorcet winner or in the top-cycle of T. If x is the Condorcet winner then $\#\{y \in X : xTy\} = \#X - 1$. Let $X = \{x, x_1, \ldots, x_{n-1}\}$ and $T \in \mathscr{T}(X)$ be as follows: xTx_1, for all $i \in \{1, \ldots, n-2\}$ x_iTx_{i+1} and for all $i \in \{2, \ldots, n-1\}$ x_iTx. The top-cycle of T is equal to X and x beats only x_1. So, $x \in \gamma(T)$ with $\#\{y \in X : xTy\} = 1$. Note that by definition of a Condorcet winner and top-cycle x can not be the winner of a tournament if it doesn't beat anyone, i.e., $n_\gamma(x) \neq 0$. Hence $n_\gamma(x) = 1$ and as x was arbitrary it is true for all $x \in X$.

$n_U = 1$: We will use the two steps principle. Let $X = \{x, x_1, \ldots, x_{n-1}\}$ and $T \in \mathscr{T}(X)$ be as follows: xTx_1, for all $i \in \{2, \ldots, n-1\}$ x_1Tx_i and for all $i \in \{2, \ldots, n-1\}$ x_iTx. That is, at T, x beats x_1 by a path of length one and all the other alternatives by a path of length two as x_1 beats every alternative other than x by a path of one. So, $x \in U(T)$ with $\#\{y \in X : xTy\} = 1$. By definition of the uncovered set $n_U(x) \neq 0$. Hence $n_U(x) = 1$ and as x was arbitrary it is true for all $x \in X$.

$n_{IU} = 1$: Let $X = \{x, x_1, \ldots, x_{n-1}\}$ and $T \in \mathscr{T}(X)$ be as follows: xTx_1, for all $i \in \{2, \ldots, n-1\}$ x_1Tx_i, for all $i \in \{2, \ldots, n-1\}$ x_iTx and for all $i, j \in \{2, \ldots, n-1\}$ x_iTx_j if $i > j$. By two steps principle $U^1(T) = U(T\ |_X) = \{x, x_1, x_2\}$ and $U^2(T) = U^1(T) = \{x, x_1, x_2\}$. Hence, $IU(T) = \{x, x_1, x_2\}$. So, $x \in IU(T)$ with $\#\{y \in X : xTy\} = 1$. By definition of the iterated uncovered set $n_{IU}(x) \neq 0$. Hence $n_{IU}(x) = 1$ and as x was arbitrary it is true for all $x \in X$.

$n_{MC} = 1$: Let $X = \{x, x_1, \ldots, x_{n-1}\}$ and $T \in \mathscr{T}(X)$ be as follows: xTx_1, for all $i \in \{2, \ldots, n-1\}$ x_1Tx_i, for all $i \in \{2, \ldots, n-1\}$ x_iTx and for all $i, j \in \{2, \ldots, n-1\}$ x_iTx_j if $i > j$. Consider the set $Y = \{x, x_1, x_2\}$. Note that Y is a covering set for T and stops being a covering set if any of the alternatives is deleted from the set. Moreover consider $Y' = \{x, x_1, x_2, x_i\}$ where $i \in \{3, \ldots, n-1\}$. Note that $x_i \notin U(T\ |_{Y \cup \{x_i\}})$ as x_i can't beat x_2 by two steps principle. This is true for all $x_i \in \{x_3, \ldots, x_{n-1}\}$ and hence Y is a minimal covering set for T, $MC(T) = \{x, x_1, x_2\}$. So, $x \in MC(T)$ with $\#\{y \in X : xTy\} = 1$. By definition of the minimal covering set $n_{MC}(x) \neq 0$. Hence $n_{MC}(x) = 1$ and as x was arbitrary it is true for all $x \in X$. □

The minimal Maskin monotonic extensions of top-cycle, uncovered set, iterated uncovered set and minimal covering set is denoted by $\bar{\gamma}$, $\bar{U}$, $I\bar{U}$ and $\bar{MC}$, respectively.

Theorem 3 *For all $T \in \mathscr{T}(X)$, $\bar{\gamma}(T) = \bar{U}(T) = I\bar{U}(T) = \bar{MC}(T) = X \backslash CL(T)$ where $CL(T) = \{x \in X : yTx$ for all $y \in X\}$ is the Condorcet loser at T.*

Proof The result follows from Theorem 2 and Lemma 1. □

The last refinement that we will consider is the Copeland rule. The *Copeland rule* is the solution $COP : \mathscr{T}(X) \to \underline{X}$ defined for each $T \in \mathscr{T}(X)$ as $COP(T) = \{x \in X : \#\{y \in X : xTy\} \geq \#\{y \in X : zTy\}$ for all $z \in X\}$.

Lemma 2 $n_{COP} = \begin{cases} \frac{\#X-1}{2} & \text{if } \#X \text{ is odd} \\ \frac{\#X}{2} & \text{if } \#X \text{ is even} \end{cases}$.

Proof First note that in a tournament T there are $\binom{\#X}{2} = \frac{\#X(\#X-1)}{2}$ relations. So an alternative x is a winner at a tournament with respect to the Copeland rule if it beats at least $\frac{\binom{\#X}{2}}{\#X} = \frac{\#X-1}{2}$ alternatives. This is the case where each alternative beats equal number of other alternatives. Hence x beating less alternatives will result in some other alternative having a higher Copeland score than x, i.e., $x \notin COP(T)$. If $\frac{\#X-1}{2}$ is not an integer then the minimum number of alternatives that should be beaten will be the integer part of $\frac{\#X-1}{2}$ plus 1, i.e., $\frac{\#X}{2}$. □

Theorem 4 *$x \in \overline{COP}$ if and only if* $\#\{y \in X : xTy\} \geq \begin{cases} \frac{\#X-1}{2} & \text{if } \#X \text{ is odd} \\ \frac{\#X}{2} & \text{if } \#X \text{ is even} \end{cases}$.

Proof The result follows from Theorem 2 and Lemma 2. □

4 Minimal Maskin Monotonic Extensions of SCRs Generated by Tournament Solutions

Every tournament solution generates a social choice rule via the majority relation. Let N be a set of individuals with $\#N = n \geq 3$ is odd and X be a set of alternatives with $\#X \geq 3$. A preference profile is an n-tuple, $P = (P_1, \ldots, P_n)$ where each P_i is a linear order on X which represents agent i's preferences on X. The set of all linear order profiles on X is denoted by $\mathscr{L}(X)^N$. A *social choice rule* (SCR) is a mapping $F : \mathscr{L}(X)^N \to \underline{X}$. The *lower contour set of x for i at P* is $L_i(x, T) = \{y \in X : xP_i y\}$. An SCR F satisfies *Maskin monotonicity* if for any $P, P' \in \mathscr{L}(X)^N$ and $x \in F(T)$, one has $x \in F(P')$ whenever $L_i(x, P) \subseteq L_i(x, P')$ for all $i \in N$. For each $P \in \mathscr{L}(X)^N$ the majority relation on X is defined as follows: $x\mu(P)y$ if and only if $\#\{i \in N : xP_i y\} > \frac{\#N}{2}$ for all distinct $x, y \in X$. Note that μ is a complete and asymmetric binary relation on X as $\#N$ is odd and induces a tournament relation on X at each profile P. Now we can define the SCR generated by a tournament solution, $C : \mathscr{T}(X) \to \underline{X}$, as $F_C(P) = C(\mu(P))$ at each $P \in \mathscr{L}(X)^N$. Our first result establishes the relation between the Maskin monotonicity of a tournament solution and the Maskin monotonicity of the social choice rule that is generated by the tournament solution. An analogous result for cover monotonicity is given by Özkal-Sanver and Sanver [13].[7]

Theorem 5 *Let C be any tournament solution and F_C be the social choice rule induced by C. C is Maskin monotonic if and only if F_C is Maskin monotonic.*

Proof To show the "only if" part, let C be Maskin monotonic. Take any $P, P' \in \mathscr{L}(X)^N$ with $L_i(x, P) \subseteq L_i(x, P')$ for all $i \in N$, and any $x \in F_C(P)$. Note that $F_C(P) = C(\mu(P))$. Moreover, $L(x, \mu(P)) \subseteq L(x, \mu(P'))$. As C Maskin

[7] The cover monotonicity condition in Özkal-Sanver and Sanver [13] is the adaptation of its original version defined by Özkal-Sanver and Sanver [12] in the standard social choice framework.

monotonic, we have $x \in C(\mu(P')) = F_C(P')$. Hence F_C is Maskin monotonic. For the "if" part, suppose C fails Maskin monotonicity. So there exists $T, T' \in \mathscr{T}(X)$ with $L(x,T) \subseteq L(x,T')$ while $x \in C(T)$ and $x \notin C(T')$. We need to construct $P, P' \in \mathscr{L}(X)^N$ such that $\mu(P) = T$, $\mu(P') = T'$ and $L_i(x,P) \subseteq L_i(x,P')$ for all $i \in N$. Such profiles were constructed by Özkal-Sanver and Sanver [13], Theorem 5) in the spirit of [8]. So, take $P, P' \in \mathscr{L}(X)^N$ as defined in [13], Theorem 5). Then we have $x \in F_C(P) = C(T)$ with $L_i(x,P) \subseteq L_i(x,P')$ for all $i \in N$ which by Maskin monotonicity of F_C implies $x \in F_C(P') = C(T')$, contradiction. Hence C is Maskin monotonic. □

We know from Özkal-Sanver and Sanver [13] that no Condorcet consistent social choice rule is Maskin monotonic. Hence the social choice rules that are generated via the Condorcet consistent tournament solutions fail to satisfy Maskin monotonicity. In this section we will determine the minimal Maskin monotonic extensions of social choice rules that are generated by tournament solutions. First we define the minimal Maskin monotonic extension of a social choice rule. Let $F'(P) = F(P) \cup h(P)$ for all $P \in \mathscr{L}(X)^N$ be Maskin monotonic with $h : \mathscr{L}(X)^N \to 2^X$. We call F' *a Maskin monotonic extension of* F and denote the set of all Maskin monotonic extensions of F by $ME(F) = \{F' \mid F' \text{ is a Maskin monotonic extension of } C\}$. The *minimal Maskin monotonic extension* $\bar{F}$ of a social choice rule F is defined as $\bar{F} = \bigcap_{F' \in ME(F)} F'$.[8] The minimal Maskin monotonic extension of a social choice rule that is generated by a Condorcet consistent solution C will be denoted by $\bar{F}_C$. Given $P \in \mathscr{L}(X)^N$, $x \in \bar{F}_C(P)$ if either $x \in F_C(P)$ or there exists $P' \in \mathscr{L}(X)^N$ such that $x \in F_C(P')$ and $L_i(x,P') \subseteq L_i(x,P)$ for all $i \in N$. One natural thing to look at is the monotonicity properties of the social choice rules that are generated by the minimal Maskin monotonic extensions of the Condorcet consistent tournament solutions that we investigated in the previous section. The Maskin monotonicity of $F_{\bar{C}}$ follows from the previous theorem. However, it is not the minimal Maskin monotonic extension of F_C which we will establish after introducing some more results. The following lemmas show how the inclusion property between two tournament solutions is carried over to the minimal Maskin monotonic extensions of the social choice rules generated by them.

Lemma 3 *Let F, G be SCRs with $F \subseteq G$. Then $\bar{F} \subseteq \bar{G}$.*

Proof Take any $x \in \bar{F}(P)$. By definition of $\bar{F}$, $x \in F'(P)$ for all $F' \in ME(F)$. By definition $G \subseteq \bar{G}$, where $\bar{G}$ is Maskin monotonic, i.e., $\bar{G} \in ME(F)$. Hence, $x \in \bar{G}(P)$. □

Lemma 4 *Let C, C' be two tournament solutions. If $C \subseteq C'$ then $F_C \subseteq F_{C'}$.*

Proof Let $P \in \mathscr{L}(X)^N$. Take any $x \in F_C(P)$. Then we have $x \in F_C(P) = C(\mu(P)) \subseteq C'(\mu(P)) = F_{C'}(P)$. □

[8] Note that the minimal Maskin monotonic extension of F is unique.

Theorem 6 *If $C \subseteq C'$ then $\bar{F_C} \subseteq \bar{F_{C'}}$.*

Proof The result follows from Lemmas 3 and 4. □

Next, we will consider $F_\gamma(P) = \gamma(\mu(P))$, the social choice rule generated by top-cycle, γ, and give a characterization of the alternatives in F_γ.

Theorem 7 *For any $P \in \mathscr{L}(X)^N$, $x \in \bar{F}_\gamma(P)$ if and only if there exists $y, z \in X\backslash\{x\}$, $y \neq z$, and $j \in N$ such that $x\mu(P)y$ and $y, z \in L_j(x, P)$.*

Proof Let $x \in \bar{F}_\gamma(P)$. First suppose $x \in F_\gamma(P)$. Then there exists $y \in X$ such that $x\mu(P)y$, as otherwise x would be Condorcet loser. Moreover, there exists $z \in X\backslash\{x, y\}$ with $y, z \in L_j(x, P)$ for some $j \in N$: Suppose not. Then for $\frac{\#N+1}{2}$ agents we have $zP_i xP_i y$ for all $z \in X\backslash\{x, y\}$ which contradicts with $x \in F_\gamma(P)$. Next, let $x \in \bar{F}_\gamma(P)\backslash F_\gamma(P)$. Then there exists $P' \in \mathscr{L}(X)^N$ such that $x \in F_\gamma(P')$ and $L_i(x, P') \subseteq L_i(x, P)$ for all $i \in N$, but $x \notin F_\gamma(P)$. Note that x can not be a Condorcet winner at P': Suppose it is. Then $x\mu(P')y$ for all $y \in X$. By the lower contour set inclusions, we have $x\mu(P)y$ for all $y \in X$ which implies x is a Condorcet winner at P and $F_\gamma(P) = x$, contradiction. So, x should be a member of the unique top-cycle at P'. By the above argument, x is not a Condorcet winner but a member of the unique top-cycle at P'. Then there exist $y, z \in X$ such that $x\mu(P')y$ and $y\mu(P')z$. Suppose the sets, $\{i \in N : xP'_i y\}$ and $\{i \in N : yP'_i z\}$ are disjoint. As, $\#N$ is odd, we have $\#\{i \in N : xP_i y\} + \#\{i \in N : xP_i y\} \geq \frac{\#N+1}{2} + \frac{\#N+1}{2} > \#N$, contradiction. So $\{i \in N : xP'_i y\}$ and $\{i \in N : yP'_i z\}$ are not disjoint and there exists $j \in N$ such that $xP'_j yP'_j z$. Then as the lower contour set of x at P' is preserved at P for each agent $i \in N$ we have $x\mu(P)y$ and $y, z \in L_j(x, P)$.

Conversely, suppose there exists $y, z \in X\backslash\{x\}$, $y \neq z$, and $j \in N$ such that $x\mu(P)y$ and $y, z \in L_j(x, P)$. Let $P' \in \mathscr{L}(X)^N$ be defined as follows: Let $N_x = \{i \in N : xP'_i y\}$ with $\#N_x = \frac{\#N+1}{2}$. For some $j \in N_x$, let $L_j(x, P') = \{x, y, z\}$, for all $i \in N_x\backslash\{j\}$ $L_i(x, P') = \{x, y\}$ and for all $i \in N\backslash N_x$ $L_i(x, P') = \{x\}$. For all $i \in N\backslash N_x$ y is top ranked and z is second ranked, and for all $i \in N_x\backslash\{j\}$ z is top-ranked. All the other alternatives ranked randomly. Now we have $x\mu(P')y\mu(P')z\mu(P')x_1\mu(P')\ldots x_k\mu(P')x$ which implies that $x \in F_\gamma(P')$. Moreover, $L_i(x, P') \subseteq L_i(x, P)$ for all $i \in N$. As $\bar{F}_C$ is Maskin monotonic we conclude that $x \in \bar{F}_\gamma$. □

Now we can get back to the question that was posed earlier; the relation between the social choice rule that is generated via the minimal Maskin monotonic extension of a Condorcet consistent tournament solution and the minimal Maskin monotonic extension of a social choice rule that is generated via a Condorcet consistent tournament solution.

Theorem 8 *For any Condorcet consistent C, $\bar{F}_C(P) \subseteq F_{\bar{C}}(P)$ for all $P \in \mathscr{L}(X)^N$ while the inclusion is strict for at least one P.*

Proof $\bar{F}_C(P) \subseteq F_{\bar{C}}(P)$: Let $P \in \mathscr{L}(X)^N$ and $x \in \bar{F}_C(P)$. Then we have either, $x \in F_C(P)$ or $x \notin F_C(P)$ and $x \in F_C(P')$ for some $P' \in \mathscr{L}(X)^N$ with $L_i(x, P') \subseteq L_i(x, P)$ for all $i \in N$. Let $T = \mu(P)$. Then $x \in F_C(P) =$

$C(T) \subseteq \bar{C}(T) = F_{\bar{C}}(P)$. Next, suppose $x \in F_C(P')$ for some $P' \in \mathscr{L}(X)^N$ with $L_i(x, P') \subseteq L_i(x, P)$ for all $i \in N$ and let $T = \mu(P)$ and $T' = \mu(P')$. Then $x \in C(T') \subseteq \bar{C}(T')$ and by the lower contour set inclusions we have, for all $y \in X\backslash\{x\}$, $x\mu(P')y$ implies $x\mu(P)y$, i.e, $xT'y$ implies xTy. So, $L(x, T') \subseteq L(x, T)$ and by Maskin monotonicity of $\bar{C}$ we have $x \in \bar{C}(T)$. Hence, $x \in F_{\bar{C}}(P)$.

For the second part let $X = \{x_1, \ldots, x_n\}$ and T' be a tournament with $x_1T'x_2T' \ldots T'x_nT'x_1$. Then $\gamma(T') = X$ and $\emptyset \neq C(T') \subseteq X$. Without loss of generality let $x_1 \in C(T') \subseteq \bar{C}(T')$. Let T be such that x_nTx_i for all $x_i \in X\backslash\{x_n\}$ and x_1Ty if $x_1T'y$. By monotonicity of $\bar{C}$, we have $x \in \bar{C}(T)$. Note that by our construction of P, $x_n \notin L_i(x_1, P)$ for all $i \in N$. Then by Theorem 7, $x_1 \notin \bar{F}_\gamma(P)$ and by Theorem 6 $x_1 \notin \bar{F}_C(P)$. □

Next we will determine the minimal Maskin monotonic extensions of the social choice rules which are generated by the uncovered set, the iterated uncovered set and the minimal covering set. The social choice rules that are generated will be denoted by $F_U(P) = U(\mu(P))$, $F_{IU}(P) = IU(\mu(P))$, $F_{MC}(P) = MC(\mu(P))$, respectively and their minimal Maskin monotonic extensions will be denoted by $\bar{F}_U$, $\bar{F}_{IU}$, $\bar{F}_{MC}$, respectively. First we will introduce some more notation. For all $x \in X$ and $P \in \mathscr{L}(X)^N$ let $M(x, P) = \{y : x\mu(P)y\}$ be the set of all alternatives that x majority beats at P. For all $x, y \in X$ and $P \in \mathscr{L}(X)^N$ let $I(x, y, P) = \{i \in N : xP_iy\}$ and $I(x, P) = \bigcup_{y \in M(x,P)} \{i \in N : xP_iy\}$ be the set of agents that belong to at least one majority of x over another alternative y.

Now we introduce our condition that characterizes $\bar{F}_U$, $\bar{F}_{IU}$ and $\bar{F}_{MC}$.

Condition I $(x, P) \in X \times \mathscr{L}(X)^N$ satisfies Condition I if either

1. $\#M(x, P) = 1$ and there exists $T \subseteq I(x, P)$ such that $\#T \leq \frac{n+1}{2}$ and $\bigcup_{i \in T} L_i(x, P) = X$, or
2. $\#M(x, P) > 1$ and $\bigcup_{i \in I(x,P)} L_i(x, P) = X$ holds.

Theorem 9 *For any $(x, P) \in X \times \mathscr{L}(X)^N$ if $x \in \bar{F}_U(P)$ then (x, P) satisfies Condition I.*

Proof Let $P \in \mathscr{L}(X)^N$ and $x \in \bar{F}_U(P)$. First assume that $x \in F_U(P)$. By definition of F_U, we have $M(x, P) \neq \emptyset$, i.e., $\#M(x, P) \geq 1$. For (1), assume $\#M(x, P) = 1$ with $M(x, P) = \{y\}$. By two steps principle, $x \in F_U(P)$ and $M(x, P) = \{y\}$ implies that $y\mu(P)z$ for all $z \in X\backslash\{x, y\}$, i.e., $M(y, P) = X\backslash\{x\}$.

Claim: For all $z \in X\backslash\{x, y\}$, $I(x, y, P) \cap I(y, z, P) \neq \emptyset$. Proof of the claim: First note that $\#I(x, y, P) \geq \frac{n+1}{2}$ as $x\mu(P)y$ and $\#I(y, z, P) \geq \frac{n+1}{2}$ as $y\mu(P)z$ for all $z \in X\backslash\{x, y\}$. So if $I(x, y, P)$ and $I(y, z, P)$ are disjoint $\#I(x, y, P) + \#I(y, z, P) = n + 1 > n$, contradiction which completes the proof of the claim. So, for each $z \in X\backslash\{x, y\}$ there exists $i \in I(x, y, P) = I(x, P)$ such that xP_iyP_iz, i.e., $z \in L_i(x, P)$. Hence there exists $T \subseteq I(x, P)$ such that $\bigcup_{i \in T} L_i(x, P) = X$. Now, suppose for all $T \subseteq I(x, P)$ with $\bigcup_{i \in T} L_i(x, P) = X$,

we have $\#T > \frac{n+1}{2}$, i.e., $\#T \geq \frac{n+3}{2}$. Denote the set of all such T by $\overline{T}$ and let $T' \in \overline{T}$ be such that $\#T' = \min_{T \in \overline{T}} \#T$. As $\bigcup_{i \in T'} L_i(x, P) = X$, $\#T' \geq \frac{n+3}{2}$ and by minimality of T', there exists $z \in X \backslash \{x, y\}$ and $j \in T'$ such that $xP_j yP_j z$ and $zP_i xP_i y$ for all $i \in T' \backslash \{j\}$ with $\#T' \backslash \{j\} \geq \frac{n+3}{2} - 1 = \frac{n+1}{2}$. But then we have $z\mu(P)y$ which contradicts with $M(y, P) = X \backslash \{x\}$. So, there exists $T \subseteq I(x, P)$ with $\#T \leq \frac{n+1}{2}$ where $\bigcup_{i \in T} L_i(x, P) = X$. For (2), assume $\#M(x, P) > 1$. Suppose $\bigcup_{i \in I(x,P)} L_i(x, P) \neq X$. Then there exists $z \in X$ such that $z \notin L_i(x, P)$ for all $i \in I(x, P)$. Note that for all $y \in M(x, P)$, $I(x, y, P) \geq \frac{n+1}{2}$ and so $I(x, P) \geq \frac{n+1}{2}$. So for at least $\frac{n+1}{2}$ individuals $zP_i x$ and $zP_i y$ for all $y \in M(x, P)$. But then $z\mu x$ and $z\mu y$ for all $y \in M(x, P)$, i.e., z covers x which contradicts with $x \in F_U(P)$. Hence $\bigcup_{i \in I(x,P)} L_i(x, P) = X$. So we showed that if $x \in F_U(P)$, then either 1 or 2 holds. Next, let $x \in \bar{F}_U(P) \backslash F_U(P)$. Then there exists $P' \in \mathscr{L}(X)^N$ such that $x \in F_U(P')$ and $L_i(x, P') \subseteq L_i(x, P)$ for all $i \in N$, but $x \notin F_U(P)$. As $x \in F_U(P')$ the above argument applies and either (1) or (2) holds for P'. Note that if $x\mu(P')y$ and $\bigcup_{i \in I(x,P')} L_i(x, P') = X$ then $x\mu(P)y$ and $\bigcup_{i \in I(x,P)} L_i(x, P) = X$ because the lower contour set for each individual at P' is preserved at P. So, either (1) or (2) holds for P. □

Theorem 10 *If $(x, P) \in X \times \mathscr{L}(X)^N$ satisfies Condition I, then $x \in \bar{F}_{MC}(P)$.*

Proof First suppose (1) holds. Let $M(x, P) = \{y\}$ and let $T' \subseteq N$ be such that $T \subseteq T' \subseteq I(x, y, P) = I(x, P)$ with $\#T' = \frac{n+1}{2}$. Let $P' \in \mathscr{L}(X)^N$ be such that:

i $-L_i(x, P') \subseteq L_i(x, P)$ for all $i \in N$,
ii $-L_i(y, P') = L_i(x, P) \backslash \{x\}$ for all $i \in T'$,
iii $-L_i(y, P') = X$ for all $i \in N \backslash T'$.

Claim 1 $x \in \bar{F}_{MC}(P')$. Proof of the claim: Suppose $x \notin \bar{F}_{MC}(P')$. First note that by (i), $M(x, P') = \{y\}$. Let $z \in X \backslash \{x, y\}$. By (1) we have $\bigcup_{i \in T} L_i(x, P) = X$ which together with $T \subseteq T'$ and (ii) implies that $\bigcup_{i \in T'} L_i(y, P') = X \backslash \{x\}$. Then there exists $j \in T'$ such that $z \in L_j(y, P')$, i.e., $yP'_j z$. Moreover, by (iii) $yP'_i z$ for all $i \in N \backslash T'$ where $\#N \backslash T' = \frac{n-1}{2}$. So $yP'_i z$ for at least $\frac{n+1}{2}$ individuals which implies $y\mu(P')z$. As z was arbitrary $y\mu(P')z$ for all $z \in X \backslash \{x, y\}$. So y is a Condorcet winner for the set $X \backslash \{x\}$. Hence, $F_{MC}(P') = y$. However, $F_U(P'|_{\{x,y\}}) = \{x\}$, which contradicts with $F_{MC}(P') = y$. Hence, $x \in F_{MC}(P') \subseteq \bar{F}_{MC}(P')$. Now by construction of P' (i) and Maskin monotonicity of $\bar{F}_{MC}$ we conclude that $x \in \bar{F}_{MC}(P)$ and complete the proof of Claim 1.

Next suppose (2) holds. Let $M(x, P) = \{y_1, \ldots, y_s\}$ with $s > 1$ and $Z = \{z \in X : z\mu(P)x\}$. First, note that if $Z = \varnothing$ then x is the Condorcet winner and we are done. So, suppose $Z \neq \varnothing$.

For the proof we will construct a profile P^{sm} in several steps:

Step1: Define $P^0 \in \mathscr{L}(X)^N$ as follows: $L_i(x, P^0) = L_i(x, P)$ for all $i \in N$ and for all $\{y, z\} \in L_i(x, P)$ with $i \in I(x, P)$, $y \in M(x, P)$ and $z \in Z$, yP_iz.

Let $I^j = I(x, y_j, P^{j-1}) \setminus \bigcup_{y \in M(x,P)\setminus\{y_j\}} I(x, y, P^{j-1})$. Define $T^j \subseteq N$ for all $j \in \{1, \ldots, s\}$ as follows:

1. For $j = 1$:
   i- $T^1 \subseteq I(x, y_1, P^0)$, ii- $\#T^1 = \frac{n+1}{2}$, iii- $I^1 \subseteq T^1$.
2. For $j \in \{2, \ldots, s\}$:
   i- $T^j \subseteq I(x, y_j, P^{j-1})$, ii- $\#T^j = \frac{n+1}{2}$, iii- $I^j \subseteq T^j$, iv- $I(x, y_j, P^{j-1}) \cap (N \setminus T^{j-1}) \subseteq T^j$.

Step2-s: Define $P^j \in \mathscr{L}(X)^N$ for all $j \in \{1, \ldots, s\}$ as follows:

i- ${P^j}_i = {P^{j-1}}_i$ for all $i \in T^j$,
ii- $L_i(a, P^j) = L_i(a, P^{j-1}) \setminus \{y_j\}$ for all $i \in N \setminus T^j$ and for all $a \in X$.

Note that by (ii) we have $L_i(y_j, P^j) = X$ for all $i \in N \setminus T^j$. Moreover, by (i), (ii) and $\#T^j = \frac{n+1}{2}$ for all $j \in \{1, \ldots, s\}$, we have $M(x, P) = M(x, P^0) = M(x, P^j) = M(x, P^s)$.

Claim 2 For each $z \in Z$ there exists $y_j \in M(x, P)$ such that $y_j \mu(P^s) z$. Proof of the claim: Let $z \in Z$. First we show that there exists $i \in N$ and $y_j \in M(x, P)$ such that $x {P^s}_i y_j P^s{}_i z$. We know that $z \in \bigcup_{i \in I(x,P)} L_i(x, P) = X$ and by the construction of P^0 we have $x {P^0}_i y_j P^0{}_i z$ for some $i \in N$ and $y_j \in M(x, P)$. So, if we can show that $i \in T^j$ then we are done. First suppose, $i \in T^l$ where $l \in \{1, \ldots, j-1\}$. But then we have $y_l P^l{}_i z$ and $y_l P^l{}_r z$ for all $r \in N \setminus T^l$ with $\#N \setminus T^l = \frac{n-1}{2}$ and hence $y_l \mu(P^l) z$ which implies $y_l \mu(P^s) z$. Next, suppose $i \notin T^l$ for all $l \in \{1, \ldots, j-1\}$. Consider, in particular, T^{j-1}. As $x {P^0}_i y_j P^0{}_i z$ and the relative orderings of x and y_j is preserved in the first j-1 steps for all $i \in N$ we have $i \in I(x, y_j, P^0) = I(x, y_j, P^{j-1})$. Moreover, $i \in N \setminus T^{j-1}$. Then, by the construction of T^j, 2-(iv), we conclude that $i \in T^j$. Hence $x P^s{}_i y_j P^s{}_i z$ exists for some $i \in T^j$. Now, at step-j, we have $x P^j{}_i y_j P^j{}_i z$ for some $i \in T^j$ and $y_j P^j{}_l z$ for all $l \in N \setminus T^j$ with $\#N \setminus T^j = \frac{n-1}{2}$. So, $y_j \mu(P^j) z$. Since the relative orderings of y_j and z is preserved in the latter steps we have $y_j \mu(P^s) z$. As $z \in Z$ was chosen arbitrarily we conclude that for each $z \in Z$ there exists $y_j \in M(x, P)$ such that $y_j \mu(P^s) z$ and complete the proof of Claim 2.

We introduce more notation before finalizing our construction: Let ${y_j}^- = \{y_1, \ldots, y_{j-1}\}$, $Z_j = \{z \in Z : y_j \mu(P^s) z\}$ for all $j \in \{1, \ldots, s\}$, ${Z_s}' = Z_s$, ${Z_j}' = Z_j \setminus \bigcup_{i=j+1,\ldots,s} Z_i$ for all $j \in \{1, \ldots, s-1\}$, and ${Z_j}'' = Z \setminus \bigcup_{i=j,\ldots,s} Z_i$ for all $j \in \{1, \ldots, s\}$. For simplicity, we denote $Y = M(x, P^s) = \{y_1, \ldots, y_s\}$. Next we define P^{sm} as follows:

i -$L_i(y_j, P^{sm}) = L_i(y_j, P^s) \setminus {y_j}^-$ for all $y_j \in Y$ and for all $i \in N$ such that $y_j, y_k \in L_i(x, P^s)$ or $y_j, y_k \notin L_i(x, P^s)$ where $y_k \in {y_j}^-$,

ii -$L_i(z_k, P^{sm}) = L_i(z_k, P^s) \cup Z_j$" for all $z_k \in Z_j{}'$, for all $j \in \{1, \ldots, s\}$ and for all $i \in N$ such that $z_k, z" \in L_i(x, P^s)$ or $z_k, z" \notin L_i(x, P^s)$ where $z" \in Z_j$".[9]

Note that by (i) and (ii) we have $L_i(x, P^{sm}) = L_i(x, P^s)$ for all $i \in N$.

Claim 3 $x \in \bar{F}_{MC}(P^{sm})$. Proof of the claim: Suppose not, i.e., $x \notin \bar{F}_{MC}(P^{sm})$ and $x \notin F_{MC}(P^{sm})$. Let $F_{MC}(P^{sm}) = \bar{Y} \cup \bar{Z}$ where $\bar{Y} \subseteq Y$ and $\bar{Z} \subseteq Z$. By definition of minimal covering set we have $x \notin F_U(P^{sm}|_{\bar{Y} \cup \bar{Z} \cup \{x\}})$. As $L_i(x, P^{sm}) = L_i(x, P^s) \subseteq L_i(x, P)$ for all $i \in N$, and $\bar{Y} \subseteq Y$, $x\mu(P^{sm})y$, for all $y \in \bar{Y}$. So, there exists $\bar{z} \in \bar{Z}$ such that x can not reach in two steps. That is, there does not exist $y \in \bar{Y}$ such that $y\mu(P^{sm})\bar{z}$. But we know that, by Claim 2 and the construction of P^{sm} there exists $y \in Y$ such that $y\mu(P^{sm})\bar{z}$. So, $y \in Y \backslash \bar{Y}$. Let $y = y_j$ and $\bar{z} \in Z_j$. First suppose $\bar{z} \notin Z_i$ for all $i \neq j$. Then $y_j\mu(P^{sm})\bar{z}\mu(P^{sm})y'$ for all $y' \in Y \backslash \{y_j\}$. Moreover, for all $z' \in Z_i$ with $i > j$ we have $y_j\mu(P^{sm})y_i\mu(P^{sm})z'$ and for all $z' \in Z_l$ with $l < j$ we have $y_j\mu(P^{sm})\bar{z}\mu(P^{sm})z'$ by P^{sm} (1) and (2). Note that by Claim 1 $\bigcup_j Z_j = Z$. Then by the 2 steps principle $y_j \in F_U(P^{sm}|_{\bar{Y} \cup \bar{Z}\{y_j\}})$ which contradicts with $F_{MC}(P^{sm}) = \bar{Y} \cup \bar{Z}$. Hence, $x \in F_{MC}(P^{sm})$. Next suppose $\bar{z} \in \bigcap_{i \in \bar{S}} Z_i$ where $S \subseteq \{1, \ldots, s\}$. Let $j = minS$. For all $i < j$ we have $y_j\mu(P^{sm})\bar{z}\mu(P^{sm})y_i$ by minimality of j in S and P^{sm} (2). For all $l > j$ we have $y_j\mu(P^{sm})y_l$ by P^{sm} (1). For all $z' \in Z$ the above argument applies and we conclude $x \in F_{MC}(P^{sm})$. Then $x \in \bar{F}_{MC}(P^{sm})$ which completes the proof of Claim 3.

Now, by construction of P^s and P^{sm} we have $L_i(x, P^{sm}) \subseteq L_i(x, P^s) \subseteq L_i(x, P)$ for all $i \in N$ and by Maskin monotonicity of $\bar{F}_{MC}$ we conclude that $x \in \bar{F}_{MC}(P)$. □

Theorem 11 *For all* $(x, P) \in X \times \mathscr{L}(X)^N$, $x \in \bar{F}_U(P) = \bar{F}_{IU}(P) = \bar{F}_{MC}(P)$ *if and only if* (x, P) *satisfies Condition I.*

Proof By Theorems 9 and 10 we have $\bar{F}_U \subseteq \bar{F}_{MC}$. Moreover, $\bar{F}_{MC} \subseteq \bar{F}_U$ by Theorem 6. Then we have $\bar{F}_U = \bar{F}_{MC}$. Finally by Theorem 6 we have $\bar{F}_U = \bar{F}_{IU} = \bar{F}_{MC}$ and they are all equivalent to Condition I by Theorems 9 and 10. □

[9]The following example illustrates the construction of P^{sm} from a given pair (x, P) satisfying Condition I-(2).

Example Let $\#N = 3$, $X = \{x, y_1, y_2, z_1, z_2\}$ and $P \in \mathscr{L}(X)^N$ be defined as follows: $z_2P_1y_2P_1xP_1z_1P_1y_1$, $z_1P_2xP_2y_2P_2z_2P_2y_1$ and $z_1P_3z_2P_3xP_3y_1P_3y_2$. Note that $M(x, P) = \{y_1, y_2\}$, $Z(x, P) = \{z_1, z_2\}$ and (x, P) satisfies Condition I (2). Then we obtain P^0 as follows: $z_2P^0{}_1y_2P^0{}_1xP^0{}_1y_1P^0{}_1z_1$, $z_1P^0{}_2xP^0{}_2y_2P^0{}_2y_1P^0{}_2z_2$ and $z_1P^0{}_3z_2P^0{}_3xP^0{}_3y_1P^0{}_3y_2$. Next we construct P^1. Note that $I^1 = \{1\}$ and let $T^1 = \{1, 2\}$. Then we obtain P^1 as follows: $z_2P^1{}_1y_2P^1{}_1xP^1{}_1y_1P^1{}_1z_1$, $z_1P^1{}_2xP^1{}_2y_2P^1{}_2y_1P^1{}_2z_2$ and $y_1P^1{}_3z_1P^1{}_3z_2P^1{}_3xP^1{}_3y_2$. For P^2, note that $I^2 = \{3\}$ and $T^2 = \{2, 3\}$. Then we obtain P^2 as follows: $y_2P^2{}_1z_2P^2{}_1xP^2{}_1y_1P^2{}_1z_1$, $z_1P^2{}_2xP^2{}_2y_2P^2{}_2y_1P^2{}_2z_2$ and $y_1P^2{}_3z_1P^2{}_3z_2P^2{}_3xP^2{}_3y_2$. Finally, P^{2m} is obtained as follows: $y_2P^{2m}{}_1z_2P^{2m}{}_1xP^{2m}{}_1y_1P^{2m}{}_1z_1$, $z_1P^{2m}{}_2xP^{2m}{}_2y_1P^{2m}{}_2y_2P^{2m}{}_2z_2$ and $y_1P^{2m}{}_3z_2P^{2m}{}_3z_1P^{2m}{}_3xP^{2m}{}_3y_2$.

The following example shows that Condition I is strictly stronger than the condition that characterizes the minimal Maskin monotonic extension of the social choice rule which is generated by top-cycle.

Example Let $X = \{x, y, z, t\}$ and P be defined as follows: $xP_i y$ for all $i \in N$, $tP_i x$ for all $i \in N$ and $xP_1 z$. Then $x \in \bar{F}_\gamma(P)$ but $x \notin \bar{F}_U(P)$ which is characterized by Condition I.

5 Conclusion

In this paper we first studied the minimal Maskin monotonic extensions of Condorcet consistent solutions. As it turns out, given a neutral Condorcet consistent tournament solution the minimum number of alternatives that has to be beaten to be a winner at some tournament identifies the alternatives that are in the extension at this tournament. For the top-cycle, the uncovered set, the iterated uncovered set and the minimal covering set this number is equal to 1 which implies that at each tournament all the alternatives except the Condorcet looser is contained in the minimal monotonic extension. For the Copeland rule, however, this number depends on the number of alternatives over which the tournament is defined and is greater than 1 if there are 4 or more alternatives.

We also determined the minimal Maskin monotonic extensions of the social choice rules that are generated by some Condorcet consistent solutions, namely, the top-cycle, the uncovered set, the iterated uncovered set and the minimal covering set. In the social choice environment the equivalence of the minimal Maskin monotonic extensions carries over for the uncovered set, the iterated uncovered set and the minimal covering set. The minimal Maskin monotonic extension of the top-cycle turns out to be coarser than the three mentioned above. Moreover, due to the finer structure in the social choice environment we showed that the social choice rules that are generated via the minimal Maskin monotonic extensions of Condorcet consistent tournament solutions are coarser than the minimal Maskin monotonic extensions of the social choice rules that are generated via Condorcet consistent tournament solutions.

The minimal Maskin monotonic extension of other Condorcet consistent solutions and the minimal monotonic extensions of the social choice rules that are generated via them are some of the future research topics in the field.

References

1. Baigent N, Klamler C (2004) Transitive closure, proximity and intransitivities. Econ Theory 23:175–181
2. Copeland A (1951) A reasonable social welfare function. University of Michigan seminar on the applications of mathematics to social sciences

3. Dutta B (1988) Covering sets and a new condorcet choice correspondence. J Econ Theory 44(1):63–80
4. Erdem O, Sanver MR (2005) Minimal monotonic extensions of scoring rules. Soc Choice Welf 25(1):31–42
5. Kara T, Sönmez T (1996) Nash implementation of matching rules. J Econ Theory 68(2):425–439
6. Laslier JF (1997) Tournament solutions and majority voting. Springer, New York
7. Maskin E (1999) Nash equilibrium and welfare optimality. Rev Econ Stud 66(1):23–38
8. McGarvey DC (1953) A theorem on the construction of voting paradoxes. Econometrica 21(4):608–610. doi:http://dx.doi.org/10.2307/1907926. articleType: research-article/Full publication date: Oct., 1953/Copyright textcopyright 1953 The Econometric Society
9. Miller N (1980) A new solution set for tournaments and majority voting: further graph-theoretical approaches to the theory of voting. Am J Polit Sci 24:68–96
10. Moulin H (1983) The strategy of social choice. Advanced textbooks in economics, vol 18. North Holland, Amsterdam
11. Muller E, Satterthwaite MA (1977) The equivalence of strong positive association and incentive compatibility. J Econ Theory 14:412–418
12. Özkal-Sanver I, Sanver MR (2006) Nash implementation via hyperfunctions. Soc Choice Welf 26(3):607–623
13. Özkal-Sanver I, Sanver MR (2010) A new monotonicity condition for tournament solutions. Theory Decis 69:439–452
14. Sen A (1995) The implementation of social choice functions via social choice correspondences; a general formulation and a limit result. Soc Choice Welf 12:277–292
15. Shepsle K, Weingast B (1984) Uncovered sets and sophisticated voting outcomes with implications for agenda institution. Am J Polit Sci 28:49–74
16. Thomson W (1999) Monotonic extensions on economic domains. Rev Econ Des 4(1):13–33

Single-Profile Choice Functions and Variable Societies: Characterizing Approval Voting

Hanji Wu, Yongsheng Xu, and Zhen Zhong

Abstract We study approval voting in a setting with a fixed profile of individuals' choices and variable societies. Four properties each linking choices made by a group of individuals to choices by its various subgroups are introduced, and are used for characterizing approval voting.

Keywords Approval voting • Choice function • Variable societies

1 Introduction

Approval voting is an important voting method that has been used in many contexts and works as follows: when a group of individuals deciding on several alternatives and assuming that each chooses his 'approved' ones, the alternatives that get the most 'votes' among the available alternatives emerge as the winners.

Since the introduction of approval voting (see [2]), there has been a number of axiomatic studies on its behavior. It is fair to say that all the axiomatic studies in the literature are based on multi profiles of preferences or choice functions with either a fixed society or variable societies. See Xu [9] for a survey on axiomatizations of approval voting in the literature.

In this paper, we take a different approach from the existing ones to study approval voting. In our framework, we work with a fixed profile of individuals' choices while allow various societies to be formed. A similar framework has been employed by Xu and Zhong [10] to study simple majority rule. Approval voting is thus investigated from a perspective of linking the society's choices with choices made by its various sub-societies. It is then natural to see how choices made by various sub-societies can be linked to the choice by the society as a whole. In particular, we can ask questions like the following: when the choices of two

H. Wu • Y. Xu (✉)
Department of Economics, Georgia State University, Atlanta, GA, USA
e-mail: wuhanji@yahoo.com; yxu3@gsu.edu

Z. Zhong
Research Institute of Finance and Banking, the People's Bank of China, Beijing 100800, China
e-mail: zhongzhen96@gmail.com

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_9

disjoint sub-societies have some alternatives in common, would those common alternatives continue to be chosen by the society joined by the two sub-societies? what happens to the choices by the society joined by the two disjoint sub-societies when their choices have nothing in common? We show that approval voting can be characterized by the following properties (see formal definitions of these properties in Sect. 3): (1) a society consisting of one individual should reflect this individual's choices, (2) when the choices of two disjoint sub-societies have some alternatives in common, the choices of the society joined by the two sub-societies should be given by those commonly chosen alternatives of the two sub-societies, (3) when an alternative is not chosen by two disjoint sub-societies, this alternative should not be a chosen by the society formed by the two sub-societies, and (4) when an individual's choices have nothing in common with the choices of a sub-society and when they form a new society, the choices of the new society should include those alternatives chosen by the former sub-society. In a sense, approval voting is characterized by two types of properties: (1) how the choices of a society consisting of just one individual are linked to the choices of this individual, and (2) how the choices of a society formed by two distinct societies are linked to the respective choices of the two societies.

The remainder of the paper is organized as follows. In Sect. 2, we introduce the basic notation and definitions. Section 3 presents a set of properties and axiomatic derivation of approval voting. The paper is concluded in Sect. 4 by offering some brief remarks.

2 Notation and Definitions

Let there be $n \geq 2$ individuals, and let $N = \{1, \cdots, n\}$ denote the set of individuals in the society. A is to denote a set of finite alternatives with two or more alternative. Throughout this paper, we assume that A is given and fixed.

For each $i \in N$, $C_i(A)$ stands for individual i's choice set over A. It is assumed that $C_i(A) \subseteq A$ and $C_i(A) \neq \emptyset$ for all $i \in N$. For each $i \in N$, $C_i(A)$ is interpreted as the alternatives approved by individual i from the set A.

Non-empty subsets of N are denoted by $S, T, \cdots$, and are called *coalitions*. For any coalition S, $\#S$ denotes the cardinality of S. The set of all non-empty coalitions is to be denoted by $\mathcal{K}$.

Let $\alpha^N(A) \equiv \{C_1(A), \cdots, C_i(A), \ldots, C_n(A)\}$ denote a *profile* of individuals' choices over A. In this paper, we consider $\alpha^N(A)$ as **fixed**. For any coalition $S \in \mathcal{K}$, let $\alpha^S(A)$ denote the set $\{C_i(A) \in \alpha^N(A) : i \in S\}$.

An *aggregation rule* f assigns, for each $\alpha^S(A) \in \bigcup_{T \in \mathcal{K}} \alpha^T(A)$, a non-empty choice set over A: $C(S, A) = f(\alpha^S(A))$, where $\emptyset \neq C(S, A) \subseteq A$ is called the choice set of the coalition over the set A.

For each coalition S, let $N(x, S, A) \equiv \#\{i \in S : x \in C_i(A) \text{ for some } i \in S\}$. An aggregation rule f is said to be *Approval Voting* if and only if, for all coalition S, $x \in C(S, A) \Leftrightarrow N(x, S, A) \geq N(y, S, A)$ for all $y \in A$.

3 Axioms and a Characterization Result

We first consider the following axioms that are to be imposed on an aggregation rule.

Self Determination (SD): For all $i \in N$, $C(\{i\}, A) = C_i(A)$.
Monotonicity (M): For all coalitions $S \in \mathcal{K}$ and all $i \in N \setminus S$, if $[C_i(A) \cap C(S, A) \neq \emptyset]$ then $C(S \cup \{i\}, A) = C(S, A) \cap C_i(A)$.
Unanimous Rejection (UR): For all coalition $S \in \mathcal{K}$, all individual $i \in N \setminus S$, and all $x \in A$, if $[x \notin C_i(A)$ and $x \notin C(S, A)]$ then $x \notin C(S \cup \{i\}, A)$.
Positive Association (PA): For all coalitions $S \in \mathcal{K}$ and all $i \in N \setminus S$, if $[C_i(A) \cap C(S, A) = \emptyset]$ then $C(S, A) \subseteq C(S \cup \{i\}, A)$.

(SD) is fairly straightforward and requires that, when a society consists of a single individual i, the choice set of this society coincides with this single individual's choice set. It thus reflects the idea of self determination.

(M) says that, when an individual i is added to a coalition S to form a new society $S \cup \{i\}$, if some alternative happens to be chosen by both the coalition S and the individual i, then the choice set of the new society, $S \cup \{i\}$, consists exactly of those alternatives that are chosen by both S and i. (M) thus reflects the idea that unanimity between a coalition and an individual should be respected. Stronger versions of (M) known as Reinforcement or Consistency have been proposed by several authors including Fine and Fine, Smith and Young[3, 4, 8, 11, 12] for different contexts.

(UR) says that if an alternative is not chosen by a coalition S and by an individual i, then this alternative cannot be in the choice set of the coalition formed by S and i. This axiom reflects again the idea of respecting unanimous choices made by individuals and coalitions.

Finally, (PA) states that if a society is formed by adding a new member to a coalition and the choice set by this individual has nothing in common with the choice set of the coalition, then the choice set of the new society must supersede the choices of the existing coalition. To a certain degree, (PA) gives a 'favorable' treatment to the choices of an existing coalition when a new member is added to this coalition when forming a new coalition.

With the help of the above axioms, we now state and prove our result, a characterization of approval voting in our framework.

Theorem 1 *An aggregation rule f is approval voting if and only if it satisfies (SD), (M), (UR) and (PA).*

Proof First, it can be checked easily that approval voting satisfies (SD), (M) and (UR). We now show that approval voting satisfies (PA) as well. Let $S \in \mathcal{K}$ and $i \in N \setminus S$, and suppose that $C_i(A) \cap C(S, A) = \emptyset$. We need to show that, if the aggregation rule is approval voting, then $C(S, A) \subseteq C(S \cup \{i\}, A)$. Since $x \in C(S, A)$, we have $N(x, S, A) \geq N(y, S, A)$ for all $y \in A$ and $N(x, S, A) \geq 1$. Note that $C_i(A) \cap C(S, A) = \emptyset$. It then follows that $N(x, S \cup \{i\}, A) \geq N(y, S \cup \{i\}, A)$

for all $y \in A$ implying that $x \in C(S \cup \{i\}, A)$. Therefore, (PA) is satisfied by approval voting.

Next, we show that, if an aggregation rule f satisfies (SD), (M), (UR) and (PA), then it must be approval voting. Let f satisfy (SD), (M), (UR) and (PA). Let $\alpha^N(A) \equiv \{C_1(A), \cdots, C_i(A), \ldots, C_n(A)\}$ be given. We shall use mathematical induction (on the number of individuals in a coalition) to show that,

$$\text{for all } S \in \mathcal{K},\ C(S, A) = \{x \in A : N(x, S, A) \geq N(y, S, A)\ \forall y \in A\} \qquad (*)$$

To begin with, note that, for all $i \in N$, by (SD), $C(\{i\}, A) = C_i(A)$ follows easily. Thus, (*) holds for any coalition $S \in \mathcal{K}$ with $\#S = 1$.

Suppose (*) holds for any coalition $S \in \mathcal{K}$ with $n > \#S = k \geq 1$. We next show that (*) holds for any $S \in \mathcal{K}$ with $\#S = k + 1$. Let $T \in \mathcal{K}$ be such that $T = S \cup \{j\}$, $j \in N \setminus S$, and $n > \#S = k \geq 1$. In what follows, we show that $C(T, A) = \{x \in A : N(x, T, A) \geq N(y, T, A)$ for all $y \in A\}$. Let $C^*(T, A) = \{x \in A : N(x, T, A) \geq N(y, T, A)$ for all $y \in A\}$.

Our first task is to show that $C(T, A) \subseteq C^*(T, A)$. Suppose to the contrary that $C(T, A) \not\subseteq C^*(T, A)$. Then, there exists $a \in A$ such that $[a \in C(T, A)$ and $a \notin C^*(T, A)]$. Since $a \notin C^*(T, A)$ and $C^*(T, A) \neq \emptyset$, it must be the case that $N(a, T, A) < N(z, T, A)$ for some $z \in C^*(T, A)$. It then follows that, for some $p \in T$, $z \in C_p(A)$ and $a \notin C_p(A)$. Consider the coalition $S' = T \setminus \{p\}$. Note that $\#S' = k$ and $C(S', A) = C^*(S', A)$ from the induction hypothesis. We consider two cases: (i) $a \in C(S', A)$ and $a \notin C(S', A)$. Case (i), $a \in C(S', A)$. Note that $[N(a, T, A) < N(z, T, A)$, $z \in C^*(T, A)$, $z \in C_p(A)$ and $a \notin C_p(A)]$. It then follows that $N(a, S', A) = N(z, S', A)$. Consequently, $z \in C(S', A)$. Noting that $C(S', A) \cap C_p(A) \neq \emptyset$, by (M), $C(T, A) = C(S', A) \cap C_p(A)$, implying that $a \notin C(T, A)$, a contradiction. Case (ii), $a \notin C(S', A)$. Note that $T = S' \cap \{p\}$ and $a \notin C_p(A)$. By (UR), $a \notin C(T, A)$, another contradiction. Therefore, $C(T, A) \subseteq C^*(T, A)$.

To complete the proof, we show that $C^*(T, A) \subseteq C(T, A)$. Let $x \in C^*(T, A)$. We consider two cases: case (i), $[x \in C_i(T)$ for all $i \in T]$; and case (ii), $x \notin C_q(A)$ for some $q \in T$. Case (i), $[x \in C_i(T)$ for all $i \in T]$. From induction hypothesis, $x \in C(S, A)$ where $S = T \setminus \{p\}$ and $p \in T$. Note that $x \in C_p(A)$. By (M), it then follows that $x \in C(T, A)$. Case (ii), $x \notin C_q(A)$ for some $q \in T$. Consider the coalition $S' = T \setminus \{q\}$. Note that $x \in C^*(T, A)$. It must be the case that $x \in C^*(S', A)$ implying that $N(x, S', A) \geq N(y, S', A)$ for all $y \in A$ and $x \in C(S', A)$, which follows from the induction hypothesis. Note that it must be true that $C(S', A) \cap C_q(A) = \emptyset$. This is because, if $C(S', A) \cap C_q(A) \neq \emptyset$, then, for some $z \in C(S', A)$, $z \in C_q(A)$, and consequently, $N(z, T, A) > N(x, T, A)$ follows from $x \notin C_q(A)$. Since $C(S', A) \cap C_q(A) = \emptyset$, by PA, $C(S', A) \subseteq C(T, A)$. Note that $x \in C(S', A)$. We then obtain that $x \in C(T, A)$. Therefore, $C^*(T, A) \subseteq C(T, A)$.

Thus, we have shown that $C(T, A) \subseteq C^*(T, A)$ and $C^*(T, A) \subseteq C(T, A)$. Therefore, $C(T, A) = C^*(T, A)$. Thus, (*) is established. This completes the proof. $\diamond$

Proposition 1 *The axioms figured in Theorem 1 are independent.*

Proof Let $x_0 \in A$ be given and let $C^*(\cdot, A)$ be the choice set given by approval voting. Consider the following aggregation rules:

f_1: for all $S \in \mathcal{K}$, $C_1(S, A) = A$

f_2: for all $S \in \mathcal{K}$, $C_2(S, A) = \bigcup_{i \in S} C_i(A)$

f_3: for all $S \in \mathcal{K}$, $C_3(S, A) = \begin{cases} A, \text{ if } C_i(A) \cap C_j(A) = \emptyset \text{ for all distinct } i, j \in S \\ C^*(S, A), \text{ if otherwise} \end{cases}$

f_4: for all $S \in \mathcal{K}$, if $S = \{i\}$ for some $i \in$ then $C_4(S, A) = C_i(A)$, and if $\#S \geq 2$, then $(C_4(S, A) = \{x_0\}$ if $[C_i(A) \cap C_j(A) = \emptyset$ for all distinct $i, j \in S$ and $x_0 \in C_i(A)$ for some $i \in S]$, and $C_4(S, A) = C^*(S, A)$ if otherwise).

It can be checked that f_1 satisfies (M), (UR) and (PA) while violates (SD), f_2 satisfies (SD), (UR) and (PA) but violates (M), f_3 satisfies (SD), (M) and (PA) while violates (UR), and f_4 satisfies (SD), (M) and (UR) but violates (PA). $\diamond$

4 Concluding Remarks

In this paper, we have developed an alternative framework to study approval voting axiomatically. The main feature of our framework is that we work with a single-profile choice functions and variable societies. In such a framework, we have studied approval voting from the perspective that links the choices of a society to the choices of its sub-societies. To put our contribution in perspective, we locate our contribution to the literature by grouping various characterizations of approval voting into the following categories:

1. Variable societies and multi profile of preferences: Fishburn [5, 6] , Sertel [7].
2. Fixed society and multi profile of choice functions: Baigent and Xu [1].
3. Variable societies and a single profile of choice functions: this paper.

Acknowledgements This paper is dedicated to Nick Baigent. Y. Xu would like to thank Nick Baigent for his mentoring and support. An earlier version of the paper was presented at the CEPET Workshop in Honour of Nick Baigent, Udine, Italy, June 2–4, 2010, and we would like to thank the audience for questions and comments. We are also grateful to two referees for their helpful comments.

References

1. Baigent N, Xu Y (1991) Independent, necessary and sufficient conditions for approval voting. Math Soc Sci 21:21–29
2. Brams S, Fishburn P (1978) Approval voting. Am Polit Sci Rev 72(3):831–847
3. Fine B, Fine K (1974) Social choice and individual ranking I. Rev Econ Stud 41:303–322
4. Fine B, Fine K (1974) Social choice and individual ranking II. Rev Econ Stud 41:459–475

5. Fishburn P (1978) Axioms for approval voting: direct proof. J Econ Theory 19:180–185
6. Fishburn P (1978) Symmetric and consistent aggregation with dichotomous preferences. In: Laffont J (ed) Aggregation and revelation of preferences. North- Holland, Amsterdam
7. Sertel M (1988) Characterizing approval voting. J Econ Theory 45:207–211
8. Smith J (1973) Aggregation of preferences with variable electorate. Econometrica 41:1027–1041
9. Xu Y (2010) Axiomatizations of approval voting, Chap 5. In: Laslier J-F, Remzi Sanver M (eds) Handbook on approval voting. Springer, New York
10. Xu Y, Zhong Z (2010) Single profile of preferences with variable societies: a characterization of simple majority rule. Econ Lett 107:119–121
11. Young HP (1974) An axiomatization of Borda's rule. J Econ Theory 9:43–52
12. Young HP (1975) Social choice scoring functions. SIAM J App Math 28:824–838

Nondictatorial Arrovian Social Welfare Functions: An Integer Programming Approach

Francesca Busetto, Giulio Codognato, and Simone Tonin

Abstract In the line opened by Kalai and Muller (J Econ Theory 16:457–469, 1977), we explore new conditions on preference domains which make it possible to avoid Arrow's impossibility result. In our main theorem, we provide a complete characterization of the domains admitting nondictatorial Arrovian social welfare functions with ties (i.e. including indifference in the range) by introducing a notion of strict decomposability. In the proof, we use integer programming tools, following an approach first applied to social choice theory by Sethuraman et al. (Math Oper Res 28:309–326, 2003; J Econ Theory 128:232–254, 2006). In order to obtain a representation of Arrovian social welfare functions whose range can include indifference, we generalize Sethuraman et al.'s work and specify integer programs in which variables are allowed to assume values in the set $\{0, \frac{1}{2}, 1\}$: indeed, we show that there exists a one-to-one correspondence between the solutions of an integer program defined on this set and the set of all Arrovian social welfare functions—without restrictions on the range.

Keywords Arrovian social welfare function • Integer programming • Nondictatorial domain

F. Busetto (✉)
Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Udine, Via Tomadini 30, 33100 Udine, Italy
e-mail: francesca.busetto@uniud.it

G. Codognato
Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Udine, Via Tomadini 30, 33100 Udine, Italy

EconomiX, Université de Paris Ouest Nanterre la Défense, 200 Avenue de la République, 92001 Nanterre Cedex, France
e-mail: giulio.codognato@uniud.it

S. Tonin
Adam Smith Business School, University of Glasgow, Main Building, Glasgow G128QQ, UK
e-mail: s.tonin.1@research.gla.ac.uk

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*, Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_10

1 Introduction

Arrow [1] established his celebrated impossibility theorem for Arrovian Social Welfare Functions (ASWFs)—that is social welfare functions satisfying the hypotheses of Pareto optimality and independence of irrelevant alternatives—defining them on the unrestricted domain of preference orderings. As is well known, this result holds also for ASWFs defined on the domain of all antisymmetric preference orderings. Kalai and Muller [3] dealt with the problem of introducing restrictions on this latter domain of individual preferences in order to overcome Arrow's impossibility result.[1] They gave the first complete characterization of the domains of antisymmetric preference orderings which admit nondictatorial ASWFs "without ties"—that is ASWFs which do not admit indifference between distinct alternatives in their range. They did this by means of two theorems: in their Theorem 1, they showed that there exists a n-person nondictatorial ASWF for a given domain of antisymmetric preference orderings if and only if there exists a 2-person nondictatorial ASWF for the same domain; in their Theorem 2, they gave the domain characterization, by introducing the concept of decomposability.

In this paper, we proceed along the way opened by Kalai and Muller, and explore new conditions on preference domains which allow for the existence of nondictatorial ASWFs. In fact, Kalai and Muller's Theorem 2 provides a complete characterization of the domains of antisymmetric preference orderings admitting nondictatorial ASWFs without ties and of those admitting dictatorial ASWFs without ties. The problem of characterizing the domains of antisymmetric preference orderings admitting nondictatorial ASWFs "with ties"—that is ASWFs which admit indifference between distinct alternatives in their range—has so far been left open. Here, we overcome this problem: in our main theorem, we provide a complete characterization of these domains by introducing the notion of strict decomposability.

We develop our analysis on nondictatorial ASWFs by using the tools of integer programming, first applied to the traditional field of social choice theory by Sethuraman et al. [5, 6]. As remarked by these authors, integer programming is a powerful analytical tool, which makes it possible to derive, in a systematic and simple way, many of the already known theorems on ASWFs, and to prove new results.

In particular, Sethuraman et al. developed Integer Programs (IPs) in which variables assume values only in the set $\{0, 1\}$. Binary IPs of this kind are suitable to be used as an auxiliary tool to represent ASWFs without ties: a fundamental theorem in [5] establishes a one-to-one correspondence, on domains of antisymmetric preference orderings, between the set of feasible solutions of their main binary IP and the set of ASWFs without ties. In both papers mentioned above, Sethuraman et al. used binary integer programming to analyze, among other issues, neutral

[1] Maskin [4] independently investigated the same issue.

and anonymous ASWFs. Moreover, in the 2003 paper, they opened the way to a reconsideration, in terms of integer programming, of the work by Kalai and Muller [3]. In particular, they provided a simplified version of Kalai and Muller's Theorem 1 by using a binary IP.

In this paper, we extend Sethuraman et al's approach in order to obtain a general representation of ASWFs, without restrictions on the range. To this end, we specify IPs in which variables are allowed to assume values in the set $\{0, \frac{1}{2}, 1\}$. We call these programs "ternary IPs," with some abuse with respect to the current specialized literature.[2] Indeed, we provide a theorem establishing that there exists a one-to-one correspondence between the set of feasible solutions of a ternary IP and the set of all ASWFs. Then, we exploit these generalized integer programs as a basic tool to show our characterization theorem on ASWFs with ties.

This new characterization result raises the question of which is the relationship between decomposable and strictly decomposable domains. We point out a redundant condition in the notion of decomposability proposed by Kalai and Muller [3] and conclude our analysis showing that all strictly decomposable domains are decomposable whereas the converse relation does not hold.

2 Notation and Definitions

Let E be any initial finite subset of the natural numbers with at least two elements and let $|E|$ be the cardinality of E, denoted by n. Elements of E are called agents.

Let $\mathcal{E}$ be the collection of all subsets of E. Given a set $S \in \mathcal{E}$, let $S^c = E \setminus S$.

Let $\mathcal{A}$ be a set such that $|\mathcal{A}| \geq 3$. Elements of $\mathcal{A}$ are called alternatives.

Let $\mathcal{A}^2$ denote the set of all ordered pairs of alternatives.

Let $\mathcal{R}$ be the set of all the complete and transitive binary relations on $\mathcal{A}$, called preference orderings.

Let Σ be the set of all antisymmetric preference orderings.

Let Ω denote a nonempty subset of Σ. An element of Ω is called admissible preference ordering and is denoted by $\mathbf{p}$. We write $x\mathbf{p}y$ if x is ranked above y under $\mathbf{p}$.

A pair $(x, y) \in \mathcal{A}^2$ is called trivial if there are not $\mathbf{p}, \mathbf{q} \in \Omega$ such that $x\mathbf{p}y$ and $y\mathbf{q}x$. Let *TR* denote the set of trivial pairs. We adopt the convention that all pairs $(x, x) \in \mathcal{A}^2$ are trivial.

A pair $(x, y) \in \mathcal{A}^2$ is nontrivial if it is not trivial. Let *NTR* denote the set of nontrivial pairs.

[2] We have to stress that we still apply the basic tools of integer linear programming and that the programs we introduce could be equivalently defined on the set $\{0, 1, 2\}$. Nonetheless, here we prefer to follow Sethuraman et al. [6], and keep using the value $\frac{1}{2}$ in order to incorporate indifference between social alternatives into the analysis.

Let Ω^n denote the n-fold Cartesian product of Ω. An element of Ω^n is called a preference profile and is denoted by $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n)$, where $\mathbf{p}_i$ is the antisymmetric preference ordering of agent $i \in E$.

A Social Welfare Function (SWF) on Ω is a function $f : \Omega^n \to \mathcal{R}$.

f is said to be "without ties" if $f(\Omega^n) \cap (\mathcal{R} \setminus \Sigma) = \emptyset$.

f is said to be "with ties" if $f(\Omega^n) \cap (\mathcal{R} \setminus \Sigma) \neq \emptyset$.

Given $\mathbf{P} \in \Omega^n$, let $P(f(\mathbf{P}))$ and $I(f(\mathbf{P}))$ be binary relations on $\mathcal{A}$. We write $xP(f(\mathbf{P}))y$ if, for $x, y \in \mathcal{A}$, $xf(\mathbf{P})y$ but not $yf(\mathbf{P})x$ and $xI(f(\mathbf{P}))y$ if, for $x, y \in \mathcal{A}$, $xf(\mathbf{P})y$ and $yf(\mathbf{P})x$.

A SWF on Ω, f, satisfies Pareto Optimality (PO) if, for all $(x, y) \in \mathcal{A}^2$ and for all $\mathbf{P} \in \Omega^n$, $x\mathbf{p}_i y$, for all $i \in E$, implies $xP(f(\mathbf{P}))y$.

A SWF on Ω, f, satisfies Independence of Irrelevant Alternatives (IIA) if, for all $(x, y) \in NTR$ and for all $\mathbf{P}, \mathbf{P}' \in \Omega^n$, $x\mathbf{p}_i y$ if and only if $x\mathbf{p}'_i y$, for all $i \in E$, implies, $xf(\mathbf{P})y$ if and only if $xf(\mathbf{P}')y$, and $yf(\mathbf{P})x$ if and only if $yf(\mathbf{P}')x$.

An Arrovian Social Welfare Function (ASWF) on Ω is a SWF on Ω, f, which satisfies PO and IIA.

An ASWF on Ω, f, is dictatorial if there exists $j \in E$ such that, for all $(x, y) \in NTR$ and for all $\mathbf{P} \in \Omega^n$, $x\mathbf{p}_j y$ implies $xP(f(\mathbf{P}))y$. f is nondictatorial if it is not dictatorial.

Given $(x, y) \in \mathcal{A}^2$ and $S \in \mathcal{E}$, let $d_S(x, y)$ denote a variable such that $d_S(x, y) \in \{0, \frac{1}{2}, 1\}$.

An Integer Program (IP) on Ω consists of a set of linear constraints, related to the preference orderings in Ω, on variables $d_S(x, y)$, for all $(x, y) \in NTR$ and for all $S \in \mathcal{E}$, and of the further conventional constraints that $d_E(x, y) = 1$ and $d_\emptyset(y, x) = 0$, for all $(x, y) \in TR$.

Let d denote a feasible solution (henceforth, for simplicity, only "solution") to an IP on Ω. d is said to be a binary solution if variables $d_S(x, y)$ reduce to assume values in the set $\{0, 1\}$, for all $(x, y) \in NTR$, and for all $S \in \mathcal{E}$. It is said to be a "ternary" solution, otherwise.

A solution d is dictatorial if there exists $j \in E$ such that $d_S(x, y) = 1$, for all $(x, y) \in NTR$ and for all $S \in \mathcal{E}$, with $j \in S$. d is nondictatorial if it is not dictatorial.

An ASWF on Ω, f, and a solution to an IP on the same Ω, d, are said to correspond if, for each $(x, y) \in NTR$ and for each $S \in \mathcal{E}$, $xP(f(\mathbf{P}))y$ if and only if $d_S(x, y) = 1$, $xI(f(\mathbf{P}))y$ if and only if $d_S(x, y) = \frac{1}{2}$, $yP(f(\mathbf{P}))x$ if and only if $d_S(x, y) = 0$, for all $\mathbf{P} \in \Omega^n$ such that $x\mathbf{p}_i y$, for all $i \in S$, and $y\mathbf{p}_i x$, for all $i \in S^c$.

3 Arrovian Social Welfare Functions and Ternary Integer Programming: A Correspondence Theorem

The first formulation of an IP on Ω was proposed by Sethuraman et al. [5], for the case where $d_S(x, y) \in \{0, 1\}$, for all $(x, y) \in NTR$ and for all $S \in \mathcal{E}$. Moreover, in both their 2003 and 2006 papers, they used binary IPs on Ω to provide

a representation of ASWFs different from the axiomatic one previously used in the Arrow's tradition.

In this section, we extend Sethuraman et al.'s approach, specifying two integer programs in which variables $d_S(x, y)$ are allowed to assume values in the set $\{0, \frac{1}{2}, 1\}$. We will show that these ternary programs on Ω can be used to provide a general representation of ASWFs, with and without ties in the range. Our first IP on Ω—called IP1—consists of the following set of constraints:

$$d_E(x, y) = 1, \tag{1}$$

for all $(x, y) \in NTR$;

$$d_S(x, y) + d_{S^c}(y, x) = 1, \tag{2}$$

for all $(x, y) \in NTR$ and for all $S \in \mathcal{E}$;

$$d_{A \cup U \cup V}(x, y) + d_{B \cup U \cup W}(y, z) + d_{C \cup V \cup W}(z, x) \leq 2, \tag{3}$$

if $d_{A \cup U \cup V}(x, y), d_{B \cup U \cup W}(y, z), d_{C \cup V \cup W}(z, x) \in \{0, 1\}$;

$$d_{A \cup U \cup V}(x, y) + d_{B \cup U \cup W}(y, z) + d_{C \cup V \cup W}(z, x) = \frac{3}{2}, \tag{4}$$

if $d_{A \cup U \cup V}(x, y) = \frac{1}{2}$ or $d_{B \cup U \cup W}(y, z) = \frac{1}{2}$ or $d_{C \cup V \cup W}(z, x) = \frac{1}{2}$, for all triples of alternatives x, y, z and for all disjoint and possibly empty sets $A, B, C, U, V, W \in \mathcal{E}$ whose union includes all agents and which satisfy the following conditions, drawn from [5], and hereafter referred to as Conditions $(*)$:

$$A \neq \emptyset \text{ only if there exists } \mathbf{p} \in \Omega \text{ such that } x\mathbf{p}z\mathbf{p}y,$$

$$B \neq \emptyset \text{ only if there exists } \mathbf{p} \in \Omega \text{ such that } y\mathbf{p}x\mathbf{p}z,$$

$$C \neq \emptyset \text{ only if there exists } \mathbf{p} \in \Omega \text{ such that } z\mathbf{p}y\mathbf{p}x,$$

$$U \neq \emptyset \text{ only if there exists } \mathbf{p} \in \Omega \text{ such that } x\mathbf{p}y\mathbf{p}z,$$

$$V \neq \emptyset \text{ only if there exists } \mathbf{p} \in \Omega \text{ such that } z\mathbf{p}x\mathbf{p}y,$$

$$W \neq \emptyset \text{ only if there exists } \mathbf{p} \in \Omega \text{ such that } y\mathbf{p}z\mathbf{p}x.$$

In fact, we propose now a result which establishes a one-to-one correspondence between the set of the solutions to IP1 on a given Ω and the set of all ASWFs on the same Ω.

Theorem 1 *Consider a domain Ω. Given an ASWF on Ω, f, there exists a unique solution to IP1 on Ω, d, which corresponds to f. Given a solution to IP1 on Ω, d, there exists a unique ASWF on Ω, f, which corresponds to d.*

Proof Consider a domain Ω and an ASWF on Ω, f. Determine d as follows. Given $(x, y) \in NTR$ and $S \in \mathcal{E}$, consider $\mathbf{P} \in \Omega^n$ such that $x\mathbf{p}_i y$, for all $i \in S$, and $y\mathbf{p}_i x$, for all $i \in S^c$. Let $d_S(x, y) = 1$ if $xP(f(\mathbf{P}))y$, $d_S(x, y) = \frac{1}{2}$ if $xI(f(\mathbf{P}))y$, $d_S(x, y) = 0$ if $yP(f(\mathbf{P}))x$. Then, for each $(x, y) \in NTR$ and for each $S \in \mathcal{E}$, we have $xP(f(\mathbf{P}))y$ if and only if $d_S(x, y) = 1$, $xI(f(\mathbf{P}))y$ if and only if $d_S(x, y) = \frac{1}{2}$, $yP(f(\mathbf{P}))x$ if and only if $d_S(x, y) = 0$, for all $\mathbf{P} \in \Omega^n$ such that $x\mathbf{p}_i y$, for all $i \in S$, and $y\mathbf{p}_i x$, for all $i \in S^c$, as f satisfies IIA. d satisfies (1), as $f(\mathbf{P})$ satisfies PO, and (2), as $f(\mathbf{P})$ is a complete binary relation on $\mathcal{A}$, for all $\mathbf{P} \in \Omega^n$. Consider a triple x, y, z, and disjoint and possibly empty sets $A, B, C, U, V, W \in \mathcal{E}$ whose union includes all agents and which satisfy Conditions $(*)$. Moreover, consider $\mathbf{P} \in \Omega^n$. Then, by Conditions $(*)$, we have: $x\mathbf{p}_i y$, for all $i \in A \cup U \cup V$; $y\mathbf{p}_i x$, for all $i \in (A \cup U \cup V)^c$; $y\mathbf{p}_i z$, for all $i \in B \cup U \cup W$; $z\mathbf{p}_i y$, for all $i \in (B \cup U \cup W)^c$; $z\mathbf{p}_i x$, for all $i \in C \cup V \cup W$; $x\mathbf{p}_i z$, for all $i \in (C \cup V \cup W)^c$. Suppose that $d_{A\cup U\cup V}(x, y), d_{B\cup U\cup W}(y, z), d_{C\cup V\cup W}(z, x) \in \{0, 1\}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > 2.$$

Then, we have $xP(f(\mathbf{P}))yP(f(\mathbf{P}))z$ and $zP(f(\mathbf{P}))x$, a contradiction. Suppose that $d_{A\cup U\cup V}(x, y) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) < \frac{3}{2}.$$

Consider the following three cases. First, $d_{B\cup U\cup W}(y, z) = 0$ and $d_{C\cup V\cup W}(z, x) = 0$. Then, we have $zP(f(\mathbf{P}))yI(f(\mathbf{P}))x$ and $xP(f(\mathbf{P}))z$, a contradiction. Second, $d_{B\cup U\cup W}(y, z) = \frac{1}{2}$ and $d_{C\cup V\cup W}(z, x) = 0$. Then, we have $xI(f(\mathbf{P}))yI(f(\mathbf{P}))z$ and $xP(f(\mathbf{P}))z$, a contradiction. Third, $d_{B\cup U\cup W}(y, z) = 0$ and $d_{C\cup V\cup W}(z, x) = \frac{1}{2}$. Then, we have $zI(f(\mathbf{P}))xI(f(\mathbf{P}))y$ and $zP(f(\mathbf{P}))y$, a contradiction. Suppose now that $d_{A\cup U\cup V}(x, y) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > \frac{3}{2}.$$

Consider the following three cases. First, $d_{B\cup U\cup W}(y, z) = 1$ and $d_{C\cup V\cup W}(z, x) = 1$. Then, we have $xI(f(\mathbf{P}))yP(f(\mathbf{P}))z$ and $zP(f(\mathbf{P}))x$, a contradiction. Second, $d_{B\cup U\cup W}(y, z) = \frac{1}{2}$ and $d_{C\cup V\cup W}(z, x) = 1$. Then, we have $xI(f(\mathbf{P}))yI(f(\mathbf{P}))z$ and $zP(f(\mathbf{P}))x$, a contradiction. Third, $d_{B\cup U\cup W}(y, z) = 1$ and $d_{C\cup V\cup W}(z, x) = \frac{1}{2}$. Then, we have $xI(f(\mathbf{P}))yP(f(\mathbf{P}))z$ and $zI(f(\mathbf{P}))x$, a contradiction. Therefore, d satisfies (3) and (4). Hence, d is a solution to IP1 on Ω which corresponds to f. Suppose that d is not unique. Then, there exist a solution to IP1 on Ω, d', $(x, y) \in NTR$, and $S \in \mathcal{E}$ such that $d_S(x, y) \neq d'_S(x, y)$. Consider $\mathbf{P} \in \Omega^n$ such that $x\mathbf{p}_i y$, for all $i \in S$, and $y\mathbf{p}_i x$, for all $i \in S^c$. Then, we have $xP(f(\mathbf{P}))y$ and $xI(f(\mathbf{P}))y$, or, $yP(f(\mathbf{P}))x$ and $xI(f(\mathbf{P}))y$, or, $xP(f(\mathbf{P}))y$ and $yP(f(\mathbf{P}))x$, a contradiction. But then, d is unique. Now, consider a solution to IP1 on Ω, d. Determine f as follows. Given $(x, y) \in TR$, let $xP(f(\mathbf{P}))y$, for all $\mathbf{P} \in \Omega^n$.

Given $(x, y) \in NTR$ and $\mathbf{P} \in \Omega^n$, let $S \in \mathcal{E}$ be the set of agents such that $x\mathbf{p}_i y$, for all $i \in S$, and $y\mathbf{p}_i x$, for all $i \in S^c$. Let $xP(f(\mathbf{P}))y$ if $d_S(x, y) = 1$, $xI(f(\mathbf{P}))y$ if $d_S(x, y) = \frac{1}{2}$, and $yP(f(\mathbf{P}))x$ if $d_S(x, y) = 0$. $f(\mathbf{P})$ is a complete binary relation on $\mathcal{A}$, for all $\mathbf{P} \in \Omega^n$, by construction and by (2). Now, we show that $f(\mathbf{P})$ is also a transitive binary relation on $\mathcal{A}$, for all $\mathbf{P} \in \Omega^n$. Consider a triple x, y, z and a preference profile $\mathbf{P} \in \Omega^n$. Then, there exist three nonempty sets H, I, J such that $x\mathbf{p}_i y$, for all $i \in H$, $y\mathbf{p}_i x$, for all $i \in H^c$, $y\mathbf{p}_i z$, for all $i \in I$, $z\mathbf{p}_i y$, for all $i \in I^c$, $z\mathbf{p}_i x$, for all $i \in J$, $x\mathbf{p}_i z$, for all $i \in J^c$. Let $A = H \setminus (I \cup J)$, $B = I \setminus (H \cup J)$, $C = J \setminus (H \cup I)$, $U = H \cap I$, $V = H \cap J$, $W = I \cap J$. Then, $A, B, C, U, V, W \in \mathcal{E}$ are disjoint sets of agents whose union includes all agents and which satisfy Conditions $(*)$. Moreover, they satisfy $A \cup U \cup V = H$, $B \cup U \cup W = I$, $C \cup V \cup W = J$. Consider the following eight cases. First, $xP(f(\mathbf{P}))yP(f(\mathbf{P}))z$ and $zP(f(\mathbf{P}))x$. Then, $d_{A\cup U\cup V}(x, y) = 1$, $d_{B\cup U\cup W}(y, z) = 1$, $d_{C\cup V\cup W}(z, x) = 1$, and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > 2,$$

contradicting (3). Second, $xP(f(\mathbf{P}))yP(f(\mathbf{P}))z$ and $xI(f(\mathbf{P}))z$. Then, $d_{C\cup V\cup W}(z, x) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > \frac{3}{2},$$

contradicting (4). Third, $xI(f(\mathbf{P}))yP(f(\mathbf{P}))z$ and $zP(f(\mathbf{P}))x$. Then, $d_{A\cup U\cup V}(x, y) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > \frac{3}{2},$$

contradicting (4). Fourth, $xI(f(\mathbf{P}))yP(f(\mathbf{P}))z$ and $xI(f(\mathbf{P}))z$. Then, $d_{A\cup U\cup V}(x, y) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > \frac{3}{2},$$

contradicting (4). Fifth, $xP(f(\mathbf{P}))yI(f(\mathbf{P}))z$ and $zP(f(\mathbf{P}))x$. Then, $d_{B\cup U\cup W}(y, z) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > \frac{3}{2},$$

contradicting (4). Sixth, $xP(f(\mathbf{P}))yI(f(\mathbf{P}))z$ and $xI(f(\mathbf{P}))z$. Then, $d_{B\cup U\cup W}(y, z) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > \frac{3}{2},$$

contradicting (4). Seventh, $xI(f(\mathbf{P}))yI(f(\mathbf{P}))z$ and $xP(f(\mathbf{P}))z$. Then, $d_{A\cup U\cup V}(x, y) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) < \frac{3}{2},$$

contradicting (4). Eighth, $xI(f(\mathbf{P}))yI(f(\mathbf{P}))z$ and $zP(f(\mathbf{P}))x$. Then, $d_{A\cup U\cup V}(x, y) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > \frac{3}{2},$$

contradicting (4). f satisfies PO as, for all $(x, y) \in TR$, we have $xP(f(\mathbf{P}))y$, for all $\mathbf{P} \in \Omega^n$; moreover, for all $(x, y) \in NTR$ and for all $\mathbf{P} \in \Omega^n$, $x\mathbf{p}_i y$, for all $i \in E$, implies $xP(f(\mathbf{P}))y$, by (1). f satisfies IIA as, for each $(x, y) \in NTR$ and for each $S \in \mathcal{E}$, we have $xP(f(\mathbf{P}))y$ if and only if $d_S(x, y) = 1$, $xI(f(\mathbf{P}))y$ if and only if $d_S(x, y) = \frac{1}{2}$, and $yP(f(\mathbf{P}))x$ if and only if $d_S(x, y) = 0$, for all $\mathbf{P} \in \Omega^n$ such that $x\mathbf{p}_i y$, for all $i \in S$, and $y\mathbf{p}_i x$, for all $i \in S^c$. Hence, f is an ASWF on Ω, which corresponds to d. Suppose that f is not unique. Then, there exists an ASWF on Ω, f', $(x, y) \in NTR$ and $\mathbf{P} \in \Omega^n$ such that we have $xf(\mathbf{P})y$ but not $xf'(\mathbf{P})y$. Let $S \in \mathcal{E}$ be the set such that $x\mathbf{p}_i y$, for all $i \in S$, and $y\mathbf{p}_i x$, for all $i \in S^c$. Then, $d_S(x, y) = 1$ and $d_S(x, y) = 0$, or, $d_S(x, y) = \frac{1}{2}$ and $d_S(x, y) = 0$, a contradiction. But then, f is unique. ■

We introduce now a second ternary IP on Ω, which we will call IP2. It consists of constraints (1), (2), and the following four logically independent constraints[3]:

$$d_S(x, y) \leq d_S(x, z), \tag{5}$$

if $d_S(x, y) \in \{0, 1\}$;

$$d_S(x, y) < d_S(x, z), \tag{6}$$

if $d_S(x, y) = \frac{1}{2}$, for all triples x, y, z such that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$, and for all $S \in \mathcal{E}$;

$$d_S(x, y) + d_S(y, z) \leq 1 + d_S(x, z), \tag{7}$$

[3] In building IP2, we take inspiration from a binary IP on Ω, introduced by Sethuraman et al. [5], which incorporates a reformulation of Kalai and Muller's condition of decomposability. It can be shown that the set of constraints proposed by Sethuraman et al. exhibits problems of logical dependence (see Busetto and Codognato [2]), which are eliminated in our IP2. These problems parallel some logical redundancies inherent in Kalai and Muller's notion of decomposability, which we will point out in Sect. 4.

if $d_S(x, y), d_S(y, z) \in \{0, 1\}$;

$$d_S(x, y) + d_S(y, z) = \frac{1}{2} + d_S(x, z), \tag{8}$$

if $d_S(x, y) = \frac{1}{2}$ or $d_S(y, z) = \frac{1}{2}$, for all triples x, y, z such that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$, and for all $S \in \mathcal{E}$.

In the remainder of this section, we prove two propositions which establish the relationships between IP1 and IP2.

Proposition 1 *If d is a solution to IP1 on Ω, then it is a solution to IP2 on the same Ω.*

Proof Let d be a solution to IP1 on Ω. Consider a triple x, y, z and $S \in \mathcal{E}$. Suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ which satisfy $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$. Let $U = S$, $W = S^c$, and $A = B = C = V = \emptyset$. Then, A, B, C, U, V, W are sets whose union includes all agents and which satisfy Conditions $(*)$. Suppose that $d_S(x, y) \in \{0, 1\}$ and $d_S(x, y) > d_S(x, z)$. Consider the following two cases. First, $d_S(x, z) \in \{0, 1\}$. Then,

$$d_U(x, y) + d_{U \cup W}(y, z) + d_W(z, x) > 2,$$

contradicting (3). Second, $d_S(x, z) = \frac{1}{2}$. Then,

$$d_U(x, y) + d_{U \cup W}(y, z) + d_W(z, x) > \frac{3}{2},$$

contradicting (4). Therefore, d satisfies (5). Suppose now that $d_S(x, y) = \frac{1}{2}$ and $d_S(x, y) \geq d_S(x, z)$. Then,

$$d_U(x, y) + d_{U \cup W}(y, z) + d_W(z, x) > \frac{3}{2},$$

contradicting (4). Therefore, d satisfies (6). Consider a triple x, y, z and $S \in \mathcal{E}$. Suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Let $C = S^c$, $U = S$, and $A = B = V = W = \emptyset$. Then, A, B, C, U, V, W are sets whose union includes all agents and which satisfy Conditions $(*)$. Suppose that $d_S(x, y), d_S(y, z) \in \{0, 1\}$ and $d_S(x, y) + d_S(y, z) > 1 + d_S(x, z)$. Consider the following two cases. First, $d_S(x, z) \in \{0, 1\}$. Then,

$$d_U(x, y) + d_U(y, z) + d_C(z, x) > 2,$$

contradicting (3). Second, $d_S(x, z) = \frac{1}{2}$. Then,

$$d_U(x, y) + d_U(y, z) + d_C(z, x) > \frac{3}{2},$$

contradicting (4). Therefore, d satisfies (7). Suppose now that $d_S(x, y) = \frac{1}{2}$ and $d_S(x, y) + d_S(y, z) < \frac{1}{2} + d_S(x, z)$. Then,

$$d_U(x, y) + d_U(y, z) + d_C(z, x) < \frac{3}{2},$$

contradicting (4). Suppose that $d_S(x, y) = \frac{1}{2}$ and $d_S(x, y) + d_S(y, z) > \frac{1}{2} + d_S(x, z)$. Then,

$$d_U(x, y) + d_U(y, z) + d_C(z, x) > \frac{3}{2},$$

contradicting (4). Therefore, d satisfies (8). Hence, d is a solution to IP2 on Ω. ■

The following result shows that the converse of Proposition 3 holds—and IP1 and IP2 coincide—when $n = 2$.

Proposition 2 *Let $n = 2$. If d is a solution to IP2 on Ω, then it is a solution to IP1 on the same Ω.*

Proof Let $n = 2$. Let d be a solution to IP2 on Ω. Consider a triple x, y, z and disjoint and possibly empty sets $A, B, C, U, V, W \in \mathcal{E}$ whose union includes all agents and which satisfy Conditions $(*)$. Suppose that $d_{A\cup U\cup V}(x, y), d_{B\cup U\cup W}(y, z), d_{C\cup V\cup W}(z, x) \in \{0, 1\}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) > 2.$$

Consider the case where $A \neq \emptyset$ and $W \neq \emptyset$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}z\mathbf{p}y$ and $y\mathbf{q}z\mathbf{q}x$. Suppose that $A = \{1\}$ and $W = \{2\}$. Then,

$$d_{\{2\}}(y, z) + d_{\{2\}}(z, x) > 1 + d_{\{2\}}(y, x),$$

contradicting (7). The cases where $B \neq \emptyset$, $V \neq \emptyset$, and $C \neq \emptyset$, $U \neq \emptyset$ lead, *mutatis mutandis*, to the same contradiction. Consider the case where $U \neq \emptyset$ and $V \neq \emptyset$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}x\mathbf{q}y$. Suppose that $U = \{1\}$ and $V = \{2\}$. Then,

$$d_{\{2\}}(z, x) > d_{\{2\}}(z, y),$$

contradicting (5). The cases where $V \neq \emptyset$, $W \neq \emptyset$, and $U \neq \emptyset$, $W \neq \emptyset$, lead, *mutatis mutandis*, to the same contradiction. Therefore, d satisfies (3). Suppose that $d_{A\cup U\cup V}(x, y) = \frac{1}{2}$ and

$$d_{A\cup U\cup V}(x, y) + d_{B\cup U\cup W}(y, z) + d_{C\cup V\cup W}(z, x) < \frac{3}{2}.$$

Consider the case where $A \neq \emptyset$ and $B \neq \emptyset$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}z\mathbf{p}y$ and $y\mathbf{q}x\mathbf{q}z$. Suppose that $A = \{1\}$ and $B = \{2\}$. Then, $d_{\{2\}}(y, x) = \frac{1}{2}$ and

$$d_{\{2\}}(y, x) \geq d_{\{2\}}(y, z),$$

contradicting (6). The case where $A \neq \emptyset$ and $C \neq \emptyset$ leads, *mutatis mutandis*, to the same contradiction. Consider the case where $A \neq \emptyset$ and $W \neq \emptyset$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}z\mathbf{p}y$ and $y\mathbf{q}z\mathbf{q}x$. Suppose that $A = \{1\}$ and $W = \{2\}$. Suppose that $d_{\{2\}}(y, z) = 0$ and $d_{\{2\}}(z, x) = 0$. Then,

$$d_{\{1\}}(x, z) + d_{\{1\}}(z, y) > 1 + d_{\{1\}}(x, y),$$

contradicting (7). Suppose that $d_{\{2\}}(y, z) = \frac{1}{2}$ and $d_{\{2\}}(z, x) = 0$. Then,

$$d_{\{2\}}(y, z) + d_{\{2\}}(z, x) < \frac{1}{2} + d_{\{2\}}(y, x),$$

contradicting (8). Consider the case where $U \neq \emptyset$ and $C \neq \emptyset$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Suppose that $U = \{1\}$ and $C = \{2\}$. Then, $d_{\{1\}}(x, y) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) < \frac{1}{2} + d_{\{1\}}(x, z),$$

contradicting (8). The case where $V \neq \emptyset$ and $B \neq \emptyset$ leads, *mutatis mutandis*, to the same contradiction. Suppose that $d_{A \cup U \cup V}(x, y) = \frac{1}{2}$ and

$$d_{A \cup U \cup V}(x, y) + d_{B \cup U \cup W}(y, z) + d_{C \cup V \cup W}(z, x) > \frac{3}{2}.$$

Consider the case where $A \neq \emptyset$ and $W \neq \emptyset$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}z\mathbf{p}y$ and $y\mathbf{q}z\mathbf{q}x$. Suppose that $A = \{1\}$ and $W = \{2\}$. Suppose that $d_{\{2\}}(y, z) = 1$ and $d_{\{2\}}(z, x) = 1$. Then,

$$d_{\{2\}}(y, z) + d_{\{2\}}(z, x) > 1 + d_{\{2\}}(y, x),$$

contradicting (7). Suppose that $d_{\{2\}}(y, z) = \frac{1}{2}$ and $d_{\{2\}}(z, x) = 1$. Then,

$$d_{\{2\}}(y, z) + d_{\{2\}}(z, x) > \frac{1}{2} + d_{\{2\}}(y, x),$$

contradicting (8). Consider the case where $U \neq \emptyset$ and $C \neq \emptyset$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Suppose that $U = \{1\}$ and $C = \{2\}$.Then,

$d_{\{1\}}(x, y) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) > \frac{1}{2} + d_{\{1\}}(x, z),$$

contradicting (8). The case where $V \neq \emptyset$ and $B \neq \emptyset$ leads, *mutatis mutandis*, to the same contradiction. Consider the case where $U \neq \emptyset$ and $W \neq \emptyset$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$. Suppose that $U = \{1\}$ and $W = \{2\}$. Then, $d_{\{1\}}(x, y) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) \geq d_{\{1\}}(x, z),$$

contradicting (6). The case where $V \neq \emptyset$ and $W \neq \emptyset$ leads, *mutatis mutandis*, to the same contradiction. Therefore, d satisfies (4). Hence, d is a solution to IP1 on Ω. ■

4 Nondictatorial Arrovian Social Welfare Functions with Ties and Integer Programming: A New Characterization Theorem

In this section, we use the integer programs developed above to deal with the issues concerning the dictatorship property of ASWFs. As already reminded, Arrow's impossibility theorem is established for ASWFs admitting ties in their range and defined on the unrestricted domain of preference orderings.

Kalai and Muller [3] were the first who overcome Arrow's impossibility theorem by providing a complete characterization of the domains of antisymmetric preference orderings which admit nondictatorial ASWFs without ties. They did this by means of two theorems. In their Theorem 1, they showed that, for a given domain Ω, there exists a nondictatorial ASWF without ties for $n > 2$ if and only if, for the same Ω, there exists a nondictatorial ASWF without ties for $n = 2$. In their Theorem 2, they gave the domain characterization, based on the following notion of decomposability, henceforth called KM-decomposability.

Ω is said to be KM-decomposable if there exists a set R, with $TR \subsetneqq R \subsetneqq \mathcal{A}^2$, satisfying the following conditions.

Condition I *For every two pairs* $(x, y), (x, z) \in NTR$, *if there exist* $\mathbf{p}, \mathbf{q} \in \Omega$ *for which* $x\mathbf{p}y\mathbf{p}z$ *and* $y\mathbf{q}z\mathbf{q}x$, *then* $(x, y) \in R$ *implies that* $(x, z) \in R$.

Condition II *For every two pairs* $(x, y), (x, z) \in NTR$, *if there exist* $\mathbf{p}, \mathbf{q} \in \Omega$ *for which* $x\mathbf{p}y\mathbf{p}z$ *and* $y\mathbf{q}z\mathbf{q}x$, *then* $(z, x) \in R$ *implies that* $(y, x) \in R$.

Condition III *For every two pairs* $(x, y), (x, z) \in NTR$, *if there exists* $\mathbf{p} \in \Omega$ *for which* $x\mathbf{p}y\mathbf{p}z$, *then* $(x, y) \in R$ *and* $(y, z) \in R$ *imply that* $(x, z) \in R$.

Condition IV For every two pairs $(x, y), (x, z) \in NTR$*, if there exists* $\mathbf{p} \in \Omega$ *for which* $x\mathbf{p}y\mathbf{p}z$*, then* $(z, x) \in R$ *implies that* $(y, x) \in R$ *or* $(z, y) \in R$*.*

It is useful to reproduce here Kalai and Muller's characterization theorem for ASWFs without ties. It can be stated as follows.

Theorem 2 *There exists a nondictatorial ASWF without ties on* Ω*,* f*, for* $n \geq 2$*, if and only if* Ω *is KM-decomposable.*

The fundamental aim of this section is taking a step forward along the way opened by Kalai and Muller: our main theorem establishes a characterization of the domains of antisymmetric preference orderings admitting nondictatorial ASWFs with ties.

In order to prove it, we need to establish some preliminary results. To begin with, let us reconsider Kalai and Muller's Theorem 1: Sethuraman et al. [5] provided a reformulation of this theorem in terms of integer programming. More precisely, they established a biunivocal relation between the nondictatorial solutions of a binary IP on Ω, for $n = 2$, and its nondictatorial solutions for $n > 2$. Here, we extend this result to the case of ternary solutions to IP1.

Theorem 3 *There exists a nondictatorial ternary solution to IP1 on* Ω*,* d*, for* $n = 2$*, if and only if there exists a nondictatorial ternary solution to IP1 on* Ω*,* d^**, for* $n > 2$*.*

Proof Let d be a nondictatorial ternary solution to IP1 on Ω for $n = 2$. Determine d^* as follows. Given $(x, y) \in NTR$ and $S \in \mathcal{E}$, let $d^*_S(x, y) = 1$ if $1, 2 \in S$; $d_S(x, y) = 0$ if $1, 2 \in S^c$; $d^*_S(x, y) = d_{\{1\}}(x, y)$ and $d^*_{S^c}(y, x) = d_{\{2\}}(y, x)$ if $1 \in S$ and $2 \in S^c$. Then, it is straightforward to verify that d^* satisfies (1)–(4) and that is nondictatorial. Hence, d^* is a nondictatorial ternary solution to IP1 on Ω, for $n > 2$. Conversely, let d^* be a nondictatorial ternary solution to IP1 on Ω for $n > 2$. Determine d as follows. Consider $(u, v) \in NTR$ and $\bar{S} \in \mathcal{E}$ such that $d^*_{\bar{S}}(u, v) = \frac{1}{2}$. Given $(x, y) \in NTR$, let $d_{\{1,2\}}(x, y) = 1$, $d_{\emptyset}(x, y) = 0$, $d_{\{1\}}(x, y) = d^*_{\bar{S}}(x, y)$, $d_{\{2\}}(y, x) = d^*_{\bar{S}^c}(y, x)$. Then, it is straightforward to verify that d satisfies (1) and (2). Moreover, by Proposition 1, d satisfies (5)–(8) as d^* is a solution to IP1 on Ω. But then, d is a solution to IP2 on Ω and this, in turn, implies that it is a solution to IP1 on Ω, by Proposition 2. Finally, d is nondictatorial as $d_{\{1\}}(u, v) = \frac{1}{2}$. Hence, d is a nondictatorial ternary solution to IP1 on Ω, for $n = 2$. ■

From Theorem 3, we obtain the following corollary, which extends Kalai and Muller's Theorem 1 to the case of ASWFs with ties. It is an immediate consequence of our Theorem 1 in Sect. 3.

Corollary *There exists a nondictatorial ASWF with ties on* Ω*,* f*, for* $n = 2$*, if and only if there exists a nondictatorial ASWF with ties on* Ω*,* f^**, for* $n > 2$*.*

At this point, we need to introduce a reformulation of the concept of KM-decomposability suitable to be applied within the analytical context of a ternary IP on Ω. We will show below that this reformulation is equivalent to the original

version proposed by Kalai and Muller. Our concept is based on the existence of two sets, $R_1, R_2 \in \mathcal{A}^2$—instead of only one—satisfying the restrictions introduced here.

Given a set $R \subset \mathcal{A}^2$, consider the following conditions on R.

Condition 1 *For all triples* x, y, z, *if there exist* $\mathbf{p}, \mathbf{q} \in \Omega$ *satisfying* $x\mathbf{p}y\mathbf{p}z$ *and* $y\mathbf{q}z\mathbf{q}x$, *then* $(x, y) \in R$ *implies that* $(x, z) \in R$.

Condition 2 *For all triples* x, y, z, *if there exist* $\mathbf{p}, \mathbf{q} \in \Omega$ *satisfying* $x\mathbf{p}y\mathbf{p}z$ *and* $z\mathbf{q}y\mathbf{q}x$, *then* $(x, y) \in R$ *and* $(y, z) \in R$ *imply that* $(x, z) \in R$.

A domain Ω is said to be decomposable if there exist two sets R_1 and R_2, with $\emptyset \subsetneqq R_i \subsetneqq NTR$, $i = 1, 2$, such that, for all $(x, y) \in NTR$, we have $(x, y) \in R_1$ if and only if $(y, x) \notin R_2$; moreover, R_i, $i = 1, 2$, satisfies Conditions 1 and 2.

With regard to this definition of a decomposable domain, let us notice the main differences with Kalai and Muller's original notion, introduced to make it compatible with the integer programming analytical setting: Conditions 1 and 2 differ from the corresponding Conditions I and III as the former refer to triples, rather than pairs, of alternatives. Moreover, Condition 2 is reformulated in terms of a pair of preference orderings, instead of only one. This is consistent with the formulation of our constraints (7) and (8), which are in fact a reinterpretation of Condition 2 in terms of integer programming. Also, our notion of decomposability does not require that R_1 and R_2 contain TR, whereas Kalai and Muller's one requires that R contains TR. In particular, let us stress that our definition requires that R_1 and R_2 satisfy only two conditions—instead of four, as in Kalai and Muller's version. As the next proposition makes it clear, this implies a redundancy of Kalai and Muller's Conditions II and IV. Nevertheless, as anticipated above, the following proposition establishes that the two concepts are equivalent.

Proposition 3 Ω *is KM-decomposable if and only if it is decomposable.*

Proof Let Ω be KM-decomposable. Then, there exists a set R, with $TR \subsetneqq R \subsetneqq \mathcal{A}^2$, which satisfies Conditions I–IV. By Lemma 4 in Kalai and Muller, there exists a set $\bar{R}$, with $TR \subsetneqq \bar{R} \subsetneqq \mathcal{A}^2$, such that, for all $(x, y) \in NTR$, we have $(x, y) \in R$ if and only if $(y, x) \notin \bar{R}$, and which satisfies Conditions I–IV. Let $R_1 = R \setminus TR$ and $R_2 = \bar{R} \setminus TR$. Then, $\emptyset \subsetneqq R_i \subsetneqq NTR$, $i = 1, 2$, and, for all $(x, y) \in NTR$, we have $(x, y) \in R_1$ if and only if $(y, x) \notin R_2$. Consider a triple x, y, z and suppose there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$. Moreover, suppose that $(x, y) \in R_1$ and $(x, z) \notin R_1$. Then, $(x, y) \in R$ and $(x, z) \notin R$ as $(x, z) \in NTR$, contradicting Condition I. Hence, R_i, $i = 1, 2$, satisfies Condition 1. Consider a triple x, y, z and suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Moreover, suppose that $(x, y), (y, z) \in R_1$ and $(x, z) \notin R_1$. Then, $(x, y), (y, z) \in R$, and $(x, z) \notin R$ as $(x, z) \in NTR$, contradicting Condition III. Hence, R_i, $i = 1, 2$, satisfies Condition 2. We have proved that Ω is decomposable. Conversely, suppose that Ω is decomposable. Then, there exist two sets R_1 and R_2, with $\emptyset \subsetneqq R_i \subsetneqq NTR$, $i = 1, 2$, such that, for all $(x, y) \in NTR$, we have $(x, y) \in R_1$ if and only if $(y, x) \notin R_2$; moreover, R_i, $i = 1, 2$, satisfies Conditions 1 and 2. Let $R = R_1 \cup TR$. Consider two pairs $(x, y), (x, z) \in NTR$ and suppose there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and

$y\mathbf{q}z\mathbf{q}x$. Moreover, suppose that $(x, y) \in R$ and $(x, z) \notin R$. Then, $(x, y) \in R_1$ and $(x, z) \notin R_1$ as $(x, y), (x, z) \in NTR$, contradicting Condition 1. Hence, R satisfies Condition I. Now, suppose that $(z, x) \in R$ and $(y, x) \notin R$. Then, $(x, y) \in R_2$ and $(x, z) \notin R_2$ as $(x, y), (x, z) \in NTR$, contradicting Condition 1. Hence, R satisfies Condition II. Consider two pairs $(x, y), (x, z) \in NTR$ and suppose there exists $\mathbf{p} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$. Moreover, suppose that $(x, y), (y, z) \in R$, and $(x, z) \notin R$. There exists $\mathbf{q} \in \Omega$ such that $z\mathbf{q}x$ as $(x, z) \in NTR$. Consider the case where $y\mathbf{q}z\mathbf{q}x$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$, $(x, y) \in R$, and $(x, z) \notin R$, contradicting Condition I. Consider the case where $z\mathbf{q}x\mathbf{q}y$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}x\mathbf{q}y$, $(y, z) \in R$, and $(x, z) \notin R$, contradicting Condition II. Consider the case where $z\mathbf{q}y\mathbf{q}x$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$, $(x, y), (y, z) \in R_1$, and $(x, z) \notin R_1$ as $(x, y), (y, z), (x, z) \in NTR$, contradicting Condition 2. Hence, R satisfies Condition III. Consider two pairs $(x, y), (x, z) \in NTR$ and suppose there exists $\mathbf{p} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$. Moreover, suppose that $(z, x) \in R$ and $(y, x), (z, y) \notin R$. There exists $\mathbf{q} \in \Omega$ such that $z\mathbf{q}x$ as $(x, z) \in NTR$. Consider the case where $z\mathbf{q}x\mathbf{q}y$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}x\mathbf{q}y$, $(z, x) \in R$, and $(z, y) \notin R$, contradicting Condition I. Consider the case where $y\mathbf{q}z\mathbf{q}x$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$, $(z, x) \in R$, and $(y, x) \notin R$, contradicting Condition II. Consider the case where $z\mathbf{q}x\mathbf{q}y$. Then, there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$, $(x, y), (y, z) \in R_2$, and $(x, z) \notin R_2$ as $(x, y), (y, z), (x, z) \in NTR$, contradicting as $(x, y), (y, z), (x, z) \in NTR$, contradicting Condition 2. Hence, R satisfies Condition IV. We have proved that Ω is KM-decomposable. ■

In order to obtain our characterization theorem for ASWFs with ties, we need to restrict further the condition of decomposability introduced above. Then, we introduce a new notion, which we define as "strict decomposability." The next section will be devoted to establish the exact relationship between the two notions of decomposability and strict decomposability.

Then, given a set $R \subset \mathcal{A}^2$, consider the following conditions on R.

Condition 3 There exists a set $R^* \subset \mathcal{A}^2$, with $R \cap R^* = \emptyset$, such that, for all triples x, y, z, if there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$, then $(x, y) \in R^*$ implies that $(x, z) \in R$.

Condition 4 There exists a set $R^* \subset \mathcal{A}^2$, with $R \cap R^* = \emptyset$, such that, for all triples of alternatives x, y, z, if there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$, then $(x, y) \in R$ and $(y, z) \in R^*$ imply that $(x, z) \in R$, and $(x, y) \in R^*$ and $(y, z) \in R$ imply that $(x, z) \in R$.

A domain Ω is said to be strictly decomposable if and only if there exist four sets R_1, R_2, R_1^*, and R_2^*, with $R_i \subsetneq NTR$, $\emptyset \subsetneq R_i^* \subset NTR$, $i = 1, 2$, such that, for all $(x, y) \in NTR$, we have $(x, y) \in R_1$ if and only if $(x, y) \notin R_1^*$ and $(y, x) \notin R_2$; $(x, y) \in R_1^*$ if and only if $(y, x) \in R_2^*$; moreover, $R_i, i = 1, 2$, satisfies Condition 1; R_i and R_i^*, $i = 1, 2$, satisfy Condition 2; each pair (R_i, R_i^*), $i = 1, 2$, satisfies Conditions 3 and 4.

On the basis of the notion of strict decomposability, we provide now the characterization of domains admitting nondictatorial ternary solutions to IP1.

Theorem 4 *There exists a nondictatorial ternary solution to IP2 on* Ω, d, *for* $n = 2$, *if and only if* Ω *is strictly decomposable.*

Proof Let d be a nondictatorial ternary solution to IP2 on Ω, for $n = 2$. Let $R_1 = \{(x, y) \in NTR : d_{\{1\}}(x, y) = 1\}$, $R_2 = \{(x, y) \in NTR : d_{\{2\}}(x, y) = 1\}$, $R_1^* = \{(x, y) \in NTR : d_{\{1\}}(x, y) = \frac{1}{2}\}$, $R_2^* = \{(x, y) \in NTR : d_{\{2\}}(x, y) = \frac{1}{2}\}$. Consider $(x, y) \in NTR$. Suppose that $(x, y) \in R_1$ and $(x, y) \in R_1^*$. Then, $d_{\{1\}}(x, y) = 1$ and $d_{\{1\}}(x, y) = \frac{1}{2}$, a contradiction. Suppose that $(x, y) \in R_1$ and $(y, x) \in R_2$. Then, $d_{\{1\}}(x, y) = 1$ and $d_{\{2\}}(y, x) = 1$, contradicting (2). Suppose that $(x, y) \notin R_1^*$, $(y, x) \notin R_2$, and $(x, y) \notin R_1$. Then, $d_{\{1\}}(x, y) \neq \frac{1}{2}$, $d_{\{1\}}(x, y) \neq 0$, and $d_{\{1\}}(x, y) \neq 1$, a contradiction. Suppose that $(x, y) \in R_1^*$ and $(y, x) \notin R_2^*$. Then, $d_{\{1\}}(x, y) = \frac{1}{2}$ and $d_{\{2\}}(y, x) \neq \frac{1}{2}$, contradicting (2). Hence, for all $(x, y) \in NTR$, $(x, y) \in R_1$ if and only if $(x, y) \notin R_1^*$ and $(y, x) \notin R_2$; $(x, y) \in R_1^*$ if and only if $(y, x) \in R_2^*$. Suppose that $R_1 = NTR$. Then, d is dictatorial, a contradiction. Hence, $R_i \subsetneqq NTR$, $i = 1, 2$. Suppose that $R_i^* = \emptyset$, $i = 1, 2$. Then, d is a binary solution, a contradiction. Hence, $\emptyset \subsetneqq R_i^* \subset NTR$. Consider a triple x, y, z and suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$. Moreover, suppose that $(x, y) \in R_1$ and $(x, z) \notin R_1$. Then, $d_{\{1\}}(x, y) = 1$ and

$$d_{\{1\}}(x, y) > d_{\{1\}}(x, z),$$

contradicting (5). Hence, R_i, $i = 1, 2$, satisfies Condition 1. Consider a triple x, y, z and suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Moreover, suppose that $(x, y), (y, z) \in R_1$, and $(x, z) \notin R_1$. Then, $d_{\{1\}}(x, y) = 1$, $d_{\{1\}}(y, z) = 1$, and

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) > 1 + d_{\{1\}}(x, z),$$

contradicting (7). Hence, $R_i, i = 1, 2$, satisfies Condition 2. Consider a triple x, y, z and suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Moreover, suppose that $(x, y) \in R_1^*$, $(y, z) \in R_1^*$, and $(x, z) \notin R_1^*$. Then, $d_{\{1\}}(x, y) = \frac{1}{2}$, $d_{\{1\}}(y, z) = \frac{1}{2}$, and

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) \neq \frac{1}{2} + d_{\{1\}}(x, z),$$

contradicting (8). Hence, R_i^* satisfies Condition 2, $i = 1, 2$. Consider a triple x, y, z and suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$. Moreover, suppose that $(x, y) \in R_1^*$ and $(x, z) \notin R_1$. Then, $d_{\{1\}}(x, y) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) \geq d_{\{1\}}(x, z),$$

contradicting (6). Hence, each pair (R_i, R_i^*), $i = 1, 2$, satisfies Condition 3. Consider a triple x, y, z and suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Moreover, suppose that $(x, y) \in R_1$, $(y, z) \in R_1^*$, and $(x, z) \notin R_1$. Then, $d_{\{1\}}(y, z) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) \neq \frac{1}{2} + d_{\{1\}}(x, z),$$

contradicting (8). Now, suppose that $(x, y) \in R_1^*$, $(y, z) \in R_1$, and $(x, z) \notin R_1$. Then, $d_{\{1\}}(x, y) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) \neq \frac{1}{2} + d_{\{1\}}(x, z),$$

contradicting (8). Hence, each pair (R_i, R_i^*), $i = 1, 2$, satisfies Condition 4. We have proved that Ω is strictly decomposable. Conversely, suppose that Ω is strictly decomposable. Then, there exist four sets R_1, R_2, R_1^*, and R_2^*, with $R_i \subsetneqq NTR$, $\emptyset \subsetneqq R_i^* \subset NTR$, $i = 1, 2$, such that, for all $(x, y) \in NTR$, we have $(x, y) \in R_1$ if and only if $(x, y) \notin R_1^*$ and $(y, x) \notin R_2$; $(x, y) \in R_1^*$ if and only if $(y, x) \in R_2^*$; moreover, R_i, $i = 1, 2$, satisfies Condition 1; R_i and R_i^*, $i = 1, 2$, satisfy Condition 2; each pair (R_i, R_i^*), $i = 1, 2$, satisfies Conditions 3 and 4. Determine d as follows. For each $(x, y) \in NTR$, let $d_\emptyset(x, y) = 0$, $d_E(x, y) = 1$; $d_{\{i\}}(x, y) = 1$ if and only if $(x, y) \in R_i$; $d_{\{i\}}(x, y) = \frac{1}{2}$ if and only if $(x, y) \in R_i^*$; $d_{\{i\}}(x, y) = 0$ if and only if, $(x, y) \notin R_i$ and $(x, y) \notin R_i^*$, for $i = 1, 2$. Then, d satisfies (1) and (2) as, for all $(x, y) \in NTR$, $(x, y) \in R_1$ if and only if $(x, y) \notin R_1^*$ and $(y, x) \notin R_2$, $(x, y) \in R_1^*$ if and only if $(y, x) \in R_2^*$. Consider a triple x, y, z and suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$. Moreover, suppose that

$$d_{\{1\}}(x, y) > d_{\{1\}}(x, z).$$

Then, we have $(x, y) \in R_1$ and $(x, z) \notin R_1$, contradicting Condition 1. Therefore, d satisfies (5). Consider a triple x, y, z and suppose that there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Moreover, suppose that

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) > 1 + d_{\{1\}}(x, z).$$

Then, we have $(x, y), (y, z) \in R_1$ and $(x, z) \notin R_1$, contradicting Condition 2. Therefore, d satisfies (7). Consider a triple x, y, z and suppose there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$. Moreover, suppose that $d_{\{1\}}(x, y) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) \geq d_{\{1\}}(x, z).$$

Then, $(x, y) \in R_1^*$ and $(x, z) \notin R_1$, contradicting Condition 3. Therefore, d satisfies (6). Consider a triple x, y, z and suppose there exist $\mathbf{p}, \mathbf{q} \in \Omega$ satisfying

$x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$. Moreover, suppose that $d_{\{1\}}(x, y) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) > \frac{1}{2} + d_{\{1\}}(x, z).$$

Consider the following two cases. First, $d_{\{1\}}(y, z) = 1$. Then, $(x, y) \in R_1^*$, $(y, z) \in R_1$, and $(x, z) \notin R_1$, contradicting Condition 4. Second, $d_{\{1\}}(y, z) = \frac{1}{2}$. Then, $(x, y) \in R_1^*$, $(y, z) \in R_1^*$, and $(x, z) \notin R_1^*$, contradicting Condition 2. Finally, suppose that $d_{\{1\}}(x, y) = \frac{1}{2}$ and

$$d_{\{1\}}(x, y) + d_{\{1\}}(y, z) < \frac{1}{2} + d_{\{1\}}(x, z).$$

Consider the following two cases. First, $d_{\{1\}}(y, z) = 0$. Then, $(z, y) \in R_2$, $(y, x) \in R_2^*$, and $(z, x) \notin R_2$, contradicting Condition 4. Second, $d_{\{1\}}(y, z) = \frac{1}{2}$. Then, $(x, y) \in R_1^*$, $(y, z) \in R_1^*$, and $(x, z) \notin R_1^*$, contradicting Condition 2. Therefore, d satisfies (8). d is nondictatorial as $\emptyset \subsetneqq R_i^* \subset NTR$, $i = 1, 2$. Hence, d is a nondictatorial ternary solution to IP2 on Ω. ■

Our characterization theorem for ASWFs with ties immediately follows from Theorems 1 and 3. This result is a generalization of Kalai and Muller's Theorem 2 for ASWFs without ties.

Theorem 5 *There exists a nondictatorial ASWF with ties on Ω, f, for $n \geq 2$, if and only if Ω is strictly decomposable.*

Proof It is a straightforward consequence of Propositions 1 and 2, Theorems 1, 3, and 4. ■

5 The Relationship Between Decomposable and Strictly Decomposable Domains

In this section, we analyze the relationship between the notions of decomposable and strictly decomposable domain. The following example illustrates the two notions.

Example 1 Let $A = \{a, b, c, d\}$ and $\Omega = \{\mathbf{p} \in \Sigma : a\mathbf{p}b\mathbf{p}c\mathbf{p}d, c\mathbf{p}d\mathbf{p}a\mathbf{p}b, d\mathbf{p}c\mathbf{p}b\mathbf{p}a\}$. Then, Ω is decomposable and strictly decomposable.

Proof The triples x, y, z for which there exist $\mathbf{p}, \mathbf{q} \in \Omega$ such that $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$ are c,a,b; d,a,b; a,c,d; b,c,d. The triples x, y, z for which there exist $\mathbf{p}, \mathbf{q} \in \Omega$ such that $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$ are a,b,c; a,b,d; a,c,d; b,c,d. Let $R_1 = \{(a, b), (b, a), (c, d), (d, c)\}$ and $R_2 = \{(a, c), (c, a), (a, d), (d, a), (b, c), (c, b), (b, d), (d, b)\}$. Then, we have $\emptyset \subsetneqq R_i \subsetneqq NTR$, $i = 1, 2$. Moreover, for all $(x, y) \in NTR$, we have $(x, y) \in R_1$ if and only if $(y, x) \notin R_2$. R_1

vacuously satisfies Conditions 1 and 2. R_2 satisfies Condition 1 as we have: $(a,c) \in R_2$ and $(a,d) \in R_2$; $(c,a) \in R_2$ and $(c,b) \in R_2$; $(d,a) \in R_2$ and $(d,b) \in R_2$; $(b,c) \in R_2$ and $(b,d) \in R_2$. R_2 vacuously satisfies Condition 2. We have shown that Ω is decomposable. Now, let $V_1 = \{(a,b),(c,d)\}$, $V_2 = \{(a,c),(c,a),(a,d),(d,a),(b,c),(c,b),(b,d),(d,b)\}$, $V_1^* = \{(b,a),(d,c)\}$, $V_2^* = \{(a,b),(c,d)\}$. Then, we have $V_i \subsetneqq NTR$, $i = 1,2$, and $\emptyset \subsetneqq V_i^* \subset NTR$, $i = 1,2$. Moreover, for all $(x,y) \in NTR$, we have: $(x,y) \in V_1$ if and only if $(x,y) \notin V_1^*$ and $(y,x) \notin V_2$; $(x,y) \in V_1^*$ if and only if $(y,x) \in V_2^*$. V_1 vacuously satisfies Conditions 1 and 2. V_1^* vacuously satisfies Condition 2. Moreover, the pair (V_1, V_1^*) vacuously satisfies Conditions 3 and 4. V_2 satisfies Conditions 1 and 2 as $V_2 = R_2$. V_2^* vacuously satisfies Condition 2. The pair (V_2, V_2^*) vacuously satisfies Condition 3. Moreover, it satisfies Condition 4 as we have: $(a,c) \in V_2$, $(c,d) \in V_2^*$, and $(a,d) \in V_2$; $(b,c) \in V_2$, $(c,d) \in V_2^*$, and $(b,d) \in V_2$; $(a,b) \in V_2^*$, $(b,c) \in V_2$, and $(a,c) \in V_2$; $(a,b) \in V_2^*$, $(b,d) \in V_2$, and $(a,d) \in V_2$. We have shown that Ω is strictly decomposable. ■

The example above specifies a domain which is both decomposable and strictly decomposable. Nonetheless, this is not the general case. In the following, we will show, with a theorem and a further example, that a strictly decomposable domain is always decomposable, but the converse is not true.

In order to obtain these results, we preliminarily show the following theorem on the nondictatorial solutions to IP2.

Theorem 6 *If there exists a nondictatorial ternary solution to IP2 on Ω, d, for $n = 2$, then there exists a nondictatorial binary solution to IP2 on Ω, $\hat{d}$, for $n = 2$.*

Proof Let d be a ternary solution to IP2 on Ω, for $n = 2$. Determine d' as follows. Consider $\mathbf{q} \in \Sigma$. For each $(x,y) \in NTR$, let: $d'_{\emptyset}(x,y) = 0$, $d'_E(x,y)=1$; $d'_{\{i\}}(x,y) = d_{\{i\}}(x,y)$, if $d_{\{i\}}(x,y) \in \{0,1\}$, $i = 1,2$; $d'_{\{1\}}(x,y) = 1$ and $d'_{\{2\}}(y,x) = 0$, if $d_{\{1\}}(x,y) = d_{\{2\}}(y,x) = \frac{1}{2}$ and $x\mathbf{q}y$. Then, it is immediate to verify that d' is a solution to IP2 on Ω, for $n = 2$. Suppose that d' is nondictatorial. Then, $\hat{d} = d'$ is a nondictatorial binary solution to IP2 on Ω, for $n = 2$. Suppose that d' is dictatorial: say, for example, that, for all $(x,y) \in NTR$, $d_S(x,y) = 1$, for all S containing agent 1. In this case, we can say that agent 1 is the dictator for d'. Determine d'' as follows. Let $\mathbf{q}^{-1} \in \Sigma$ be an antisymmetric preference ordering such that, for each $(x,y) \in \mathcal{A}^2$, $x\mathbf{q}y$ if and only if $y\mathbf{q}^{-1}x$. For each $(x,y) \in NTR$, let: $d''_{\emptyset}(x,y) = 0$, $d''_E(x,y)=1$; $d''_{\{i\}}(x,y) = d_{\{i\}}(x,y)$, if $d_{\{i\}}(x,y) \in \{0,1\}$, $i = 1,2$; $d''_{\{1\}}(x,y) = 1$ and $d''_{\{2\}}(y,x) = 0$, if $d_{\{1\}}(x,y) = d_{\{2\}}(y,x) = \frac{1}{2}$ and $x\mathbf{q}^{-1}y$. Then, it is immediate to verify that $\hat{d} = d''$ is a binary solution to IP2 on Ω, for $n = 2$, and that agent 1 is not a dictator for d''. Suppose that agent 2 is a dictator for d''. Consider $(x,y) \in NTR$ such that $d_{\{1\}}(x,y) = d_{\{2\}}(y,x) = \frac{1}{2}$. Suppose that $y\mathbf{q}x$. This implies that $d'_{\{1\}}(x,y) = 0$ and agent 1 is not a dictator for d', a contradiction. But then, we must have that $x\mathbf{q}y$. Consider variables $d_{\{1\}}(y,x)$ and $d_{\{2\}}(x,y)$. Suppose that $d_{\{1\}}(y,x) = 1$ and $d_{\{2\}}(x,y) = 0$. Then, agent 2 is not a dictator for d'', a contradiction. Suppose that $d_{\{1\}}(y,x) = 0$ and $d_{\{2\}}(x,y) = 1$.

Then, agent 1 is not a dictator for d'. This implies that $d_{\{1\}}(y,x) = d_{\{2\}}(x,y) = \frac{1}{2}$ and this, in turn, implies that $d''_{\{2\}}(x,y) = 0$ and agent 2 is not a dictator of d'', a contradiction. Then, $\hat{d} = d''$ is a nondictatorial binary solution to IP2 on Ω, for $n = 2$. ■

Then, the following theorem can be immediately proved.

Theorem 7 *If a domain Ω is strictly decomposable, then it is decomposable.*

Proof Let Ω be a strictly decomposable domain. Then, by Theorem 4, there exists a nondictatorial ternary solution to IP2 on Ω, d, for $n = 2$. But then, by Theorem 6, there exists a nondictatorial binary solution to IP2 on Ω, $\hat{d}$, for $n = 2$. Hence, by Theorems 1 and 2, and Proposition 3, Ω is decomposable. ■

The following example shows that the converse of Theorem 7 does not hold.

Example 2 Let $A = \{a,b,c,d\}$ and $\Omega = \{\mathbf{p} \in \Sigma : a\mathbf{p}b\mathbf{p}c\mathbf{p}d, c\mathbf{p}a\mathbf{p}d\mathbf{p}b, d\mathbf{p}c\mathbf{p}b\mathbf{p}a, b\mathbf{p}d\mathbf{p}a\mathbf{p}c\}$. Then, Ω is decomposable but it is not strictly decomposable.

Proof The triples x, y, z for which there exist $\mathbf{p}, \mathbf{q} \in \Omega$ such that $x\mathbf{p}y\mathbf{p}z$ and $y\mathbf{q}z\mathbf{q}x$ are: c,a,b; c,b,a; a,b,d; a,d,b; d,a,c; d,c,a; b,c,d; b,d,c. The triples x, y, z for which there exist $\mathbf{p}, \mathbf{q} \in \Omega$ such that $x\mathbf{p}y\mathbf{p}z$ and $z\mathbf{q}y\mathbf{q}x$ are: a,b,c; c,a,b; a,b,d; a,d,b; a,c,d; c,a,d; b,c,d; c,d,b. Let $R_i = \{(a,b),(a,c),(a,d),(b,c),(b,d),(c,d)\}$, $i = 1,2$. Then, we have $\emptyset \subsetneqq R_i \subsetneqq NTR$, $i = 1,2$. Moreover, for all $(x,y) \in NTR$, we have $(x,y) \in R_1$ if and only if $(y,x) \notin R_2$. R_i, $i = 1,2$, satisfies Condition 1 as we have: $(a,b) \in R_i$ and $(a,d) \in R_i$; $(a,d) \in R_i$ and $(a,b) \in R_i$; $(b,c) \in R_i$ and $(b,d) \in R_i$; $(b,d) \in R_i$ and $(b,c) \in R_i$, $i = 1,2$. R_i, $i = 1,2$, satisfies Condition 2 as we have: $(a,b) \in R_i$, $(b,c) \in R_i$, and $(a,c) \in R_i$; $(a,b) \in R_i$, $(b,d) \in R_i$, and $(a,d) \in R_i$; $(a,c) \in R_i$, $(c,d) \in R_i$, and $(a,d) \in R_i$; $(b,c) \in R_i$, $(c,d) \in R_i$, and $(b,d) \in R_i$, $i = 1,2$. We have shown that Ω is decomposable. Now suppose that Ω is strictly decomposable. Then, there exist four sets V_1, V_2, V_1^*, and V_2^*, with $V_i \subsetneqq NTR$, $\emptyset \subsetneqq V_i^* \subset NTR$, $i = 1,2$, such that, for all $(x,y) \in NTR$, we have: $(x,y) \in V_1$ if and only if $(x,y) \notin V_1^*$ and $(y,x) \notin V_2$; $(x,y) \in V_1^*$ if and only if $(y,x) \in V_2^*$. Moreover, V_i, $i = 1,2$, satisfies Condition 1; V_i and V_i^*, $i = 1,2$, satisfy Condition 2; each pair (V_i, V_i^*), $i = 1,2$, satisfies Conditions 3 and 4. Suppose that $(a,b) \in V_1^*$ and $(b,a) \in V_2^*$. Then, $(a,d) \in V_1$ as the pair (V_1, V_1^*) satisfies Condition 3. But then, $(a,b) \in V_1$ as V_1 satisfies Condition 1, a contradiction. Suppose that $(a,c) \in V_1^*$ and $(c,a) \in V_2^*$. Then, $(c,b) \in V_2$ as the pair (V_2, V_2^*) satisfies Condition 3. But then, $(c,a) \in V_2$ as V_2 satisfies Condition 1, a contradiction. Suppose that $(a,d) \in V_1^*$ and $(d,a) \in V_2^*$. Then, $(a,b) \in V_1$ as the pair (V_1, V_1^*) satisfies Condition 3. But then, $(a,d) \in V_1$ as V_1 satisfies Condition 1, a contradiction. Suppose that $(b,c) \in V_1^*$ and $(c,b) \in V_2^*$. Then, $(b,d) \in V_1$ as the pair (V_1, V_1^*) satisfies Condition 3. But then, $(b,c) \in V_1$ as V_1 satisfies Condition 1, a contradiction. Suppose that $(b,d) \in V_1^*$ and $(d,b) \in V_2^*$. Then, $(b,c) \in V_1$ as the pair (V_1, V_1^*) satisfies Condition 3. But then, $(b,d) \in V_1$ as V_1 satisfies Condition 1, a contradiction. Suppose that $(c,d) \in V_1^*$ and $(d,c) \in V_2^*$. Then, $(d,a) \in V_2$ as the pair (V_2, V_2^*) satisfies Condition 3. But then, $(d,c) \in V_2$ as V_2 satisfies Condition 1,

a contradiction. Hence, $V_i^* = \emptyset, i = 1, 2$, a contradiction. We have shown that Ω is not strictly decomposable. ■

Acknowledgements This paper has been written to honor Nick Baigent and his distinguished contributions to social choice theory. We would like to thank an anonymous referee for his comments and suggestions. Francesca Busetto and Giulio Codognato gratefully acknowledge financial support from MIUR (PRIN 20103S5RN3).

References

1. Arrow KJ (1963) Social choice and individual values. Wiley, New York
2. Busetto F, Codognato G (2010) Nondictatorial Arrovian social welfare functions and integer programs. Working Paper n. 01-10-eco, Dipartimento di Scienze Economiche, Università degli Studi di Udine
3. Kalai E, Muller E (1977) Characterization of domains admitting nondictatorial social welfare functions and nonmanipulable voting procedures. J Econ Theory 16:457–469
4. Maskin E (1979) Fonctions de préférence collective définies sur des domaines de préférence individuelle soumis à des constraintes. Cahiers du Séminaire d'Econométrie 20:153–182
5. Sethuraman J, Teo CP, Vohra RV (2003) Integer programming and Arrovian social welfare functions. Math Oper Res 28:309–326
6. Sethuraman J, Teo CP, Vohra RV (2006) Anonymous monotonic social welfare functions. J Econ Theory 128:232–254

Distance Rationalizability of Scoring Rules

Burak Can

Abstract Collective decision making problems can be seen as finding an outcome that is "closest" to a concept of "consensus". Nitzan (1981) introduced "Closeness to Unanimity Procedure" as a first example to this approach and showed that the Borda rule is the closest to unanimity under the Kemeny (1959) distance. Elkind et al. (2009) generalized this concept as distance-rationalizability, and showed that all scoring rules can be distance rationalized via a class of distance functions, which we call scoring distances. In this paper, we propose another class of distances, i.e., weighted distances, introduced in Can (2014). This class is a generalization of the Kemeny distance that rationalizes the generalization of the Borda rule, i.e., scoring rules. Hence the results here extend those in Nitzan (1981) and reveal the broader connection between Kemeny-like distances and Borda-like voting rules.

Keywords Distance rationalizability • Scoring rules • Voting • Weighted distances

1 Introduction

Nitzan [9] introduced the *closeness to unanimity procedures* (CUPs) for collective decision making problems. Given a distance function as a measure of closeness over preference profiles, these procedures find "closest" unanimous preference profiles to the original preference profile at hand. This approach, in a sense, yields the outcome which requires the minimal total compromise towards a unanimous agreement from a utilitarian perspective.

Meskanen and Nurmi [8] use other consensus concepts such as the existence of a Condorcet winner in a profile. Then, the compromise needed is not to achieve a unanimous profile, but to achieve a profile in which a Condorcet winner exists. They show that if the consensus concept is not unanimity, but a Condorcet winner

B. Can (✉)
Department of Economics, School of Business and Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands
e-mail: b.can@maastrichtuniversity.nl

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*, Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_11

instead, then the Dodgson winner in a profile is the closest to being a Condorcet winner under a compromise defined as the Kemeny (swap) distance.

Elkind et al. [4] generalize the notion of closeness to various concepts of consensus as *distance-rationalization*.[1] They use many reasonable consensus classes apart from unanimity, and employ different distance functions to shed light on the existing voting rules and their relation to distance functions within a consensus approach.

Nitzan [9] showed that the simplest scoring rule, i.e. the Borda rule, is equivalent to *closeness to unanimity procedure* under the Kemeny distance. This means that the Borda rule is somewhat rationalized by the Kemeny distance. Elkind et al. [4] extend this result and show that non-degenerate[2] scoring rules are rationalized by a class which we shall call *scoring-distances*. They also show that degenerate scoring rules, e.g., scoring rules which may have equal scores for different positions in a ranking, can be rationalized by pseudo-distances.[3]

In this paper, we show that the non-degenerate scoring rules can also be rationalized by another class of distance functions introduced in [2], i.e., *weighted distances*. There it is shown that weighted distances are generalizations of the Kemeny distance. Hence, the connection between the "Borda" rule and the "Kemeny" distance revealed in [9], can be extended to the connection between the "scoring rules" and the "weighted distances". The main difference between weighted distances and scoring distances in [4, 5], is that the former class satisfy a condition called *decomposability*. This condition is a weakening of one of the Kemeny distance axioms, i.e., *betweenness*. Hence the rationalizability of the Borda rule (with the Kemeny distance) is naturally extended to rationalizability of scoring rules (with the weighted Kemeny distances). The results also extend to distance rationalization of degenerate scoring rules by weighted pseudo-distances.

2 Model

2.1 Preliminaries

Let N be a finite set of agents with cardinality n, and A be a finite set of alternatives with cardinality m. The set of all possible strict preferences, i.e., complete, transitive and antisymmetric binary relations over A, is denoted by $\mathscr{L}$. A generic preference is denoted by $R \in \mathscr{L}$ whereas the set of strict preferences with an alternative a at the top is denoted by $\mathscr{L}^a$. A preference profile is an n-tuple vector of preferences

[1] For a broad analysis of the connection between distance functions and voting rules see [5], for distance rationalizability of Condorcet-consistent voting rules see [6].

[2] Non-degenerate scoring rules are scoring rules that assign decreasing scores to the positions in a ranking, therefore these rules do not include plurality, k-approval rule etc.

[3] A pseudo-distance is a function which satisfies all metric conditions except identity of indiscernibles.

denoted by $p = (p(1), p(2), \ldots, p(n)) \in \mathscr{L}^N$. Given an alternative $a \in A$, we denote the profiles with a as the top alternative in each individual preference by p^a.

For $l = 1, 2, \ldots, m$, $R(l)$ denotes the alternative in the lth position in R, e.g., $R(1)$ denotes the top alternative. Given an alternative a and a preference R, we denote the position of a in R by $a(R)$, i.e., $a(R) = x$ if and only if $a = R(x)$. To denote the position of alternative a in the preference of ith individual in a profile, we abuse notation and write $a(i)$ instead of $a(p(i))$, as long as it is clear which preference profile we refer to. Two linear orders $(R, R') \in \mathscr{L}^2$ form an *elementary change*[4] in position k whenever $R(k) = R'(k+1)$, $R'(k) = R(k+1)$ and for all $t \notin \{k, k+1\}$, $R(t) = R'(t)$, i.e. $|R \setminus R'| = 1$. Given any two distinct linear orders $R, R' \in \mathscr{L}$, a vector of linear orders $\rho = (R_0, R_1, \ldots, R_k)$ is called a *path* between R and R' if $k = |R \setminus R'|$, $R_0 = R$, $R_k = R'$ and for all $i = 1, 2, \ldots k$, (R_{i-1}, R_i) forms an elementary change. For the special case where $R = R'$, we denote the unique path as $\rho = (R, R)$.

A vector $s = (s_1, s_2, \ldots, s_m)$ over positions of alternatives in a preference is called a *scoring vector* whenever $s_1 \geq s_2 \geq \ldots s_m \geq 0$. A scoring vector s is called *non-degenerate* if scores are strictly decreasing from s_1 to s_m, i.e., $s_1 > s_2 > \ldots s_m \geq 0$. The score of an alternative a in a preference R is denoted by $score(a, R)$ and is equal to $s_{a(R)}$ in the scoring vector.

A *collective choice rule*, or a voting rule, is a correspondence $\alpha : \mathscr{L}^N \to 2^A \setminus \emptyset$, which assigns each preference profile a nonempty subset of alternatives. Given a preference profile $p \in \mathscr{L}^N$, a *scoring rule*, denoted by α_s, with scoring vector s is a choice rule that assigns a summed score to each alternative in A, $\sum_{i \in N} score(a, p(i))$, and assigns to each profile the alternatives with maximal total scores,

$$\alpha_s(p) = \max_{a \in A} \sum_{i \in N} score(a, p(i))$$

Example 1 Let $s = (m-1, m-2, \ldots, 0)$, then *the Borda rule* on each preference profile is defined as:

$$\alpha_{Borda}(p) = \arg\max_{a \in A} \sum_{i \in N} score(a, p(i)) = \arg\max_{a \in A} \sum_{i \in N} (m - a(i)))$$

Let us now dwell upon the concepts of "closeness" between individual preferences and thereafter preference profiles. Let a function $\delta : \mathscr{L} \times \mathscr{L} \to \mathbf{R}$ assign a real number to each pair of preferences. A function over preferences is a *distance function* if it satisfies:

(i) Non-negativity: $\delta(R, R') \geq 0$ for all $R, R' \in \mathscr{L}$,
(ii) Identity of indiscernibles: $\delta(R, R') = 0$ if and only if $R = R'$ for all R, $R' \in \mathscr{L}$,

[4]We omit the parenthesis whenever it is clear and write R, R' instead.

(iii) Symmetry: $\delta(R, R') = \delta(R', R)$ for all $R, R' \in \mathscr{L}$.
(iv) Triangular inequality: $\delta(R, R'') \leq \delta(R, R') + \delta(R', R'')$ for all $R, R', R'' \in \mathscr{L}$.

Two well-known examples of distance functions are *the discrete distance*, and *the swap distance*. The former assigns 0 if the two preferences are identical, and 1 otherwise. The latter function was introduced by Kemeny [7] and re-characterized with logically independent conditions in [3]. The Kemeny distance counts the symmetric total number of disjoint ordered pairs in preferences, or simply the minimal number of "swaps of adjacent alternatives" required to transform one preference into another.[5] Elkind et al. [4] also refer to functions that satisfy i, iii, and iv. These functions, which lack the identity of indiscernibles condition, are called *pseudo-distance functions*. These functions may assign 0 to distances between distinct pair of rankings, e.g., $\delta(abc, cab) = 0$.

For distance rationalizability we will mainly refer to distance functions between preference profiles. Given a distance function δ over preferences, a straightforward extension of δ over preference profiles, say $p, p' \in \mathscr{L}^N$, can be defined as a function $d : \mathscr{L}^N \times \mathscr{L}^N \to \mathbf{R}$ as follows:

$$d(p, p') = \sum_{i \in N} \delta(p(i), p'(i)).$$

Note that this is a very straightforward and common extension of distances over individual preferences to distances over preference profiles, e.g., see [1]. We abuse notation for the sake of simplicity by referring to δ instead of d as long as it is clear.

2.2 Distance Rationalizability

We only consider "unanimity" as a *consensus class*. The definitions below are adapted smoothly to our notation for simplicity. For a more general notation that would be applicable to many other consensus classes, we refer the reader to [4, 6].

Definition 1 ((U,δ)-Score) The unanimity-score of an alternative a in a preference profile p under the distance function δ is the minimal distance between the profile p and any profile p^a where a is unanimity winner. Formally:

$$(U, \delta){-}score(a, p) = \min_{p^a \in \mathscr{L}^N} \delta(p, p^a).$$

Roughly speaking, $(U, \delta) - score$ of an alternative in a profile tells us how costly it is to make this alternative the best alternative in each individual preference, i.e.,

[5] In the literature, the swap distance and the Kemeny distance are interchangeably used. Kemeny [7] originally assumes the distance for each swap in a ranking to be 2, whereas in many works, for convenience, this is normalized to 1. This occurs especially when the domain of preferences is strict and there is no indifference.

the unanimity winner. Obviously there are many possible preference profiles, p^a, where the alternative a is the unanimity winner. The aforementioned score assigns the total cost to convert the original profile to one of such profiles for which the total cost is minimal. Next we reproduce the definition of distance rationalizability. We adapt again from [6] to simplify our notation.

Definition 2 A collective choice rule α, is *distance-rationalizable* via unanimity and a distance function δ, or simply *(U,δ)-rationalizable*, if for all profiles $p \in \mathscr{L}^N$, we have:

$$\alpha(p) = \arg\min_{a \in A} [(U, \delta)\text{–}score(a, p)]$$

To state verbally, a rule is (U,δ)-rationalizable if all outcomes the rule assigns to each profile are also the alternatives which have the minimal (U,δ)-scores for that profile, i.e., the least costly to make the unanimity winner with that distance function.

2.3 Weighted Distances

Can [2] introduced weighted distances as an extension of the Kemeny distance on strict rankings, which would allow for differential treatment of the position of elementary changes. For instance consider, $R = abc$, $R' = acb$, and $\bar{R} = bac$. The Kemeny distance between R and R' is 1 as well as the Kemeny distance between R and $\bar{R}$. However one might argue that the former two are less dissimilar than the latter two, i.e., $\delta_\omega(R, R') < \delta_\omega(R, \bar{R})$, because a swap at the top of rankings may be more critical than a swap at the bottom of thereof.

A weighted distance assigns weights to positions of such swaps with a weight vector on all possible swaps, e.g., $\omega = (\omega_1, \omega_2, \ldots, \omega_{m-1})$. For any two rankings that require more than a single swap, one would find the summation of sequential swaps on a shortest path between the two rankings (see Example 2 below for multiple paths). Hence a path between the two rankings is decomposed into elementary changes, and each elementary change is assigned its corresponding weight according to the weight vector.

Example 2 An example of the two possible shortest paths between $R = abc$ and $R' = cba$ would then be $\rho_1 = [abc, bac, bca, cba]$ and $\rho_2 = [abc, acb, cab, cba]$

For a technical description of the weighted distances, we refer the reader to [2]. Note that in the case of distance rationalizability, the complication regarding multiple paths between rankings do not occur. Hence, it is sufficient to illustrate a weighted distance with an example below:

Example 3 Let $R = abcd$, and $R' = dabc$. Consider the weight vector $\omega = (10, 3, 1)$ and a weighted distance δ_ω, i.e., a swap of alternatives at top creates a

distance of 10, at the middle a distance of 3, and at the bottom a distance of 1. Then:

$$\delta(R, R') = \delta(ab\mathbf{cd}, ab\mathbf{dc}) + \delta(a\mathbf{bd}c, a\mathbf{db}c) + \delta(\mathbf{ad}bc, \mathbf{da}bc)$$
$$\delta(R, R') = \omega_3 + \omega_2 + \omega_1 = 10 + 3 + 1 = 14.$$

3 Results

Nitzan [9] proved that the plurality rule is $(U, \delta_{discrete})$-*rationalizable* and that the Borda rule is (U, δ_{Kemeny})-*rationalizable*. In this paper we extend the Borda result to all scoring rules via a new class of distance functions introduced in [2]. We show that any non-degenerate[6] scoring rule is (U, δ_ω)-rationalizable where δ_ω is a weighted distance with particular non-zero weights. For degenerate scoring rules, the rationalization still holds but with weighted pseudo-distances which allows for zero weights.

The class of weighted distance functions in [2] are characterized by two conditions on top of the usual metric conditions: *positional neutrality* and *decomposability*. Both conditions[7] are in fact weakening of characterizing axioms of the Kemeny distance, which allow for differential treatment of positions in a ranking. Therefore to allow for scoring rules other than the Borda rule, some weakening on the conditions on the distance functions is necessary. The results herein, therefore extend the existing interconnectedness (of the Borda rule and the Kemeny distance) to that of "all scoring rules" and "weighted distances". Weighted distances are Kemeny-like metrics which assign weights on the position of the swaps required to convert one (strict) ranking to another. In that respect, the Kemeny distance is also a weighted distance where weights on all possible swaps, regardless of their positions, are identical. The scoring distances introduced in [4], however, are not decomposable[8] hence they do not follow a Kemeny-like pattern.

Let α_s be a scoring choice rule with the scoring vector $s = (s_1, s_2, \ldots, s_m)$. Then consider a weighted distance δ_ω with the weight vector $\omega = \Delta s = (s_1 - s_2, s_2 - s_3, \ldots, s_{m-1} - s_m)$, i.e., the weight assigned to each swap is the difference between the scores of the relevant consecutive positions. In the following theorem we explain the connection with the class of weighted distance functions and the distance rationalizability of non-degenerate scoring rules.

[6] By non-degenerate scoring rule we mean a non-degenerate scoring vector wherein $s_i > s_{i+1}$ for all $i = 1, 2, \ldots, m$.

[7] Positional neutrality is simply equal treatments of swaps of adjacent alternatives on same positions whereas decomposability requires additive summation of distances on at least one path as in Example 2.

[8] For instance consider the Borda score vector $s = (2, 1, 0)$. According to the scoring distance, the distance between $R = abc$ and $R' = cba$ would be 4, i.e.,$\delta_{scoring}(R, R') = |s_1 - s_3| + |s_2 - s_2| + |s_3 - s_1| = 2 + 0 + 2$. However when you consider the two paths between R and R' in Example 2, it is easy to see that the summation on each of the paths should add up to 6.

Theorem 1 *A non-degenerate scoring rule α_s is (U, δ)-rationalizable if $\delta = \delta_\omega$ is a weighted distance with $\omega = \Delta s$.*

Proof Let $\delta = \delta_\omega$ be a weighted distance function with a weight vector $\omega = \Delta s = (s_i - s_{i+1})_{i=1}^{m-1}$. We want to show that α_s is (U, δ_ω)-rationalizable which means for all profiles $p \in \mathscr{L}^n$, and for all alternatives $a \in A$, we have $a \in \alpha_s(p)$ if and only if (U, δ_ω)-score of a is minimal for all $a \in A$. Take any $p \in \mathscr{L}^n$ and any $a \in A$. Now for each $i \in N$, let $\bar{p}^a(i) \in \mathscr{L}^a$ be such that $\bar{p}^a(i)$ is identical to $p(i)$ except that alternative a is taken to the top while everything else remains the same. By triangular inequality of δ_ω, note that $\bar{p}^a(i) = \arg\min_{p^a \in \mathscr{L}^a} \delta_\omega(p(i), p^a)$, i.e., $\bar{p}^a(i)$ is the closest to $p(i)$ among all other preferences which have a at the top. This is simply because when constructing $\bar{p}^a(i)$, we leave everything unchanged except bringing a to the top. Hence, for the constructed preference profile $\bar{p}^a \in \mathscr{L}^N$, the alternative a is the unanimity winner and furthermore $\bar{p}^a$ is the closest to the original profile p among all other profiles $p^a \in \mathscr{L}^N$ where a is the unanimity winner.

Then, $(U, \delta_\omega) - score(a, p)$ is $\sum_{i=1}^n \delta(p(i), \bar{p}^a(i))$. By definition of a weighted distance and construction of ω, this equals to $\sum_{i=1}^n \sum_{t=1}^{a(i)-1} \omega_t = \sum_{i=1}^n \sum_{t=1}^{a(i)-1} (s_t - s_{t+1})$, which[9] in turn equals to $\sum_{i=1}^n (s_1 - s_{a(i)}) = n \times s_1 - \sum_{i=1}^n s_{a(i)}$. Note that the score of a in α_s is $\sum_{i=1}^n s_{a(i)}$. Obviously, $n \times s_1 - \sum_{i=1}^n s_{a(i)}$ is minimal if and only if $\sum_{i=1}^n s_{a(i)}$ is maximal. Hence $(U, \delta_\omega) - score(a, p)$ is minimal if and only if $a \in \alpha_s(p)$. This completes the proof as the choice of p and a is arbitrary.

An immediate corollary is on the extension of the result to degenerate scoring rules via weighted pseudo-distances. The proof follows identical reasoning with the theorem above, except where an equal score assigned by the degenerate scoring rule to two adjacent positions leads to a zero weight. This leads to violation of "identity of indiscernibles" condition hence δ_ω is a pseudo distance.

Corollary 1 *A degenerate scoring rule α_s is (U, δ)-rationalizable if $\delta = \delta_\omega$ is a weighted pseudo-distance with $\omega = \Delta s$.*

Let us finally dwell upon the significance of these results. In Example 3, one can see "positional neutrality" leading to assigning the same value so long as the swaps are at the same position. "Decomposability" is also seen in the example via the additivity of distances on pairs that require a single swap. Decomposability is a natural weakening of the original Kemeny [7] betweenness condition. This particular weakening of characterizing conditions lead to the class of weighted distances which rationalize scoring rules. As we already know "Kemeny" and "Borda" are very interconnected, it is interesting to see that a natural "generalization" of the former, i.e., the weighted distances, helps us rationalize the "generalization" of the latter, i.e., the scoring rules.

[9] Note that if a is already at the top of $p(i)$, then this formulation gives 0. The equation $\sum_{i=1}^n \sum_{t=1}^{a(i)-1} \omega_t$ sums the weights (costs) of carrying alternative a to the top in each individual preference.

4 Conclusion

In this paper we show that the relation between the Borda rule and the Kemeny distance is further extended to a relation between all scoring rules and all weighted distances. In fact the relation even spans the degenerate scoring rules in case we extend the weighted functions to pseudo-distances.

The distance rationalization of scoring rules, as mentioned in the introduction, has already been shown in [4], albeit the metrics therein do not resemble the Kemeny distance. The scoring distances proposed in that paper fails to satisfy an additivity condition, i.e., decomposability. This condition is essential in the axiomatisation of the Kemeny distance, as shown [2, 3]. This paper shows in fact that distance rationalization of the scoring rules can be achieved via the weighted distances which mimic the features of the Kemeny distance. Hence, the rationalization result of the Borda rule with the Kemeny distance is carried over naturally to a rationalization result on Borda-like rules with Kemeny-like distances.

Acknowledgements The author appreciates the crucial feedback from two anonymous referees. This research has benefited from Netherlands Organisation for Scientific Research (NWO) grant with project nr. 400-09-354.

References

1. Baigent N (1987) Preference proximity and anonymous social choice. Q J Econ 102(1):161–169
2. Can B (2014) Weighted distances between preferences. J Math Econ 51:109–115
3. Can B, Storcken T (2013) A re-characterization of the Kemeny distance. METEOR Research Memoranda RM/13/009
4. Elkind E, Faliszewski P, Slinko A (2009) On distance rationalizability of some voting rules. In: Proceedings of the 12th conference on theoretical aspects of rationality and knowledge, New York, pp 108–117. doi:10.1145/1562814.1562831. http://doi.acm.org/10.1145/1562814.1562831
5. Elkind E, Faliszewski P, Slinko A (2010) On the role of distances in defining voting rules. In: Proceedings of the 9th international conference on autonomous agents and multiagent systems: volume 1 - volume 1, International Foundation for Autonomous Agents and Multiagent Systems, AAMAS'10, Richland, SC, pp 375–382. http://dl.acm.org/citation.cfm?id=1838206.1838259
6. Elkind E, Faliszewski P, Slinko A (2012) Rationalizations of condorcet-consistent rules via distances of hamming type. Soc Choice Welf 39(4):891–905
7. Kemeny J (1959) Mathematics without numbers. Daedalus 88(4):577–591
8. Meskanen T, Nurmi H (2008) Closeness counts in social choice. In: Braham M, Steffen F (eds) Power, freedom, and voting. Springer, Berlin, pp 289–306
9. Nitzan S (1981) Some measures of closeness to unanimity and their implications. Theory Decis 13(2):129–138

Climate Change and Social Choice Theory

Norman Schofield

Abstract The enlightenment was a philosophical project to construct a rational society without the need for a supreme being . It opened the way for the creation of market democracy and rapid economic growth. At the same time economic growth is the underlying cause of climate change, and we have become aware that this may destroy our civilization. The principal underpinning of the enlightenment project is the *general equilibrium theorem (GET) of* Arrow and Debreu (Econometrica 22:265–290, 1954), asserting the existence of a Pareto optimal price equilibrium. Arrow's work in social choice can be interpreted as an attempt to construct a more general social equilibrium theorem. The current paper surveys recent results in social choice which suggests that chaos rather than equilibrium is generic.

We also consider models of belief aggregation similar to Condorcet's Jury theorem and mention Penn's Theorem on existence of a belief equilibrium.

However, it is suggested that a belief equilibrium with regard to the appropriate response to climate change depends on the creation of a fundamental social principle of "guardianship of our planetary home." It is suggested that this will involve conflict between entrenched economic interests and ordinary people, as the effects of climate change make themselves felt in many countries.

Keywords Black swan events • Climate change • Dynamical models • The enlightenment

1 Introduction

In this essay I shall consider what Israel (2012) calls the *Radical Enlightenment*, the program to establish rationality as the basis for society, opposed to monarchy, religion and the church. Radical enlighteners included Thomas Jefferson, Thomas

N. Schofield (✉)
Center in Political Economy, Washington University in Saint Louis, 1 Brookings Drive, Saint Louis, MO 63130, USA
e-mail: schofield.norman@gmail.com

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_12

Paine and James Madison. They believed that society could be based on rational constitutional principles , leading to the "probability of a fit choice." Implicit in the Radical enlightenment was the belief, originally postulated by Spinoza, that individuals could find moral bases for their choices without a need for a divine creator. An ancillary belief was that the economy would also be rational and that the principles of the radical enlightenment would lead to material growth and the eradication of poverty and misery.[1]This enlightenment philosophy has recently has had to face two troubling propositions. First are the results of Arrovian social choice theory. These very abstract results suggest that no process of social choice can be rational Second, recent events suggest that the market models that we have used to guide our economic actions are deeply flawed. Opposed to the Radical enlighteners, David Hume and Burke believed that people would need religion and nationalism to provide a moral compass to their lives. As Putnam [156] and Putnam and Campbell (2010) have noted religion is as important as it has ever been in the US. Recent models of US Elections [193] show that religion is a key dimension of politics that divides voters one from another. A consequence of the Industrial Revolution, that followed on from the Radical Enlightenment, has been the unintended consequence of climate change. Since this is the most important policy dimension that the world economy currently faces, this paper will address the question whether we are likely to be able to make wise social choices to avoid future catastrophe.

1.1 The Radical Enlightenment

It was no accident that the most important cosmologist after Ptolemy of Alexandria was Nicolaus Copernicus (1473–1543), born only a decade before Martin Luther. Both attacked orthodoxy in different ways.[2] Copernicus formulated a scientifically based heliocentric cosmology that displaced the Earth from the center of the universe. His book, *De revolutionibus orbium coelestium* (*On the Revolutions of the Celestial Spheres*, 1543), is often regarded as the starting point of the Scientific Revolution.

The ideas of Copernicus influenced many scholars: the natural philosopher, William Gilbert, who wrote on magnetism in *De Magnete* (1601); the physicist,

[1]See Pagden [149] for an argument about the significance today of the enlightenment project, but a counter argument by Gray [79–81].

[2]Weber (1904) speculated that there was a connection between the values of Protestantism and Capitalism. It may be that there are connections between the preference for scientific explanation and protestant belief about the relationship between God and humankind.

mathematician, astronomer, and philosopher, Galileo Galilei (1564–1642); the mathematician and astronomer, Johannes Kepler (1571–1630).

Philosophiæ Naturalis Principia Mathematica (1687), by the physicist, mathematician, astronomer and natural philosopher, Isaac Newton (1642–1726) is considered to be the most influential book in the history of science.[3] Margolis [123] argues that, after Newton, a few scholars realized that the universe exhibits laws that can be precisely written down in mathematical form. Moreover, we have, for some mysterious reason, the capacity to conceive of exactly those mathematical forms that do indeed govern reality. We believe that this mysterious connection between mind and reality was the basis for Newton's philosophy. While celestial mechanics had been understood by Ptolemy to be the domain most readily governed by these forms, Newton's work suggested that *all* reality was governed by mathematics. The influence of Newton can perhaps be detected in the work of the philosopher, mathematician, and political scientist, Marie Jean Antoine Nicolas de Caritat, Marquis de Condorcet (1743–1794), known as Nicolas de Condorcet. His work in formal social choice theory [52] was discussed in [189] connection with the arguments about democracy by Madison and Jefferson. The work on Moral Sentiment by the Scottish Enlightenment writers, Francis Hutcheson (1694–1746), David Hume (1711–1776), Adam Smith (1723–1790) and Adam Ferguson (1723–1816), also influenced Jefferson and Madison. Between Copernicus and Newton, the writings of Thomas Hobbes (1588–1679), René Descartes (1596–1650), John Locke (1632–1704), Baruch Spinoza (1632–1677), and Gottfied Liebnitz (1646–1716) laid down foundations for the modern search for rationality in life.[4] Hobbes was more clearly influenced by the scientific method, particularly that of Galileo, while Descartes, Locke, Spinoza, and Liebniz were all concerned in one way or another with the imperishability of the soul.[5] The mathematician, Liebniz, in particular was concerned with an

> [E]xplanation of the relation between the soul and the body, a matter which has been regarded as inexplicable or else as miraculous.

Without the idea of a soul it would seem difficult to form a general scheme of ethics.[6] Indeed, the progress of science and the increasing secularization of society have caused many to doubt that our society can survive. Hawking and Mlodonow

[3] See Feingold (2004).

[4] For Hobbes, see Rogow (1986). For Descartes, see Gaukroger (1995). For Spinoza and Liebnitz see Stewart (2006) and Goldstein (2006). See also Israel (2012) for the development of the Radical Enlightenment.

[5] It is of interest that the English word "soul" derives from Old English *sáwol* (first used in the eighth century poem, *Beowulf*).

[6] Hawking and Mlodinow (2010) assert that God did not create the Universe, perhaps implying that the soul does not exist. However they do say that they understand Isaac Newton's belief that God did "create" and "conserve" order in the universe. See other books by Dawkins [55] (2008) and Hitchens (2007) on the same theme, as well as Wright (2009) on the evolution of the notion of God.

(2010) argue for a strong version of this universal mathematical principle, called *model-dependent realism*, citing its origins in Pythagoras (580 BCE to 490 BCE), Euclid (383-323 BCE and Archimedes (287-212 BCE), and the recent developments in mathematical physics and cosmology.

They argue that it is only through a mathematical model that we can properly perceive reality. However, this mathematical principle faces two philosophical difficulties. One stems from the [74, 220] undecidability theorems. The first theorem asserts that mathematics cannot be both complete and consistent, so there are mathematical principles that in principle cannot be verified. Turing's work, though it provides the basis for our computer technology also suggests that not all programs are computable. The second problem is associated with the notion of *chaos* or *catastrophe*.

Since the early work of Hardin [86] the "tragedy of the commons" has been recognised as a global prisoner' dilemma. In such a dilemma no agent has a motivation to provide for the collective good. In the context of the possibility of climate change, the outcome is the continued emission of greenhouses gases like carbon dioxide into the atmosphere and the acidification of the oceans. There has developed an extensive literature on the n-person prisoners' dilemma in an attempt to solve the dilemma by considering mechanisms that would induce cooperation.[7]

The problem of cooperation has also provided a rich source of models of evolution, building on the early work by Trivers [218] and Hamilton [84, 85]. Nowak [146] provides an overview of the recent developments.Indeed, the last 20 years has seen a growing literature on a game theoretic, or mathematical, analysis of the evolution of social norms to maintain cooperation in prisoners' dilemma like situations. Gintis [71], for example, provides evolutionary models of the cooperation through strong reciprocity and internalization of social norms.[8] The anthropological literature provides much evidence that, from about 500KYBP years ago, the ancestors of *homo sapiens* engaged in cooperative behavior, particularly in hunting and caring for offspring and the elderly.[9] On this basis we can infer that we probably do have very deeply ingrained normative mechanisms that were crucial, far back in time, for the maintenance of cooperation, and the fitness and thus survival

[7] See for example Hardin [87, 88], Taylor [215, 216], Axelrod and Hamilton [12], Axelrod [12, 13], Kreps et al. [109], Margolis [122].

[8] Strong reciprocity means the punishment of those who do not cooperate.

[9] Indeed, White et al. (2009) present evidence of a high degree of cooperation among very early hominids dating back about 4MYBP (million years before the present). The evidence includes anatomical data which allows for inferences about the behavioral characteristics of these early hominids.

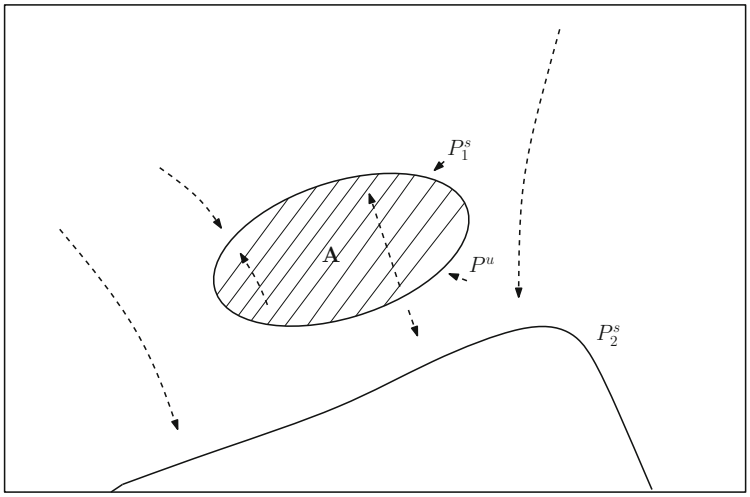


Fig. 1 Stable and unstable components of the global Pareto Set

of early hominids.[10] These normative systems will surely have been modified over the long span of our evolution.

Current work on climate change has focussed on how we should treat the future. For example Stern [206, 207], Collier [51] and Chichilnisky [45, 46] argue essentially for equal treatment of the present and the future. Dasguta [54] points out that how we treat the future depends on our current estimates of economic growth in the near future.

The fundamental problem of climate change is that the underlying dynamic system is extremely complex, and displays many positive feedback mechanisms.[11] The difficulty can perhaps be illustrated by Fig. 1. It is usual in economic analysis to focus on Pareto optimality. Typically in economic theory, it is assumed that preferences and production possibilities are generated by convex sets. However, climate change could create non-convexities. In such a case the Pareto set will exhibit stable and unstable components. Figure 1 distinguishes between a domain A, bounded by stable and unstable components P_1^s and P^u, and a second stable component P_2^s. If our actions lead us to an outcome within A, whether or not it is Paretian, then it is possible that the dynamic system generated by climate could lead to a catastrophic destruction of A itself. More to the point, our society would be trapped inside A as the stable and unstable components merged together.

[10] Gintis cites the work of Robson and Kaplan (2003) who use an economic model to estimate the correlation between brain size and life expectancy (a measure of efficiency). In this context, the increase in brain size is driven by the requirement to solve complex cooperative games against nature.

[11] See the discussion in [192].

Our society has recently passed through a period of economic disorder, where "black swan" events, low probability occurrences with high costs, have occurred with some regularity. Recent discussion of climate change has also emphasized so called "fat-tailed climate events" again defined by high uncertainty and cost.[12] The catastrophic change implied by Fig. 1 is just such a black swan event. The point to note about Fig. 1 is everything would appear normal until the evaporation of A.

Cooperation could in principle be attained by the action of a hegemonic leader such as the United States as suggested by Kindleberger [105] and Keohane and Nye [102]. In Sect. 2 we give a brief exposition of the prisoners' dilemma and illustrate how hegemonic behavior could facilitate international cooperation. However, the analysis suggests that in the present economic climate, such hegemonic leadership is unlikely.

Analysis of games such as the prisoner's dilemma usually focus on the existence of a Nash equilibrium, a vector of strategies with the property that no agent has an incentive to change strategy. Section 3 considers the family of equilibrium models based on the [28] fixed point theorem, or the more general result known as the Ky Fan theorem [62] as well as the application by Bergstrom [21, 22] to prove existence of a Nash equilibrium and market equilibrium.

Section 4 considers a generalization of the Ky Fan Theorem, and argues that the general equilibrium argument can be interpreted in terms of particular properties of a preference field, H, defined on the tangent space of the joint strategy space. If this field is continuous, in a certain well-defined sense, and *"half open"* then it will exhibit a equilibrium. This half open property is the same as the non empty intersection of a family of dual cones. We mention a Theorem by Chichilnisky [40] that a necessary and sufficient condition for market equilibrium is that a family of dual cones also has non-empty intersection.

However, preference fields that are defined in terms of coalitions need not satisfy the half open property and thus need not exhibit equilibrium. For coalition systems, it can be shown that unless there is a collegium or oligarchy, or the dimension of the space is restricted in a particular fashion, then there need be no equilibrium. Earlier results by McKelvey [125], Schofield [173], McKelvey and Schofield [128] and Saari [165] suggested that voting can be "non-equilibrating" and indeed "chaotic."[13]

Kauffman [100] commented on "chaos" or the failure of "structural stability" in the following way.

> One implication of the occurrence or non-occurrence of structural stability is that, in structurally stable systems, smooth walks in parameter space must [result in] smooth changes in dynamical behavior. By contrast, chaotic systems, which are not structurally stable, adapt on uncorrelated landscapes. Very small changes in the parameters pass through many interlaced bifurcation surfaces and so change the behavior of the system dramatically.

[12] Weitzman [225] and Chichilnisky [47]. See also Chichilnisky and Eisenberger [47] on other catastrophic events such as collision with an asteroid.

[13] See Schofield [172, 175, 176]. In a sense these voting theorems can be regarded as derivative of Arrow's Impossibility Theorem [8]. See also Arrow [9].

Chaos is generally understood as sensitive dependence on initial conditions whereas *structural stability* means that the qualitative nature of the dynamical system does not change as a result of a small perturbation.[14] I shall use the term *chaos* to mean that the trajectory taken by the dynamical process can wander anywhere.[15]

An earlier prophet of uncertainty was, of course, Keynes [104] whose ideas on "speculative euphoria and crashes" would seem to be based on understanding the economy in terms of the qualitative aspects of its coalition dynamics.[16] An extensive literature has tried to draw inferences from the nature of the recent economic events. A plausible account of market disequilibrium is given by Akerlof and Shiller [7] who argue that

> the business cycle is tied to feedback loops involving speculative price movements and other economic activity—and to the talk that these movements incite. A downward movement in stock prices, for example, generates chatter and media response, and reminds people of longstanding pessimistic stories and theories. These stories, newly prominent in their minds, incline them toward gloomy intuitive assessments. As a result, the downward spiral can continue: declining prices cause the stories to spread, causing still more price declines and further reinforcement of the stories.

It would seem reasonable that the rise and fall of the market is due precisely to the coalitional nature of decision-making, as large sets of agents follow each other in expecting first good things and then bad. A recent example can be seen in the fall in the market after the earthquake in Japan, and then recovery as an increasing set of investors gradually came to believe that the disaster was not quite as bad as initially feared.

Since investment decisions are based on these uncertain evaluations, and these are the driving force of an advanced economy, the flow of the market can exhibit singularities, of the kind that recently nearly brought on a great depression. These singularities associated with the bursting of market bubbles are time-dependent, and can be induced by endogenous belief-cascades, rather than by any change in economic or political fundamentals [53].

Similar uncertainty holds over political events. The fall of the Berlin Wall in 1989 was not at all foreseen. Political scientists wrote about it in terms of "belief cascades"[17] as the coalition of protesting citizens grew apace. As the very recent democratic revolutions in the Middle East and North Africa suggest, these

[14]The theory of chaos or complexity is rooted in Smale's fundamental theorem [198] that structural stability of dynamical systems is not "generic" or typical whenever the state space has more than two dimensions.

[15]In their early analysis of chaos, Li and Yorke [115] showed that in the domain of a chaotic transformation f it was possible for almost any pair of positions (x, y) to transition from x to $y = f^r(x)$, where f^r means the r times reiteration of f.

[16]See Minsky [135, 136] and Keynes's earlier work in 1921.

[17]Karklins and Petersen [99] and Lohmann [116]. See also Bikhchandani et al. [23].

coalitional movements are extremely uncertain.[18] In particular, whether the autocrat remains in power or is forced into exile is as uncertain as anything Keynes discussed. Even when democracy is brought about, it is still uncertain whether it will persist.[19]

Section 5 introduces the [52] Jury Theorem. This theorem suggests that majority rule can provide a way for a society to attain the truth when the individuals have common goals. Schofield [187, 189] has argued that Madison was aware of this theorem while writing Federalist X [120] so it can be taken as perhaps the ultimate justification for democracy. However, models of belief aggregation that are derived from the Jury Theorem can lead to belief cascades that bifurcate the population. In addition, if the aggregation process takes place on a network, then centrally located agents, who have false beliefs, can dominate the process.[20]

In Sect. 6 we introduce the idea of a belief equilibrium, and then go on to consider the notion of "punctuated equilibrium" in general evolutionary models. Again however, the existence of an equilibrium depends on a fixed point argument, and thus on a half open property of the "cones" by which the developmental path is modeled. This half open property is equivalent to the existence of a social direction gradient defined everywhere. In Sect. 7 we introduce the notion of a "moral compass" that may provide a teleology to guide us in making wise choices for the future, by providing us with a social direction gradient. Section 8 concludes.

2 The Prisoners' Dilemma, Cooperation and Morality

> For before constitution of Sovereign Power ... all men had right to all things; which necessarily causeth Warre. [94].

Kindleberger [105] gave the first interpretation of the international economic system of states as a "Hobbesian" prisoners' dilemma, which could be solved by a leader, or "hegemon."

> A symmetric system with rules for counterbalancing, such as the gold standard is supposed to provide, may give way to a system with each participant seeking to maximize its short-term gain. ... But a world of a few actors (countries) is not like [the competitive system envisaged by Adam Smith]. ... In advancing its own economic good by a tariff, currency depreciation, or foreign exchange control, a country may worsen the welfare of its partners by more than its gain. Beggar-thy-neighbor tactics may lead to retaliation so that each country ends up in a worse position from having pursued its own gain ...
> This is a typical non-zero sum game, in which any player undertaking to adopt a long range solution by itself will find other countries taking advantage of it ...

[18]The response by the citizens of these countries to the demise of Osama bin Laden on May 2, 2011, is in large degree also unpredictable.

[19]See for example Carothers [33] and Collier [50].

[20]Golub and Jackson [76].

In the 1970s, Keohane and Nye [102] rejected "realist" theory in international politics, and made use of the idea of a hegemonic power in a context of "complex interdependence" of the kind envisaged by Kindleberger. Although they did not refer to the formalism of the prisoners' dilemma, it would appear that this notion does capture elements of complex interdependence. To some extent, their concept of a hegemon is taken from realist theory rather than deriving from the game-theoretic formalism.

The essence of the theory of hegemony in international relations is that if there is a degree of inequality in the strengths of nation states then a hegemonic power may maintain cooperation in the context of an n-country prisoners' dilemma. Clearly, the British Empire in the 1800s is the role model for such a hegemon [63].

Hegemon theory suggests that international cooperation was maintained after World War II because of a dominant cooperative coalition. At the core of this cooperative coalition was the United States; through its size it was able to generate collective goods for this community, first of all through the Marshall Plan and then in the context first of the post-world war II system of trade and economic cooperation, based on the Bretton Woods agreement and the Atlantic Alliance, or NATO. Over time, the United States has found it costly to be the dominant core of the coalition In particular, as the relative size of the U.S. economy has declined. Indeed, the global recession of 2008–2010 suggests that problems of debt could induce "beggar thy neighbor strategies", just like the 1930s.

The future utility benefits of adopting policies to ameliorate these possible changes depend on the discount rates that we assign to the future. Dasgupta [54] gives a clear exposition of how we might assign these discount rates. Obviously enough, different countries will in all likelihood adopt very different evaluations of the future. Developing countries like the BRICs (Brazil, Russia, India and China) will choose growth and development now rather than choosing consumption in the future.

There have been many attempts to "solve" the prisoners' dilemma in a general fashion. For example Binmore [24] suggests that in the iterated nPD there are many equilibria with those that are *fair* standing out in some fashion. However, the criterion of "fairness" would seem to have little weight with regard to climate change. It is precisely the poor countries that will suffer from climate change, while the rapidly growing BRICS believe that they have a right to choose their own paths of development.

An extensive literature over the last few years has developed Adam Smith's ideas as expressed in the *Theory of Moral Sentiments* (1984 [1759]) to argue that human beings have an innate propensity to cooperate. This propensity may well have been the result of co-evolution of language and culture [26, 71].

Since language evolves very quickly [58, 129], we might also expect moral values to change fairly rapidly, at least in the period during which language itself was evolving. In fact there is empirical evidence that cooperative behavior as well as

notions of fairness vary significantly across different societies.[21] While there may be fundamental aspects of morality and "altruism," in particular, held in common across many societies, there is variation in how these are articulated. Gazzaniga (2008) suggests that moral values can be described in terms of various *modules*: reciprocity, suffering (or empathy), hierarchy, in-group and outgroup coalition, and purity/ disgust. These modules can be combined in different ways with different emphases. An important aspect of cooperation is emphasized by Burkhart et al. [31] and Hrdy [95], namely cooperation between man and woman to share the burden of child rearing.

It is generally considered that hunter-gatherer societies adopted egalitarian or "fair share" norms. The development of agriculture and then cities led to new norms of hierarchy and obedience, coupled with the predominance of military and religious elites [191].

North [143], North et al. [145] and Acemoglu and Robinson [2] focus on the transition from such oligarchic societies to open access societies whose institutions or "rules of the game", protect private property, and maintain the rule of law and political accountability, thus facilitating both cooperation and economic development. Acemoglu et al. [5] argue, in their historical analyses about why "good" institutions form, that the evidence is in favor of "critical junctures."[22] For example, the "Glorious Revolution" in Britain in 1688 [144], which prepared the way in a sense for the agricultural and industrial revolutions to follow [137–139] was the result of a sequence of historical contingencies that reduced the power of the elite to resist change. Recent work by Morris [140], Fukuyama [68], Ferguson [64], Acemoglu and Robinson [4] has suggested that these fortuitous circumstances never occurred in China and the Middle East, and as a result these domains fell behind the West. Although many states have become democratic in the last few decades, oligarchic power is still entrenched in many parts of the world.[23]

At the international level, the institutions that do exist and that are designed to maintain cooperation, are relatively young. Whether they succeed in facilitating cooperation in such a difficult area as climate change is a matter of speculation. As we have suggested, international cooperation after World War II was only possible because of the overwhelming power of the United States. In a world with oligarchies in power in Russia, China, and in many countries in Africa, together with political disorder in almost all the oil producing counties in the Middle East, cooperation would appear unlikely.

To extend the discussion, we now consider more general theories of social choice.

[21] See Henrich et al. [90, 91], which reports on experiments in fifteen "small-scale societies," using the game theoretic tools of the "prisoners' dilemma," the "ultimatum game," etc.

[22] See also Acemoglu and Robinson [3].

[23] The popular protests in N.Africa and the Middle East in 2011 were in opposition to oligarchic and autocratic power.

3 Existence of a Choice

The above discussion has considered a very simple version of the prisoner's dilemma.The more general models of cooperation typically use variants of evolutionary game theory, and in essence depend on proof of existence of Nash equilibrium, using some version of the Brouwer's fixed point theorem [28].

Brouwer's theorem asserts that any continuous function $f : B \to B$ from the finite dimensional ball,B (or indeed any compact convex set in $\mathbb{R}^w$) into itself, has the *fixed point property*. That is, there exists some $x \in B$ such that $f(x) = x$.

We will now consider the use of variants of the theorem, to prove existence of an equilibrium of a general choice mechanism. We shall argue that the condition for existence of an equilibrium will be violated if there are cycles in the underlying mechanism.

Let W be the set of alternatives and let X be the set of all subsets of W. A *preference correspondence*, P, on W assigns to each point $x \in W$, its *preferred set* $P(x)$. Write $P: W \to X$ or $P: W \twoheadrightarrow W$ to denote that the image of x under P is a set (possibly empty) in W. For any subset V of W, the restriction of P to V gives a correspondence $P_V: V \twoheadrightarrow V$. Define $P_V^{-1}: V \twoheadrightarrow V$ such that for each $x \in V$,

$$P_V^{-1}(x) = \{y : x \in P(y\} \cap V.$$

$P_V^{-1}(x) = \{y : x \in P(y\} \cap V$. The sets $P_V(x)$, $P_V^{-1}(x)$ are sometimes called the *upper* and *lower* preference sets of P on V. When there is no ambiguity we delete the suffix V. The *choice* of P from W is the set

$$C(W, P) = \{x \in W: P(x) = \varnothing\}.$$

Here $\varnothing$ is the empty set. The choice of P from a subset, V, of W is the set

$$C(V, P) = \{x \in V: P_V(x) = \varnothing\}.$$

Call C_P a *choice function* on W if $C_P(V) = C(V, P) \neq \varnothing$ for every subset V of W. We now seek general conditions on W and P which are sufficient for C_P to be a choice function on W. Continuity properties of the preference correspondence are important and so we require the set of alternatives to be a topological space.

Definition 1 Let W, Y be two topological spaces. A correspondence $P: W \twoheadrightarrow Y$ is

(i) *Lower demi-continuous* (*ldc*) iff, for all $x \in Y$, the set

$$P^{-1}(x) = \{y \in W: x \in P(y)\}$$

is open (or empty) in W.

(ii) *Acyclic* if it is impossible to find a cycle $x_t \in P(x_{t-1}), x_{t-1} \in P(x_{t-2}), .., x_1 \in P(x_t)$.

(iii) *Lower hemi-continuous* (*lhc*) iff, for all $x \in W$, and any open set $U \subset Y$ such that $P(x) \cap U \neq \varnothing$ there exists an open neighborhood V of x in W, such that $P(x') \cap U \neq \varnothing$ for all $x' \in V$.

Note that if P is ldc then it is lhc.

We shall use lower demi-continuity of a preference correspondence to prove existence of a choice.

We shall now show that if W is compact, and P is an acyclic and ldc preference correspondence $P: W \twoheadrightarrow W$, then $C(W, P) \neq \varnothing$. First of all, say a preference correspondence $P: W \twoheadrightarrow W$ satisfies the *finite maximality property* (FMP) on W iff for every finite set V in W, there exists $x \in V$ such that $P(x) \cap V = \varnothing$.

Lemma 1 ([221]) *If W is a compact, topological space and P is an ldc preference correspondence that satisfies FMP on W, then $C(W, P) \neq \varnothing$.*

This follows readily, using compactness to find a finite subcover, and then using FMP.

Corollary 1 *If W is a compact topological space and P is an acyclic, ldc preference correspondence on W, then $C(W, P) \neq \varnothing$.*

As Walker [221] noted, when W is compact and P is ldc, then P is acyclic iff P satisfies FMP on W, and so either property can be used to show existence of a choice. A second method of proof is to show that C_P is a choice function is to substitute a convexity property for P rather than acyclicity.

Definition 2 **(i)** If W is a subset of a vector space, then the *convex hull* of W is the set, Con$[W]$, defined by taking all convex combinations of points in W.

(ii) W is *convex* iff $W =$ Con$[W]$. (The empty set is also convex.)

(iii) W is *admissible* iff W is a compact, convex subset of a topological vector space.

(iv) A preference correspondence $P: W \twoheadrightarrow W$ is *semi-convex* iff, for all $x \in W$, it is the case that $x \notin \text{Con}(P(x))$.

Fan [62] has shown that if W is admissible and P is ldc and semi-convex, then $C(W, P)$ is non-empty.

Choice Theorem ([21, 62]) If W is an admissible subset of a Hausdorff topological vector space, and $P: W \twoheadrightarrow W$ a preference correspondence on W which is ldc and semi-convex then $C(W, P) \neq \varnothing$.

The proof uses the KKM lemma due to [106].

The original form of the Theorem by Fan made the assumption that $P: W \twoheadrightarrow W$ was *irreflexive* (in the sense that $x \notin P(x)$ for all $x \in W$) and *convex*. Together these two assumptions imply that P is semi-convex. Bergstrom [21] extended Fan's original result to give the version presented above.[24]

[24] See also Shafer and Sonnenschein [195] who use this result to extend the Arrow Debreu equilibrium existence theorem [10].

Note that the Fan Theorem is valid without restriction on the dimension of W. Indeed, Aliprantis and Brown (1983) have used this theorem in an economic context with an infinite number of commodities to show existence of a price equilibrium. Bergstrom [22] also showed that when W is finite dimensional then the Fan Theorem is valid when the continuity property on P is weakened to lhc and used this theorem to show existence of a Nash equilibrium of a game $G = \{(P_1, W_1), .P_i, W_i), ..(P_n, W_n) : i \in N\}$. Here the ith strategy space is finite dimensional W_i and each individual has a preference P_i on the joint strategy space $P_i : W^N = W_1 \times W_2 \ldots \times W_n \twoheadrightarrow W_i$. The Fan Theorem can be used, in principle to show existence of an equilibrium in complex economies with externalities. Define the Nash improvement correspondence by $P_i^* : W^N \twoheadrightarrow W^N$ by $y \in P_i^*(x)$ whenever $y = (x_1, ..x_{i-1}, x_i^*, \ldots, x_n)$, $x = (x_1, .., x_{i-1}, x_i, .., x_n)$, and $x_i^* \in P_i(x)$ The joint Nash improvement correspondence is $P_N^* = \cup P_i^* : W^N \twoheadrightarrow W^N$. The Nash equilibrium of a game G is a vector $\mathbf{z} \in W^N$ such that $P_N^*(\mathbf{z}) = \varnothing$. Then the Nash equilibrium will exist when P_N^* is ldc and semi-convex and W^N is admissible.

4 Dynamical Choice Functions

We now consider a *generalized preference field* $H : W \twoheadrightarrow TW$, on a manifold W. TW is the tangent bundle above W, given by $TW = \cup\{T_x W : x \in W\}$, where $T_x W$ is the tangent space above x. If V is a neighborhood of x, then $T_V W = \cup\{T_x W : x \in V\}$ which is locally like the product space $\mathbb{R}^w \times V$. Here W is locally like $\mathbb{R}^w$.

At any $x \in W$, $H(x)$ is a *cone* in the tangent space $T_x W$ above x. That is, if a vector $v \in H(x)$, then $\lambda v \in H(x)$ for any $\lambda > 0$. If there is a smooth curve, $c : [-1, 1] \to W$, such that the differential $\frac{dc(t)}{dt} \in H(x)$, whenever $c(t) = x$, then c is called an *integral curve* of H. An integral curve of H from x=$c(o)$ to $y = \lim_{t \to 1} c(t)$ is called an H *-preference curve* from x to y. In this case we write $y \in \mathbb{H}(x)$. We say y is reachable from x if there is a piecewise differentiable $H-$preference curve from x to y, so $y \in \mathbb{H}^r(x)$ for some reiteration r.The preference field H is called *S-continuous* iff the inverse relation $\mathbb{H}^{-1}$ is ldc. That is, if x is reachable from y, then there is a neighborhood V of y such that x is reachable from all of V. The *choice* $C(W, H)$ of H on W is defined by

$$C(W, H) = \{x \in W : H(x) = \varnothing\}.$$

Say $H(x)$ is semi-convex at $x \in W$, if either $H(x) = \varnothing$ or $0 \notin \text{Con}[H(x)]$ in the tangent space $T_x W$. In the later case, there will exist a vector $v' \in T_x W$ such that $(v' \cdot v) > 0$ for all $v \in H(x)$. We can say in this case that there is, at x, a *direction gradient* d in the cotangent space $T_x^* W$ of linear maps from $T_x W$ to $\mathbb{R}$ such that $d(v) > 0$ for all $v \in H(x)$. If H is S-continuous and half-open

in a neighborhood,V, then there will exist such a continuous direction gradient $d : V \to T^*V$ on the neighborhood V[25]

We define

$$Cycle(W, H) = \{x \in W : H(x) \neq \varnothing, 0 \in \text{Con}\, H(x)\}.$$

An alternative way to characterize this property is as follows.

Definition 3 The *dual* of a preference field $H : W \twoheadrightarrow TW$ is defined by $H^* : W \twoheadrightarrow T^*W : x \twoheadrightarrow \{d \in T_x^*W : d(v) > 0 \text{ for all } v \in H(x) \subset T_xW\}$. For convenience if $H(x) = \varnothing$ we let $H^*(x) = T_xW$. Note that if $0 \notin \text{Con}\, H(x)$ iff $H^*(x) \neq \varnothing$. We can say in this case that the field is *half open* at x.

In applications, the field $H(x)$ at x will often consist of some family $\{H_j(x)\}$. As an example , let $u : W \twoheadrightarrow \mathbb{R}^n$ be a smooth utility profile and for any coalition $M \subset N$ let

$$H_M(u)(x) = \{v \in T_xW : (du_i(x)(v) > 0, \forall\, i \in M\}.$$

If $\mathbb{D}$ is a family of *decisive* coalitions, $\mathbb{D} = \{M \subset N\}$, then we define

$$H_{\mathbb{D}}(u) = \cup H_M(u) : W \twoheadrightarrow TW$$

Then the field $H_{\mathbb{D}}(u) : W \twoheadrightarrow TW$ has a dual $[H_{\mathbb{D}}(u)]^* : W \twoheadrightarrow T^*W$ given by $[H_{\mathbb{D}}(u)]^*(x) = \cap[H_M(u)(x)]^*$ where the intersection at x is taken over all $M \in \mathbb{D}$ such that $H_M(u)(x) \neq \varnothing$. We call $[H_M(u)(x)]^*$ the *co-cone of* $[H_M(u)(x)]^*$. It then follows that at $x \in Cycle(W, H_{\mathbb{D}}(u))$ then $0 \in \text{Con}[H_{\mathbb{D}}(u)(x)]$ and so $[H_{\mathbb{D}}(u)(x)]^* = \varnothing$. Thus

$$Cycle(W, H_{\mathbb{D}}(u)) = \{x \in W : [H_{\mathbb{D}}(u)]^*(x) = \varnothing\}.$$

The condition that $[H_{\mathbb{D}}(u)]^*(x) = \varnothing$ is equivalent to the condition that $\cap[H_M(u)(x)]^* = \varnothing$ and was called the *null dual condition* (at x). Schofield [173] has shown that $Cycle(W, H_{\mathbb{D}}(u))$ will be an open set and contains cycles so that a point x is reachable from itself through a sequence of preference curves associated with different coalitions. This result was an application of a more general result.

Dynamical Choice Theorem ([173]) For any S-continuous field H on compact, convex W, then

$$Cycle(W, H) \cup C(W, H) \neq \varnothing.$$

If $x \in Cycle(W, H) \neq \varnothing$ then there is a *piecewise differentiable H -preference cycle* from x to itself. If there is an open path connected neighborhood $V \subset$

[25] *ie* $d(x)(v) > 0$ for all $x \in V$, for all $v \in H(x)$, whenever $H(x) \neq \varnothing$.

Cycle(W, H) such that $H(x')$ is open for all $x' \in V$ then there is a *piecewise differentiable H -preference curve* from x to x'.□

(Here piecewise differentiable means the curve is continuous, and also differentiable except at a finite number of points). The proof follows from the previous choice theorem. The trajectory is built up from a set of vectors $\{v_1, \ldots v_t\}$ each belonging to $H(x)$ with $0 \in \text{Con}[\{v_1, \ldots v_t\}]$. If $H(x)$ is of full dimension, as in the case of a voting rule, then just as in the model of chaos by Li and York [115], trajectories defined in terms of H can wander anywhere within any open path connected component of *Cycle*(W, H).

This result has been shown more generally in [179] for the case that W is a compact manifold with non-zero Euler characteristic [27]. For example the theorem is valid if W is an even dimensional sphere. (The theorem is not true on odd dimensional spheres, as the clock face illustrates.)

Existence of Nash Equilibrium Let $\{W_1, \ldots, W_n\}$ be a family of compact, contractible, smooth, strategy spaces with each $W_i \subset \mathbb{R}^w$. A smooth profile $u{:}W^N = W_1 \times W_2 \ldots \times W_n \twoheadrightarrow \mathbb{R}^n$. Let $H_i : W_i \twoheadrightarrow TW_i$ be the induced i-preference field in the tangent space over W_i. If each H_i is S-continuous and half open in TW_i then there exists a *critical Nash equilibrium,* $\mathbf{z} \in W^N$ such that $H^N(\mathbf{z}) = (H_1 \times ..H_n)(\mathbf{z}) = \varnothing$.

This follows from the choice theorem because the product preference field, H^N, will be half-open and S-continuous. Below we consider existence of *local* Nash equilibrium. With smooth utility functions, a local Nash equilibrium can be found by checking the second order conditions on the Hessians (see [190], for an application of this technique).

Example 1 To illustrate the Choice Theorem, define the preference relation $P_{\mathbb{D}}{:}W \twoheadrightarrow W$ generated by a family of *decisive* coalitions, $\mathbb{D} = \{M \subset N\}$, so that $y \in P_{\mathbb{D}}(x)$ whenever all voters in some coalition $M \in \mathbb{D}$ prefer y to x. In particular consider the example due to [108], with $N = \{1, 2, 3\}$ and $\mathbb{D} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}$Suppose further that the preferences of the voters are characterized by the direction gradients

$$\{du_i(x){:}\, i = 1, 2, 3\}$$

as in Fig. 2. In the figure, the utilities are assume to be "Euclidean," derived from distance from a preferred point, but this assumption is not important.

As the figure makes evident, it is possible to find three points $\{a, b, c\}$ in W such that

$$u_1(a) > u_1(b) = u_1(x) > u_1(c)$$
$$u_2(b) > u_2(c) = u_2(x) > u_2(a)$$
$$u_3(c) > u_3(a) = u_3(x) > u_3(b).$$

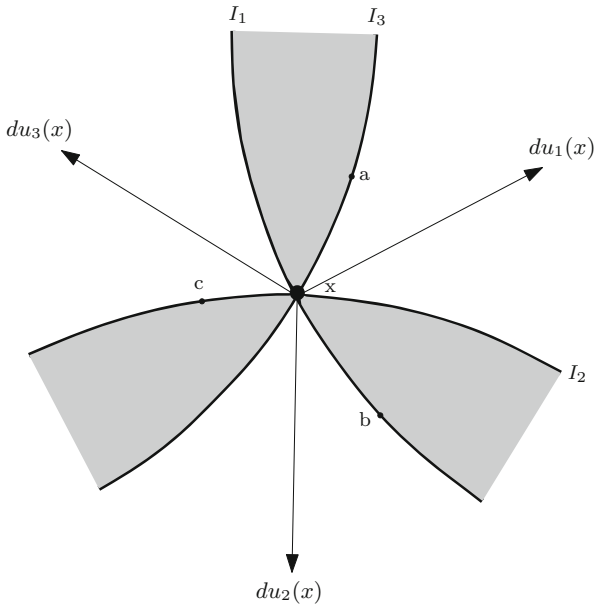


Fig. 2 Cycles in a neighborhood of x

That is to say, preferences on $\{a, b, c\}$ give rise to a *Condorcet cycle*. Note also that the set of points $P_{\mathbb{D}}(x)$, preferred to x under the voting rule, are the shaded "win sets" in the figure. Clearly $x \in \text{Con}\, P_{\mathbb{D}}(x)$, so $P_{\mathbb{D}}(x)$ is not semi-convex. Indeed it should be clear that in *any* neighborhood V of x it is possible to find three points $\{a', b', c'\}$ such that there is *local* voting cycle, with $a' \in P_{\mathbb{D}}(b'), b' \in P_{\mathbb{D}}(c'), c' \in P_{\mathbb{D}}(a')$. We can write this as

$$a' \to c' \to b' \to a'.$$

Not only is there a voting cycle, but the Fan theorem fails, and we have no reason to believe that $C(W, P_{\mathbb{D}}) \neq \varnothing$.

We can translate this example into one on preference fields by considering the preference field

$$H_{\mathbb{D}}(u) = \cup H_M(u) : W \twoheadrightarrow TW$$

where each $M \in \mathbb{D}$.

Figure 3 shows the three difference preference fields $\{H_i : i = 1, 2, 3)$ on W, as well as the intersections H_M, for $M = \{1, 2\}$ etc.
Obviously the joint preference field $H_{\mathbb{D}}(u) = \cup H_M(u) : W \twoheadrightarrow TW$ fails the half open property at x since $0 \in \text{Con}[H_{\mathbb{D}}(u)(x)]$. Although $H_{\mathbb{D}}(u)$ is S-continuous, we cannot infer that $C(W, H_{\mathbb{D}}(u)) \neq \varnothing$.

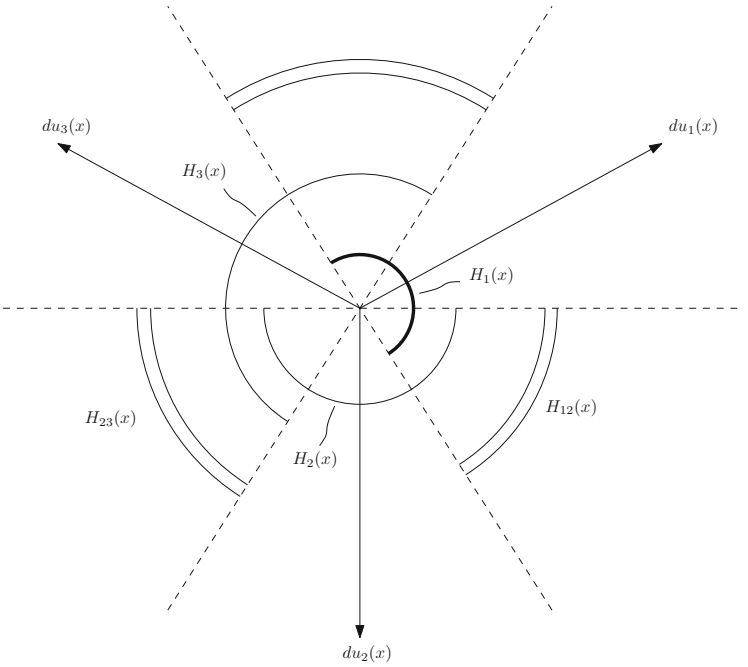


Fig. 3 The failure of half-openess of a preference field

Chichilnisky [38, 40–42] has obtained similar results for markets, where the condition that the dual is non-empty was termed *market arbitrage,* and defined in terms of global market co-cones associated with each player. Such a dual co-cone, $[H_i(u)]^*$ is precisely the set of prices in the cotangent space that lie in the dual of the preferred cone, $[H_i(u)]$, of the agent. By analogy with the above, she identifies this condition on non-emptiness of the intersection of the family of co-cones as one which is necessary and sufficient to guarantee an equilibrium.

Chichilnisky Theorem ([43]) The *limited arbitrage condition* $\cap[H_i(u)]^* \neq \varnothing$ is necessary and sufficient for existence of a competitive equilibrium.□

Chichilnisky [39, 44] also defined a topological obstruction to the non-emptiness of this intersection and showed the connection with the existence of a social choice equilibrium.

For a voting rule, $\mathbb{D}$ it is possible to guarantee that $Cycle(W, H_{\mathbb{D}}) = \varnothing$ and thus that $C(W, H_{\mathbb{D}}) \neq \varnothing$. We can do this by restricting the dimension of W.

Definition 4 **(i)** Let $\mathbb{D}$ be a family of decisive subsets of the finite society N of size n. If the collegium, $K(\mathbb{D}) = \cap\{M \in \mathbb{D}\}$ is non-empty then $\mathbb{D}$ is called *collegial* and the *Nakamura number* $\kappa(\mathbb{D})$ is defined to be ∞.

(ii) If the collegium $K(\mathbb{D})$ is empty then $\mathbb{D}$ is called *non-collegial*. Define the *Nakamura number* in this case to be $\kappa(\mathbb{D}) = \min\{|\mathbb{D}'| : \mathbb{D}' \subset \mathbb{D}$ and $K(\mathbb{D}') = \varnothing\}$.

Nakamura Theorem If $u \in U(W)^N$ and $\mathbb{D}$ has Nakamura number $\kappa(\mathbb{D})$ with $dim(W) \leq \kappa(\mathbb{D}) - 2$ then $Cycle(W, H_{\mathbb{D}}(u)) = \varnothing$ and $C(W, H_{\mathbb{D}}(u)) \neq \varnothing$.

Outline of proof Consider any subfamily $\mathbb{D}'$ of $\mathbb{D}$ with cardinality $\kappa(\mathbb{D})-1$. Then $\cap M \neq \varnothing$, so $\cap\{[H_M(u)]^*(x) : M \in \mathbb{D}'\} \neq \varnothing$. If $[H_M(u)(x)] \neq \varnothing$, we can identify each $[H_M(u)(x)]^*$ with a non-empty convex hull generated by $(du_i(x) : i \in M\}$. These sets can be projected into $T_x W$ where they are convex and compact. Since $dim(W) \leq \kappa(\mathbb{D}) - 2$, then by Helly's Theorem, we see that $\cap\{[H_M(u)]^*(x) : M \in \mathbb{D}\} \neq \varnothing$. Thus $Cycle(W, H_{\mathbb{D}}(u)) = \varnothing$ and $C(W, H_{\mathbb{D}}(u)) \neq \varnothing$.□

See Schofield [180], Nakamura [142] and Strnad [204].

For social choice defined by voting games, the Nakamura number for majority rule is 3, except when $n = 4$, in which case $\kappa(\mathbb{D}) = 4$, so the Nakamura Theorem can generally only be used to prove a "median voter" theorem in one dimension. However, the result can be combined with the Fan Theorem to prove existence of equilibrium for a political economy with voting rule $\mathbb{D}$, when the dimension of the public good space is no more than $\kappa(\mathbb{D}) - 2$ (Konishi 1996). Recent work in political economy often only considers a public good space of one dimension [2]. Note however, that if $\mathbb{D}$ is collegial, then $Cycle(W, H_{\mathbb{D}}(u)) = \varnothing$.and $C(W, H_{\mathbb{D}}(u)) \neq \varnothing$. Such a rule can be called oligarchic, and this inference provides a theoretical basis for comparing democracy and oligarchy [1]. Figure 3 showed the preference cones in a majority voting game with 3 agents and Nakamura number 3, so half openess fails in two dimensions.

Extending the equilibrium result of the Nakamura Theorem to higher dimension for a voting rule faces a difficulty caused by Bank's Theorem. We first define a *fine* topology on smooth utility functions [92, 186, 188].

Definition 5 Let $(U(W)^N, T_1)$ be the topological space of smooth utility profiles endowed with the C^1−topology. See [188] for definition.

In economic theory, the existence of isolated price equilibria can be shown to be "generic" in this topological space [56, 57, 199, 200]. In social choice no such equilibrium theorem holds. The difference is essentially because of the coalitional nature of social choice.

Banks Theorem For any non-collegial $\mathbb{D}$, there exists an integer $w(\mathbb{D}) \geq \kappa(\mathbb{D})-1$ such that $dim(W) > w(\mathbb{D})$ implies that $C(W, H_{\mathbb{D}}(u)) = \varnothing$ for all u in a dense subspace of $(U(W)^N, T_1)$ so $Cycle(W, H_{\mathbb{D}}(u)) \neq \varnothing$ generically.□

This result was essentially proved by Banks [16], building on earlier results by Plott [154], Kramer [108], McKelvey [126], Schofield [177, 178], McKelvey and Schofield [128]. See [162, 163, 165–168] for related analyses. Indeed, it can be shown that if $dim(W) > w(\mathbb{D}) + 1$ then $Cycle(W, H_{\mathbb{D}}(u))$ is generically dense [181].

The integer $w(\mathbb{D})$ can usually be computed explicitly from $\mathbb{D}$. For majority rule with n odd it is known that $w(\mathbb{D}) = 2$ while for n even, $w(\mathbb{D}) = 3$.
Although the Banks Theorem formally applies only to voting rules, [191] argues that it is applicable to any non-collegial social mechanism, say $H(u)$ and can be interpreted to imply that

$$Cycle(W, H(u)) \neq \varnothing \text{ and } C(W, H(u)) = \varnothing$$

is a generic phenomenon in coalitional systems. Because preference curves can wander anywhere in any open component of $Cycle(W, H(u))$, [174] called this *chaos*. It is not so much the sensitive dependence on initial conditions, but the aspect of indeterminacy that is emphasized. On the other hand, existence of a hegemon, as discussed in Sect. 2, is similar to existence of a collegium, suggesting that $Cycle(W, H(u))$ would be constrained in this case.

Richards (1990) has examined data on the distribution of power in the international system over the long run and presents evidence that it can be interpreted in terms of a chaotic trajectory. This suggests that the metaphor of the nPD in international affairs does characterise the ebb and flow of the system and the rise and decline of hegemony.

It is worth noting that the early versions of the Banks Theorem were obtained in the decade of the 1970s, a decade that saw the first oil crisis, the collapse of the Bretton Woods system of international political economy, the apparent collapse of the British economy, the beginning of social unrest in Eastern Europe, the revolution in Iran, and the second oilcrisis (Caryl 2011). Many of the transformations that have occurred since then can be seen as changes in beliefs, rather than preferences. Models of belief aggregation are less well developed than those dealing with preferences.[26] In general models of belief aggregation are related to what is now termed Condorcet's jury Theorem, which we now introduce.

5 Beliefs and Condorcet's Jury Theorem

The Jury theorem formally only refers to a situation where there are just two alternatives $\{1, 0\}$, and alternative 1 is the "true" option. Further, for every individual, i, it is the case that the probability that i picks the truth is ρ_{i1}, which exceeds the probability ,ρ_{i0}, that i does not pick the truth..We can assume that $\rho_{i1} + \rho_{i0} = 1$, so obviously $\rho_{i1} > \frac{1}{2}$. To simplify the proof, we can assume that ρ_{i1} is the same for every individual, thus $\rho_{i1} = \alpha > \frac{1}{2}$ for all i. We use χ_i $(= 0 \text{ or } 1)$ to refer to the choice of individual i, and let $\chi = \Sigma_{i=1}^{n} \chi_i$ be the number of individuals who select the true option 1. We use Pr for the probability operator, and E for the expectation operator. In the case that the electoral size, n, is odd, then a majority, m, is defined

[26] Results on belief aggregation include [153] and [127].

to be $m = \frac{n+1}{2}$. In the case n is even, the majority is $m = \frac{n}{2} + 1$. The probability that a majority chooses the true option is then

$$\alpha^n_{maj} = \Pr[\chi \geq m].$$

The theorem assumes that voter choice is *pairwise independent*, so that $\Pr(\chi = j)$ is simply given by the binomial expression $\binom{n}{j}\alpha^j(1-\alpha)^{n-j}$.

A version of the theorem can be proved in the case that the probabilities $\{\rho_{i1} = \alpha_i\}$ differ but satisfy the requirement that $\frac{1}{n}\Sigma^n_{i=1}\alpha_i > \frac{1}{2}$. Versions of the theorem are valid when voter choices are not pairwise independent [113].

The Jury Theorem If $1 > \alpha > \frac{1}{2}$, then $\alpha^n_{maj} \geq \alpha$, and $\alpha^n_{maj} \longrightarrow 1$ as $n \longrightarrow \infty$.

For both n being even or odd, as $n \longrightarrow \infty$, the fraction of voters choosing option 1 approaches $\frac{1}{n}E(\chi) = \alpha > \frac{1}{2}$. Thus, in the limit, more than half the voters choose the true option. Hence the probability $\alpha^n_{maj} \longrightarrow 1$ as $n \longrightarrow \infty$.□

Laplace also wrote on the topic of the probability of an error in the judgement of a tribunal. He was concerned with the degree to which jurors would make just decisions in a situation of asymmetric costs, where finding an innocent party guilty was to be more feared than letting the guilty party go free. As he commented on the appropriate rule for a jury of twelve, "I think that in order to give a sufficient guarantee to innocence, one ought to demand at least a plurality of nine votes in twelve" [114]. Schofield [169, 170] considered a model derived from the jury theorem where uncertain citizens were concerned to choose an ethical rule which would minimize their disappointment over the likely outcomes, and showed that majority rule was indeed optimal in this sense.

Models of belief aggregation extend the Jury theorem by considering a situation where individuals receive signals, update their beliefs and make an aggregate choice on the basis of their posterior beliefs [11]. Models of this kind can be used as the basis for analysing correlated beliefs.[27] and the creation of belief cascades [59].

Schofield [187, 189] has argued that Condorcet's Jury theorem provided the basis for Madison's argument in Federalist X [120] that the judgments of citizens in the extended Republic would enhance the "probability of a fit choice." However, Schofield's discussion suggests that belief cascades can also fracture the society in two opposed factions, as in the lead up to the Civil War in 1860.[28]

There has been a very extensive literature recently on cascades[29] but it is unclear from this literature whether cascades will be equilibrating or very volatile. In their formal analysis of cascades on a network of social connections, Golub and Jackson [76] use the term *wise* if the process can attain the truth. In particular they note that

[27] Schofield [169, 170], Ladha [111–113], Ladha and Miller [113].

[28] Sunstein [209, 211] also notes that belief aggregation can lead to a situation where subgroups in the society come to hold very disparate opinions.

[29] Gleick [73], Buchanan [29, 30], Gladwell [72], Johnson [97], Barabasi [17, 18], Strogatz [205], Watts [222, 223], Surowiecki [212], Ball [15], Christakis and Fowler [49]

if one agent in the network is highly connected, then untrue beliefs of this agent can steer the crowd away from the truth. The recent economic disaster has led to research on market behavior to see if the notion of cascades can be used to explain why markets can become volatile or even irrational in some sense [6, 194]. Indeed the literature that has developed in the last few years has dealt with the nature of herd instinct, the way markets respond to speculative behavior and the power law that characterizes market price movements.[30] The general idea is that the market can no longer be regarded as efficient. Indeed, as suggested by Ormerod [147] the market may be fundamentally chaotic.

"Empirical" chaos was probably first discovered by Lorenz [117, 118] in his efforts to numerically solve a system of equations representative of the behavior of weather. A very simple version is the non-linear vector equation

$$\frac{dx}{dt} = \begin{bmatrix} dx_1 \\ dx_2 \\ dx_3 \end{bmatrix} = \begin{bmatrix} -a_1(x_1 - x_2) \\ -x_1x_3 + a_2x_1 - x_2 \\ x_1x_2 - a_3x_3 \end{bmatrix}$$

which is chaotic for certain ranges of the three constants, a_1, a_2, a_3.

The resulting "butterfly" portrait winds a number of times about the left hole (as in Fig. 3),then about the right hole,then the left, etc. Thus the "phase prortrait" of this dynamical system can be described by a sequence of winding numbers $(w_l^1, w_k^1, w_l^2, w_k^2$, etc.). Changing the constants a_1, a_2, a_3 slightly changes the winding numbers. Note that the picture in Fig. 3 is in three dimensions. The butterfly wings on left and right consist of infinitely many closed loops. Figure 5 gives a version of the butterfly, namely the chaotic trajectory of the Artemis Earth Moon orbiter. The whole thing is called the Lorentz "strange attractor." A slight perturbation of this dynamic system changes the winding numbers and thus the qualitative nature of the process. Clearly this dynamic system is not structurally stable, in the sense used by Kaufmann [100]. The metaphor of the butterfly gives us pause, since all dynamic systems whether models of climate, markets, voting processes or cascades may be indeterminate or chaotic.

6 The Edge of Chaos

Recent work has attempted to avoid chaos by using the Brouwer fixed point theorem to seek existence of a *belief equilibrium* for a society N_τ of size $n_{\tau.}$ time τ. In this context we let

$$W_E = W_1 \times W_2 \ldots \times W_{n_{\tau+1.}} \times \Delta$$

[30]See, for example, Mandelbrot and Hudson [121], Shiller [196, 197], Taleb [213], Barbera [19], Cassidy [35], Fox [67].

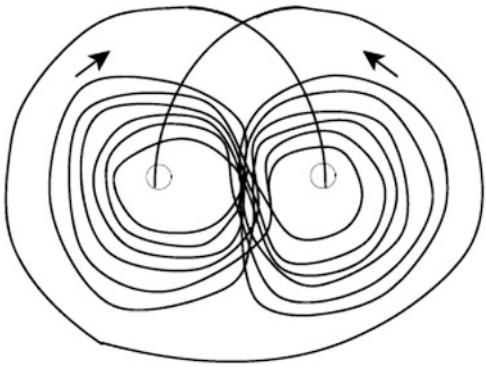


Fig. 4 The butterfly

be the economic product space, where W_i is the commodity space for citizen i and Δ is a price simplex.. Let W_E be the economic space and $W_{\mathbb{D}}$ be a space of political goods, governed by a rule $\mathbb{D}$. At time τ, $W_\tau = W_E \times W_{\mathbb{D}}$ is the political economic space.

At τ, each individual, i, is described by a utility function $u_i : W_\tau \to \mathbb{R}$, so the population profile is given by $u : W_\tau \to \mathbb{R}^{n_\tau}$. Beliefs at τ about the future $\tau + 1$ are given by a stochastic rule, $\mathbb{Q}_\tau$, that transforms the agents' utilities from those at time τ to those at time $\tau + 1$. Thus $\mathbb{Q}_\tau$ generates a new profile for $N_{\tau+1}$ at $\tau + 1$ given by $\mathbb{Q}_\sim(u) = u' : W_{\tau+1} \to \mathbb{R}^{n_\tau+1}$. The utility and beliefs of i will depend on the various sociodemographic subgroups in the society $N_{\tau.}$ that i belongs to, as well as information about the current price vector in Δ.

Thus we obtain a transformation on the function space $[W_\tau \to \mathbb{R}^{n_{fi}}]$ given by

$$[W_{fi} \to \mathbb{R}^{n_{fi}}] \to {}^{\mathbb{Q}_\sim}\to [W_{fi} \to \mathbb{R}^{n_{fi}+1}] \to [W_{fi} \to \mathbb{R}^{n_{fi}}]$$

The second transformation here is projection onto the subspace $[W_\tau \to \mathbb{R}^{n_\tau}]$ obtained by restricting to changes to the original population $N_{\tau.}$ and space.

A *dynamic belief equilibrium* at τ for $N_{\tau.}$ is fixed point of this transformation. Although the space $[W_{fi} \to \mathbb{R}^{n_{fi}}]$ is infinite dimensional, if the domain and range of this transformation are restricted to *equicontinous* functions [155], then the domain and range will be compact. Penn [153] shows that if the domain and range are convex then a generalized version of Brouwer's fixed point theorem can be applied to show existence of such a dynamic belief equilibrium. This notion of equilibrium was first suggested by Hahn [83] who argued that equilibrium is located in the mind, not in behavior.

However, the choice theorem suggests that the validity of Penn's result will depend on how the model of social choice is constructed. For example [53] consider a formal model of the market, based on the reasoning behind Keynes's "beauty contest" [104]. There are two coalitions of "bulls" and "bears". Individuals randomly sample opinion from the coalitions and use a *critical* cutoff-rule. For example if the individual is bullish and the sampled ratio of bears exceeds some

proportion then the individual flips to bearish. The model is very like that of the Jury Theorem but instead of guaranteeing a good choice the model can generate chaotic flips between bullish and bearish markets, as well as fixed points or cyclic behavior, depending on the cut-off parameters. Taleb's argument [213] about black swan events can be applied to the recent transformation in societies in the Middle East and North Africa that resemble such a cascade [214]. As in the earlier episodes in Eastern Europe, it would seem plausible that the sudden onset of a cascade is due to a switch in a critical coalition.

The notion of "criticality" has spawned in enormous literature particularly in fields involving evolution, in biology, language and culture.[31] Bak and Sneppen [14] refer to the self organized critical state as the

> "edge of chaos" since it separates a frozen inactive state from a "hot" disordered state. The mechanism of evolution in the critical state can be thought of as an exploratory search for better local fitness, which is rarely successful, but sometimes has enormous effect on the ecosystem

Flyvbjerg et al. [66] go on to say

> species sit at local fitness maxima..and occasionally a species jumps to another maximum [in doing so it] may change the fitness landscapes of other species which depend on it... Consequently they immediately jump to new maxima. This may affect yet another species in a chain reaction, a *burst* of evolutionary activity.

This work was triggered by the earlier ideas on "punctuated equilibrium" by Eldredge and Gould [61].[32]

The point to be emphasized is that the evolution of a species involves bifurcation, a splitting of the pathway. We can refer to the bifurcation as a *catastrophe* or a *singularity*. The portal or door to the singularity may well be characterized by chaos or uncertainty, since the path can veer off in many possible directions, as suggested by the bifurcating cones in Figs. 3 and 4. At every level that we consider, the bifurcations of the evolutionary trajectory seem to be locally characterized by chaotic domains. I suggest that these domains are the result of different coalitional possibilities. The fact that the trajectories can become indeterminate suggests that this may enhance the exploration of the fitness landscape.

A more general remark concerns the role of climate change. Climate has exhibited chaotic or catastrophic behavior in the past.[33] There is good reason to believe that human evolution over the last million years can only be understood in terms of "bursts" of sudden transformations [146] and that language and culture co-evolve through group or coalition selection [37]. Calvin [32] suggests that our braininess was cause and effect of the rapid exploration of the fitness landscape

[31] See for example Cavallli-Sforza and Feldman [37], Bowles et al. [25].

[32] See also Eldredge [60] and Gould (1976).

[33] Indeed as I understand the dynamical models, the chaotic episodes are due to the complex interactions of dynamical processes in the oceans, on the land, in weather, and in the heavens. These are very like interlinked *coalitions* of non-gradient vector fields.

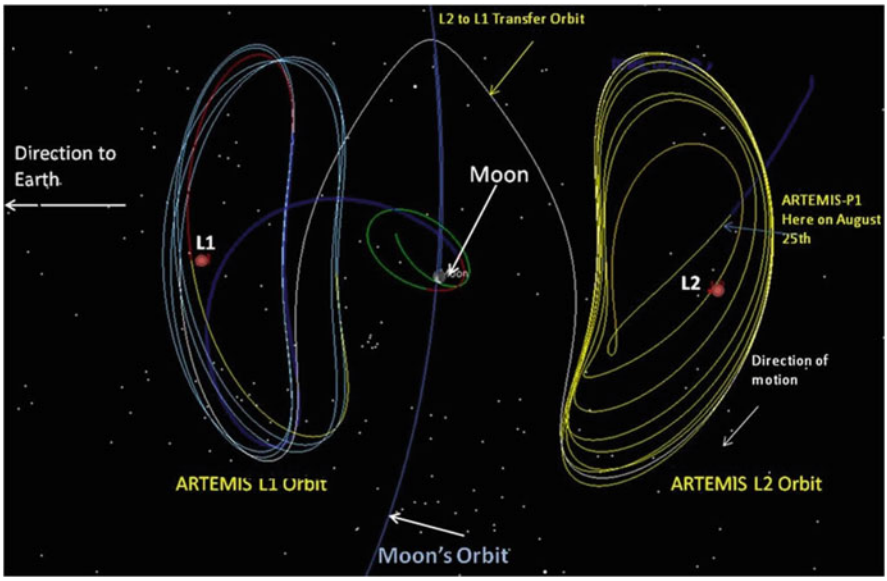


Fig. 5 A chaotic trajectory of the Artemis Earth Moon orbiter, downloaded from nasa.gov (artemis orbiter)

in response to climatic forcing. For example Fig. 6 shows the rapid changes in temperature over the last 100,000 years. It was only in the last period of stable temperature, the "holocene", the last 10,000 years that agriculture was possible.

Stringer (2012) calls the theory of rapid evolution during a period of chaotic climate change" the Social Brain hypothesis." The cave art of Chauvet, in France dating back about 36,000 years suggests that belief in the supernatural played an important part in human evolution. Indeed, we might speculate that the part of our mind that enhances technological/ mathematical development and that part that facilitates social/ religious belief are in conflict with each other.[34] We might also speculate that market behavior is largely driven by what Keynes termed *speculation*, namely the largely irrational changes of *mood* (Casti 2010). Figure 7 gives an illustration of the swings in the US stock market over the last 80 years. While the figure may not allow us to assert that it truly chaotic, there seems no evidence that it is equilibrating.

[34]This is suggested by Kahneman [98].

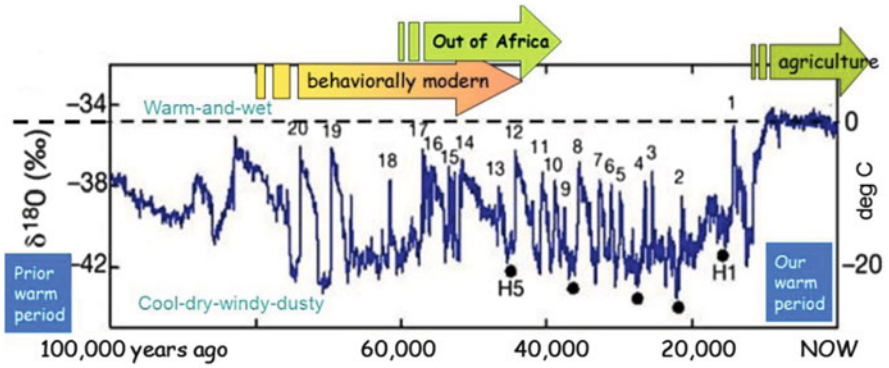


Fig. 6 Climate 100KYBP to now: chaos from 90KYBP to 10KYBP (*Source*: Global-Fever.org)

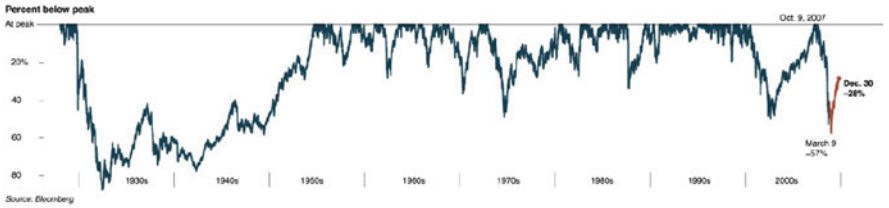


Fig. 7 Chaotic stock market prices 1930–2009 (*Source*: New York Times, Dec 31, 2009)

7 A Moral Compass

If we accept that moral and religious beliefs are as important as rational calculations in determining the choices of society, then depending on models of preference aggregation will not suffice in helping us to make decisions over how to deal with climate change. Instead, I suggest a moral compass, derived from current inferences made about the nature of the evolution of intelligence on our planetary home. The anthropic principle reasons that the fundamental constants of nature are very precisely tuned so that the universe contains matter and that galaxies and stars live long enough to allow for the creation of carbon, oxygen etc., all necessary for the evolution of life itself.[35] Gribbin [82] goes further and points out that not only is the sun unusual in having the characteristics of a structurally stable system of planets, but the earth is fortunate in being protected by Jupiter from chaotic bombardment

[35] As Smolin [203] points out, the anthropic principle has been adopted because of the experimental evidence that the expansion of the universe is accelerating. Indeed it has led to the hypothesis that there is an infinity of universes all with different laws. An alternative inference is the principle of intelligent design. My own inference is that we require a teleology as proposed in the conclusion.

but the Moon also stabilizes our planet's orbit.[36] In essence Gribbin gives good reasons to believe that our planet may well be the only planet in the galaxy that sustains intelligent life. If this is true then we have a moral obligation to act as guardians of our planetary home. Parfit [151] argues

> What matters most is that we rich people give up some of our luxuries, ceasing to overheat the Earth's atmosphere, and taking care of this planet in other ways, so that it continues to support intelligent life. If we are the only rational animals in the Universe, it matters even more whether we shall have descendants during the billions of years in which that would be possible. Some of our descendants might live lives and create worlds that, though failing to justify past suffering, would give us all, including those who suffered, reason to be glad that the Universe exists. (Parfit: 419)

8 Conclusion

Even if we believe that markets are well behaved, there is no reason to infer that markets are able to reflect the social costs of the externalities associated with production and consumption. Indeed Gore (2006) argues that the globalized market place, what he calls *Earth Inc* has the power and inclination to maintain business as usual. If this is so, then climate change will undoubtedly have dramatic adverse effects, not least on the less developed countries of the world. [37]

In principle we may be able to rely on a version of the jury theorem Rae (1960) and [169, 170, 210], which asserts that majority rule provides an optimal procedure for making collective choices under uncertainty However, for the operation of what Madison called a "fit choice" it will be necessary to overcome the entrenched power of capital. Although we now disregard Marx's attempt at constructing a teleology of economic and political development,[38] we are in need of a more complex overarching and evolutionary theory of political economy that will go beyond the notion of equilibrium and might help us deal with the future.[39]

[36]The work by Poincare in the late nineteenth century focussed on the structural stability of the solar system and was the first to conceive of the notion of chaos.

[37]Zhang et al. [230] and Hsiang et al. [96] have provided quantitative analyses of such adverse effects in the past. See also Parker [152] for an historical account of the effect of climate change in early modern Europe.

[38]See Sperber [202] for a discussion of the development of Marx's ideas, in the context of nineteenth century belief in the teleology of "progreśś" or the advance of civilization. The last 100 years has however,made it difficult to hold such beliefs.

[39]The philosopher [141] argues that without a teleogy of some kind, we are left with Darwinian evolutionary theory, which by itself cannot provide a full explanation of what we are and where we are going. See also [217] and [20].

References

1. Acemoglu D (2008) Oligarchic versus democratic societies. J Eur Econ Assoc 6:1–44
2. Acemoglu D, Robinson J (2006) Economic origins of dictatorship and democracy. Cambridge University Press, Cambridge
3. Acemoglu D, Robinson J (2008) Persistence of power, elites, and institutions. Am Econ Rev 98:267–293
4. Acemoglu D, Robinson J (2011) Why nations fail. Profile Books, London
5. Acemoglu D, Johnson S, Robinson J, Yared P (2009) Reevaluating the modernization hypothesis. J Monet Econ 56:1043–1058
6. Acemoglu D, Ozdaglar A, Tahbaz-Salehi A (2010) Cascades in networks and aggregate volatility. NBER working paper # 16516
7. Akerlof GA, Shiller RJ (2009) Animal spirits. Princeton University Press, Princeton
8. Arrow KJ (1951) Social choice and individual values. Yale University Press, New Haven
9. Arrow K (1986) Rationality of self and others in an economic system. J Bus 59:S385–S399
10. Arrow K, Debreu G (1954) Existence of an equilibrium for a competitive economy. Econometrica 22:265–290
11. Austen-Smith D, Banks J (1996) Information aggregation, rationality, and the Condorcet Jury theorem. Am Polit Sci Rev 90:34–45
12. Axelrod R (1981) The emergence of cooperation among egoists. Am Polit Sci Rev 75:306–318
13. Axelrod R (1984) The evolution of cooperation. Basic, New York
14. Bak P, Sneppen (1993) Punctuated equilibrium and criticality in a simple model of evolution. Phys Rev Lett 71(24):4083–4086
15. Ball P (2004) Critical mass. Ferrar, Strauss and Giroux, New York
16. Banks JS (1995) Singularity theory and core existence in the spatial model. J Math Econ 24:523–536
17. Barabasi A-L (2003) Linked. Plume, New York
18. Barabasi A-L (2010) Bursts. Dutton, New York
19. Barbera R (2009) The cost of capitalism: understanding market mayhem. McGraw Hill, New York
20. Bellah (2011) Religion in human evolution. Belknap, Cambridge, MA
21. Bergstrom T (1975) The existence of maximal elements and equilibria in the absence of transitivity. University of Michigan, Typescript
22. Bergstrom T (1992) When non-transitive relations take maxima and competitive equilibrium can't be beat. In: Neuefeind W, Riezman R (eds) Economic theory and international trade. Springer, Berlin
23. Bikhchandani S, Hirschleifer D, Welsh I (1992) A theory of fads, fashion, custom, and cultural change as information cascades. J Polit Econ 100:992–1026
24. Binmore K (2005) Natural justice. Oxford University Press, Oxford
25. Bowles S et al (2003) The co-evolution of individual behaviors and socal institutions. J Theor Biol 223:135–147
26. Boyd J, Richerson PJ (2005) The origin and evolution of culture. Oxford University Press, Oxford
27. Brown R (1971) The Lefschetz fixed point theorem. Scott and Foreman, Glenview, IL
28. Brouwer LEJ (1912) Uber abbildung von mannigfaltikeiten. Math Analen 71:97–115
29. Buchanan M (2001) Ubiquity. Crown, New York
30. Buchanan M (2003) Nexus. Norton, New York
31. Burkhart JM, Hrdy SB, van Schaik CP (2009) Cooperative breeding and human cognitive evolution. Evol Anthropol 18:175–186
32. Calvin WH (2003) The ascent of mind. Bantam, New York
33. Carothers T (2002) The end of the transition paradigm. J Democr 13:5–21
34. Caryl (2011) Strange Rebels 1979 and the birth of the 20th century Basic Books. New York

35. Cassidy J (2009) How markets fail: the logic of economic calamities. Farrar, Strauss and Giroux, New York
36. Casti J (2010) Mood matters Copernicus. New York
37. Cavallli-Sforza L, Feldman M (1981) Cultural transmission and evolution. Princeton University Press, Princeton, NJ
38. Chichilnisky G (1992) Social diversity, arbitrage, and gains from trade: a unified perspective on resource allocation. Am Econ Rev 84:427–434
39. Chichilnisky G (1993) Intersecting families of sets and the topology of cones in economics. Bull Am Math Soc 29:189–207
40. Chichilnisky G (1995) Limited arbitrage is necessary and sufficient for the existence of a competitive equilibrium with or without short sales. Econ Theory 5:79–107
41. Chichilnisky G (1996) Markets and games: a simple equivalence among the core, equilibrium and limited arbitrage. Metroeconomica 47:266–280
42. Chichilnisky G (1997) A topological invariant for competitive markets. J Math Econ 28:445–469
43. Chichilnisky G (1997) Limited arbitrage is necessary and sufficient for the existence of a equilibrium. J Math Econ 28:470–479
44. Chichilnisky G (1997) Market arbitrage, social choice and the core. Soc Choice Welf 14:161–198
45. Chichilnisky G (2009) The topology of fear. J Math Econ 45:807–816
46. Chichilnisky G (2009) Avoiding extinction: equal treatment of the present and the future. Working Paper: Columbia University
47. Chichilinsky G (2010) The foundations of statistics with black swans. Math Soc Sci 59:184–192
48. Chichilinsky G (2012) Sustainable markets with short sales. Econ Theory 49:293–307
49. Christakis N, Fowler JH (2011) Connected. Back Bay, New York
50. Collier P (2009) Wars, guns and votes. Harper, New York
51. Collier P (2010) The plundered planet. Oxford University Press, Oxford
52. Condorcet N (1994 [1785]) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Imprimerie Royale, Paris. Translated in part in: McLean I, Hewitt F (eds) Condorcet: foundations of social choice and political theory. Edward Elgar Publishing, Aldershot
53. Corcos et al (2002) Imitation and contrarian behavior: hyperbolic bubbles, crashes and chaos. Quant Finan 2:264–281
54. Dasgupta P (2005) Three conceptions of intergenerational Justice. In: Lillehammer H, Mellor DH (eds) Ramsey's legacy. Clarendon Press, Oxford
55. Dawkins R (1976) The selfish gene. Oxford University Press, Oxford
56. Debreu G (1970) Economies with a finite number of equilibria. Econometrica 38:387–392
57. Debreu G (1976) The application to economics of differential topology and global analysis: regular differentiable economies. Am Econ Rev 66:280–287
58. Deutscher G (2006) The unfolding of language. Holt, New York
59. Easley D, Kleinberg J (2010) Networks, crowds and markets . Cambridge University Press, Cambridge
60. Eldredge N (1976) Differential evolutionary rates. Paleobiology 2:174–177
61. Eldridge N, Gould SJ (1972) Punctuated equilibrium. In: Schopf T (ed) Models of paleobiology. Norton, New York
62. Fan K (1961) A generalization of Tychonoff's fixed point theorem. Math Ann 42:305–310
63. Ferguson N (2002) Empire: the rise and demise of the British world order. Penguin Books, London
64. Ferguson N (2011) Civilization. Penguin, London
65. Feingold M (2004) The Newtonian moment. Oxford University Press, Oxford
66. Flyvbjerg H, Sneppen K, Bak P (1993) A mean field theory for a simple model of evolution. Phys Rev Lett 71:4087–4090
67. Fox J (2009) The myth of the rational market. Harper, New York

68. Fukuyama F (2011) The origins of political order. Ferrar, Strauss and Giroux, New York
69. Gaukroger S (1995) Descartes. Oxford University Press, Oxford
70. Gazzaniger M S (2008) Human Harper. New York
71. Gintis H (2000) Strong reciprocity and human sociality. J Theor Biol 206:169–179
72. Gladwell M (2002) The tipping point. Back Bay, New York
73. Gleick J (1987) Chaos: making a new science. Viking, New York
74. Gödel K (1931) Uber formal unentscheidbare Satze der Principia Mathematica und verwandter Systeme. Monatschefte fur Mathematik und Physik 38:173–98. Translated as On formally undecidable propositions of Principia Mathematica and related systems. In: van Heijenoort J (ed) Frege and G ödel: Two Fundamental Texts in Mathematical Logic. Harvard University Press, Cambridge, MA
75. Goldstein R (2006) Betraying Spinoza. Random House, New York
76. Golub B, Jackson M (2010) Naive learning in social networks and the wisdom of crowds. Am Econ J 2:112–149
77. Gould SJ (1976) Full House. Belknap, New York
78. Gore A (2006) The Future Random. New York
79. Gray J (1995) Enlightenment's wake. Routledge, London
80. Gray J (1997) Endgames. Blackwell, London
81. Gray J (2000) False dawn. New Press, London
82. Gribbin J (2011) Alone in the universe. Wiley, NewYork
83. Hahn F (1973) On the notion of equilibrium in economics. Cambridge University Press, Cambridge .
84. Hamilton W (1964) The genetical evolution of social behavior I and II. J Theor Biol 7:1–52
85. Hamilton W (1970) Selfish and spiteful behavior in an evolutionary model. Nature 228:1218–1220
86. Hardin G (1968 [1973]) The tragedy of the commons. In: Daly HE (ed) Towards a steady state economy. Freeman, San Francisco
87. Hardin R (1971) Collective action as an agreeable prisons' dilemma. Behav Sci 16:472–481
88. Hardin R (1982) Collective action. Johns Hopkins University Press, Baltimore, MD
89. Hawking S, Mlodinow L (2010) The Grand Design. Random House New York
90. Henrich J et al (2004) Foundations of human sociality. Oxford University Press, Oxford
91. Henrich J et al (2005) Economic man' in cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behav Brain Sci 28:795–855
92. Hirsch M (1976) Differential topology. Springer, Berlin
93. Hitchens C (2007) God is not great. Hachette, New york
94. Hobbes T (2009 [1651]) In: Gaskin (ed) Leviathan; or the matter, forme, and power of a common-wealth, ecclesiastical and civil. Oxford University Press, Oxford
95. Hrdy SB (2011) Mothers and others: the evolutionary origins of mutual understanding. Harvard University Press, Cambridge, MA
96. Hsiang S et al (2013) Quantifying the influence of climate on human conflict. Sci Express 10:1126
97. Johnson S (2002) Emergence. Scribner, New York
98. Kahneman D (2011) Thinking fast and slow. Ferrar Strauss and Giroux, New York
99. Karklins R, Petersen R (1993) Decision calculus of protestors and regime change: Eastern Europe 1989. J Polit 55:588–614
100. Kauffman S (1993) The origins of order. Oxford University Press, Oxford
101. Keohane R (1984) After hegemony. Princeton University Press, Princeton, NJ
102. Keohane R, Nye R (1977) Power and interdependence. Little Brown, New York
103. Keynes JM (1921) Treatise on probability. Macmillan, London
104. Keynes JM (1936) The general theory of employment, interest and money. Macmillan, London
105. Kindleberger C (1973) The world in depression 1929–1939. University of California Press, Berkeley, CA

106. Knaster B, Kuratowski K, Mazurkiewicz S (1929) Ein beweis des fixpunktsatzes fur n-dimensionale simplexe. Fund Math 14:132–137
107. Konishi M (1996) Equilibrium in an abstract political economy. Social Choice and Welf 13:43–50
108. Kramer GH (1973) On a class of equilibrium conditions for majority rule. Econometrica 41:285–297
109. Kreps DM et al (1982) Rational cooperation in the finitely repeated prisoners' dilemma. J Econ Theory 27:245–252
110. Kurz M, Motolese M (2001) Endogenous uncertainty and market volatility. Econ Theory 17:497–544
111. Ladha K (1992) Condorcet's jury theorem, free speech and correlated votes. Am J Polit Sci 36:617–674
112. Ladha K (1993) Condorcet's jury theorem in the light of de Finetti's theorem: majority rule with correlated votes. Soc Choice Welf 10:69–86
113. Ladha K, Miller G (1996) Political discourse, factions and the general will: correlated voting and Condorcet's jury theorem. In: Schofield N (ed) Collective decision making. Kluwer, Boston
114. Laplace PS (1951 [1814]) Essai Philosophique sur les Probabilités. Gauthiers-Villars, Paris. A philosophical essay on probabilities (Trans. F. Truscott and F. Emory) Dover, New York
115. Li TY, Yorke JA (1975) Period three implies chaos. Math Mon 82:985–992
116. Lohmann S (1994) The dynamics of information cascades. World Polit 47:42–101
117. Lorenz EN (1962) The statistical prediction of solutions of dynamical equations. In: Proceedings of the international symposium om numerical weather prediction, Tokyo
118. Lorenz EN (1963) Deterministic non periodic flow. J Atmos Sci 20:130–141
119. Lorenz EN (1993) The essence of chaos. University of Washington Press, Seattle
120. Madison J (1999[1787]) Federalist X. In: Rakove J (ed) Madison: writings. Library Classics, New York
121. Mandelbrot B, Hudson R (2004) The (mis)behavior of markets. Perseus, New York
122. Margolis H (1982) Selfishness, altruism and rationality. Cambridge University Press, Cambridge
123. Margulis L, Sagan D (2002) Acquiring genomes. Basic, New York
124. Maynard Smith J (1982) Evolution and the theory of games . Cambridge University Press, Cambridge
125. McKelvey RD (1976) Intransitivities in multidimensional voting models and some implications for agenda control. J Econ Theory 12:472–482
126. McKelvey RD (1979) General conditions for global intransitivities in formal voting models. Econometrica 47:1085–1112
127. McKelvey RD, Page T (1986) Common knowledge, consensus and affregrate information. Econometrica 54:109—127
128. McKelvey RD, Schofield N (1987) Generalized symmetry conditions at a core point. Econometrica 55:923–933
129. McWhorter J (2001) The power of babel. Holt, New York
130. Merton RC (1973) Theory of rational option pricing. Bell J Econ Manag Sci 4:141–183
131. Michael E (1956) Continuous selections I. Ann Math 63:361–382
132. Miller G, Schofield N (2003) Activists and partisan realignment in the U.S. Am Polit Sci Rev 97:245–260
133. Miller G, Schofield N (2008) The transformation of the Republican and Democratic party coalitions in the United States. Perspect Polit 6:433–450
134. Milnor JW (1997) Topology from a differential viewpoint. Princeton University Press, Princeton, NJ
135. Minsky H (1975) John maynard keynes. Columbia University Press, New York
136. Minsky H (1986) Stabilizing an unstable economy. Yale University Press, New Haven
137. Mokyr J (2005) The intellectual origins of modern economic growth. J Econ Hist 65:285–351

138. Mokyr J (2010) The enlightened economy: an economic history of Britain 1700–1850. Yale University Press, New Haven
139. Mokyr J, Nye VC (2007) Distributional coalitions, the Industrial Revolution, and the origins of economic growth in Britain. South Econ J 74:50–70
140. Morris I (2010) Why the west rules. Ferrar, Strauss and Giroux, New York
141. Nagel T (2012) Mind and cosmos. Oxford University Press, Oxford
142. Nakamura K (1979) The vetoers in a simple game with ordinal preference. Int J Game Theory 8:55–61
143. North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
144. North DC, Weingast BR (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth century England. J Econ Hist 49:803–832
145. North DC, Wallis B, Weingast BR (2009) Violence and social orders: a conceptual framework for interpreting recorded human history. Cambridge University Press, Cambridge
146. Nowak M (2011) Supercooperators. Free Press, New York
147. Ormerod P (2001) Butterfly economics. Basic, New York
148. Ostrom E (1990) Governing the commons. Cambridge University Press, Cambridge
149. Pagden A (2013) The enlightenment. Random, New York
150. Pareto V (1935) The mind and society [Trattato di Sociologia Generale]. Harcourt, Brace, New York
151. Parfit D (2011) On what matters. Oxford University Press, Oxford
152. Parker G (2013) Global crisis. Yale University Press, New Haven, CT
153. Penn E (2009) A model of far-sighted voting. Am J Polit Sci 53:36–54
154. Plott CR (1967) A notion of equilibrium and its possibility under majority rule. Am Econ Rev 57:787–806
155. Pugh CC (2002) Real mathematical analysis. Springer, Berlin
156. Putnam RD, Campbell DE (2010) American grace: how religion divides and Unites Us. Simon and Schuster, New York
157. Rader T (1972) Theory of general economic equilibrium. Academic Press, New York
158. Rae D (1960) Decision Rules and Individual Values in Constitutional Choice. American Political Science Review 63:40–56
159. Richards (1990) Is strategic decisionmaking chaotic? Behavioral Science 35:219–232
160. Robson AJ, Kaplan HS (2003) The evolution of human life expectancy and intelligence in hunter-gatherer economies. American Economic Review 93:150–169
161. Rogow (1986) Thomas Hobbes Norton. New York
162. Saari D (1985) Price dynamics, social choice, voting methods, probability and chaos. In: Aliprantis D, Burkenshaw O, Rothman NJ (eds) Lecture Notes in Economics and Mathematical Systems, vol 244. Springer, Berlin
163. Saari D (1985) A chaotic exploration of aggregation paradoxes. SIAM Rev 37:37–52
164. Saari, D (1995) Mathematical complexity of simple economics. Notes Am Math Soc 42:222–230
165. Saari D (1997) The generic existence of a core for q-rules. Econ Theory 9:219–260
166. Saari D (2001) Decisions and elections: explaining the unexpected. Cambridge University Press, Cambridge
167. Saari D (2001) Chaotic elections. American Mathematical Society, Providence, RI
168. Saari D (2008) Disposing dictators, demystifying voting paradoxes. Cambridge University Press, Cambridge
169. Schofield N (1972) Is majority rule special? In: Niemi RG, Weisberg HF (eds) Probability models of collective decision-making. Charles E. Merrill Publishing Co, Columbus, OH
170. Schofield N (1972) Ethical decision rules for uncertain voters. Br J Polit Sci 2:193–207
171. Schofield N (1975) A game theoretic analysis of Olson's game of collective action. J Confl Resolut 19:441–461
172. Schofield N (1977) The logic of catastrophe. Hum Ecol 5:261–271
173. Schofield N (1978) Instability of simple dynamic games. Rev Econ Stud 45:575–594

174. Schofield N (1979) Rationality or chaos in social choice. Greek Econ Rev 1:61–76
175. Schofield N (1980) Generic properties of simple Bergson-Samuelson welfare functions. J Math Econ 7:175–192
176. Schofield N (1980) Catastrophe theory and dynamic games. Qual Quant 14:529–545
177. Schofield N (1983) Equilibria in simple dynamic games. In: Pattanaik P, Salles M (eds) Social choice and welfare, pp 269–284. North Holland, Amsterdam
178. Schofield N (1983) Generic instability of majority rule. Rev Econ Stud 50:695–705
179. Schofield N (1984) Existence of equilibrium on a manifold. Math Oper Res 9:545–557
180. Schofield N (1984) Social equilibrium and cycles on compact sets. J Econ Theory 33:59–71
181. Schofield N (1984) Classification theorem for smooth social choice on a manifold. Soc Choice Welf 1:187–210
182. Schofield N (1985) Anarchy, altruism and cooperation. Soc Choice Welf 2:207–219
183. Schofield N, Tovey C (1992) Probability and convergence for supra-majority rule with Euclidean preferences. Math Comput Model 16:41–58
184. Schofield N (1999) The heart and the uncovered set. J Econ Suppl 8:79–113
185. Schofield N (1999) A smooth social choice method of preference aggregation. In: Wooders M (ed) Topics in mathematical economics and game theory: essays in honor of R. Aumann. Fields Institute, American Mathematical Society, Providence, RI
186. Schofield N (1999) The C^1-topology on the space of smooth preferences. Soc Choice Welf 16:445–470
187. Schofield N (2002) Evolution of the constitution. Br J Polit Sci 32:1–20
188. Schofield N (2003) Mathematical methods in economics and social choice. Springer, Berlin
189. Schofield N (2006) Architects of political change. Cambridge University Press, Cambridge
190. Schofield N (2007) The mean voter theorem: necessary and sufficient conditions for convergent equilibrium. Rev Econ Stud 74:965–980
191. Schofield N (2010) Social orders. Soc Choice Welf 34:503–536
192. Schofield N (2011) Is the political economy stable or chaotic? Czech Econ Rev 5:76–93
193. Schofield N, Gallego M (2011) Leadership or chaos. Springer, Berlin
194. Schweitzer F et al (2009) Economic networks: the new challenges. Science 325:422–425
195. Shafer W, Sonnenschein H (1975) Equilibrium in abstract economies without ordered preferences. J Math Econ 2:245–248
196. Shiller R (2003) The new financial order. Princeton University Press, Princeton, NJ
197. Shiller R (2005) Irrational exuberance. Princeton University Press, Princeton, NJ
198. Smale S (1966) Structurally stable systems are not dense. Am J Math 88:491–496
199. Smale S (1974) Global analysis and economics IIA: extension of a theorem of Debreu. J Math Econ 1:1–14.
200. Smale S (1974) Global analysis and economics IV: finiteness and stability of equilibria with general consumption sets and production. J Math Econ 1:119–127
201. Smith A (1984 [1759]) The theory of moral sentiments. Liberty Fund, Indianapolis, IN
202. Sperber J (2011). Karl Marx: a nineteenth century life. Liveright, New York
203. Smolin L (2007) The trouble with physics. Houghton Mifflin, New York
204. Strnad J (1985) The structure of continuous-valued neutral monotonic social functions. Soc Choice Welf 2:181–195
205. Strogatz S (2004) Sync. Hyperion, New York
206. Stern N (2007) The economics of climate change. Cambridge University Press, Cambridge
207. Stern N (2009) The global deal. Public Affairs, New York
208. Stringer C (2012) Lone Survivors. Macmillan, London
209. Sunstein CR (2006) Infotopia. Oxford University Press, Oxford
210. Sunstein CR (2009) A constitution of many minds. Princeton University Press, Princeton NJ
211. Sunstein CR (2011) Going to extremes. Oxford University Press, Oxford
212. Surowiecki J (2005) The wisdom of crowds. Anchor, New York
213. Taleb NN (2007) The black swan. Random, New York
214. Taleb NN, Blyth M (2011) The black swan of Cairo. Foreign Affairs 90(3):33–39
215. Taylor M (1976) Anarchy and cooperation. Wiley, London

216. Taylor M (1982) Community, anarchy and liberty. Cambridge University Press, Cambridge
217. Taylor C (2007) A secular age. Belknap Press, Cambridge, MA
218. Trivers R (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35–56
219. Trivers R (1985) Social evolution. Cummings, Menlo Park, CA
220. Turing A (1937) On computable numbers with an application to the entscheidungs problem. Proc Lond Math Soc 42:230–265. Reprinted in Jack Copeland (ed) The essential turing. The Clarendon Press, Oxford
221. Walker M (1977) On the existence of maximal elements. J Econ Theory 16:470–474
222. Watts D (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci 99:5766–5771
223. Watts D (2003) Six degrees. Norton, New York
224. Weber M (1904) The Protestant Ethic and the spirit of capitalism. Reprinted in 1997 by Routledge, London
225. Weitzman M (2009) Additive damages, fat-tailed climate dynamics, and uncertain discounting. Economics 3:1–22
226. White TD et al (2009) Ardipithicus ramidus and the paleobiology of early hominids Science 326:64–86
227. Wright R (2009) The moral animal Vintage. New York
228. Zeeman EC (1977) Catastrophe theory: selected papers, 1972–1977. Addison Wesley, New York
229. Zeeman EC (1992) Evolution and Catstrophe Theory. In: Bourriau (ed) Understanding catastrophe. Cambridge University Press, Cambridge
230. Zhang DD et al (2007) Global climate change, war, and population decline in recent human history. Proc Natl Acad Sci 104(49):19214–19219

Relevant Irrelevance: The Relevance of Independence of Irrelevant Alternatives in Family Bargaining

Elisabeth Gugl

Abstract Introducing production to a family bargaining model immediately sets the stage for the axiom of Independence of Irrelevant Alternative (IIA). Requiring that bargaining solutions satisfy IIA rules out the Kalai-Smorodinsky solution, but the broad class of Generalized Utilitarian bargaining solutions satisfies this axiom. I show that In the case of utility profiles that lead to almost transferable utility, IIA has no bite because the utility possibility frontier before and after production coincide. Almost TU is an important subdomain of all utility profiles and much broader than transferable utility, but it is still restrictive. Hence IIA is a desirable axiom of family bargaining solutions. I focus on bargaining within the family but the argument for IIA as a relevant property of bargaining solutions applies to other bargaining problems as well in which goods are produced or second period renegotiation takes place.

Keywords Almost transferable utility • Axiomatic bargaining • Bargaining with production • Models of family bargaining

1 Introduction

As a student of Economics at the Karl-Franzens Universität in Graz I was originally more drawn to Macroeconomics than Microeconomics. However, that changed when I took Nick Baigent's courses and when I became familiar with Nick's research. I fell in love with Social Choice which Nick taught as the first of three topics in his Public Economics sequence. Inspired by Nick's own research I developed a keen interest in family economics while I was still at Graz and it was his encouragement that lead me to pursue a Ph.D. in the United States. I have to thank many people for helping me find my own path in economics, but Nick first ignited the spark to become an academic and my interest in axiomatic social choice.

E. Gugl (✉)
Department of Economics, University of Victoria, P.O. Box 1700, STN CSC, Victoria, BC, Canada V8W 2Y2
e-mail: egugl@uvic.ca

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*, Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_13

This paper stresses the importance of the axiom of independence of irrelevant alternatives in family bargaining models. Independence of Irrelevant Alternatives (IIA) states that if a bargaining solution (BS) picks a point in the utility possibility set that is still available when some of the previously feasible points are removed, the bargaining solution applied to the new smaller set must again pick the point that was selected in the larger set. While IIA is not always persuasive and has been dropped in favor of other axioms—as, for example, in the case of the Kalai-Smorodinsky solution (KSS)[1]—, here I discuss its appeal in the context of household decision making.

Early models of family bargaining followed the traditional setup of a bargaining problem [6, 13]. That is, given amounts of goods were to be distributed efficiently among household members.[2] However, in order to share goods they need to be produced with the inputs (most importantly the time) of the spouses, and hence the question arises whether bargaining takes place before or after production. It is not straightforward to answer this question. Different time lines have been proposed but the most common setup is that husband and wife pool their resources and settle on a distribution before they have produced anything (pre-production bargaining).[3] Then they take the actions that lead to the product mix that allows them to achieve this distribution of final utility gains. However, spouses might run into a problem when they take stock of what they have actually produced; at this stage, because they face a fixed product mix, their utility possibility set will typically be smaller than the utility possibility set considered before any action was taken.

What should spouses do if they realize that by applying the same bargaining solution to the post-production utility possibility set (UPS) it would imply a different distribution of utility gains? With a bargaining solution satisfying IIA this puzzle does not occur; once actions are taken the post-production UPS is a subset of the pre-production UPS, and the upper boundaries of both the pre- and post-production UPSs, the utility possibility frontiers (UPFs), are tangent to each other at the point on the pre-production UPF that was selected with the pre-production UPS. Hence if a bargaining solution satisfies IIA, applying the same bargaining solution to the post-production UPS will yield exactly the same result.

I focus on the Kalai-Smorodinsky solution (KSS)[4] as well as generalized utilitarian bargaining solutions (GUBS). KSS equalizes relative gains from cooperation, while the latter class of solutions consists of maximizing an additively separable function in agents' utility gains from cooperation and includes the Utilitarian and

[1] See Kalai and Smorodinsky [6].

[2] See e.g. [9–11].

[3] Another common setup is bargaining after goods have been produced, and spouses decide individually how much they contribute to production. In these models spouses' actions change the utility possibility set and the disagreement point. Spouses' strategic interaction leads to inefficient decisions. See e.g. [7]. Gugl [3] provides a discussion of these models.

[4] Manser and Brown in their seminal paper in 1980 considered both the Nash Bargaining Solution (NBS) and KSS, but already early on the NBS got more attention [11].

Nash bargaining solutions. The disagreement utility of each spouse is given by how much each of them can guarantee him or herself either when single, or in divorce, or in non-cooperative marriage.[5,6] Any GUBS satisfies IIA, while the KSS does not.[7]

In the case where there is no distinction between the pre- and post-production UPS, IIA has no bite; any bargaining solution will yield the same result no matter whether it's applied before or after production decisions have been made.[8] Drawing on results by Gugl and Leroux [4], I point out in this paper that there is no distinction between the pre- and post-production UPS when the utility functions of the spouses satisfy Almost Transferable Utility.

The model follows the basic set-up and notation of Gugl and Leroux [4], but the question asked is different. The authors ask when GUBS satisfy the solidarity property, that is, when a change in the UPS due to a change in production possibilities without changing the disagreement point of the bargaining problem will lead to either all agents gaining from the change in the UPS or all agents losing. Here the question is whether the same point on the UPF is picked when irrelevant alternatives (points not picked before) are removed. Both the present paper and Gugl and Leroux have an important feature in common; in both papers the focus is on how changes in the utility possibility set due to changes in the available product mixes impact distribution while typically in the family bargaining literature authors are concerned with changes in intrafamily distribution due to changes in the disagreement point [9].

For the sake of clarity, I restrict the number of household members to two, interpreting them as husband and wife. This is also the most common assumption in family bargaining models and it allows me to draw graphs illustrating the results discussed in this paper. The next section introduces the model and applies basic results in axiomatic bargaining theory to the specific model of family bargaining. The third section introduces Almost Transferable Utility (ATU) and revisits the role of IIA in family bargaining models with ATU. The fourth section relates the results presented here to family bargaining models with two periods and renegotiation and concludes by highlighting the advantages of GUBS over KSS in family bargaining.

While the focus of the paper is on providing a justification for imposing IIA in family bargaining models, it should be stressed that the same argument carries over to other situations in which two or more people need to produce the goods that they will later share.

---

[5] See e.g. [9–11].

[6] The latter types of disagreement points often lead to inefficient production or distribution of goods in marriage once we allow for more than one period. See e.g. [3, 8] for two-period models with renegotiation failing to achieve efficiency. Gugl and Welling [5] show that efficiency in two-period bargaining models with renegotiation is achieved if renegotiation does not interfere with the equalization of spouses's marginal rates of intertemporal substitution.

[7] See e.g. [12].

[8] A similar result is obtained if one considers family bargaining with renegotiation [5].

2 The Model and Basic Results

The model follows very closely the basic set-up and notation of Gugl and Leroux [4]. Two agents, husband and wife, denoted by $i = 1, 2$, produce a set $M = \{1, 2, 3, \ldots m\}$ of goods and $|M| \geq 2$. The set of goods is partitioned into a subset M_1 of private goods, and a subset M_2 of public goods. Thus $M = M_1 \cup M_2$ and $M_1 \cap M_2 = \emptyset$. A product mix, $y \in \mathbb{R}_+^M$ may be correspondingly partitioned into $\left(y^1, y^2\right)$. Denote by Z_i the action set of an agent and by $z_i \in Z_i$ a specific action taken by agent i. Each vector $z = (z_1, z_2) \in Z_1 \times Z_2$ generates a product mix $y(z)$. We assume that agents take an action vector z that maximizes an agreed-upon objective function. In most of the subsequent analysis we can focus on y and in the notation ignore z. Let $Y \subset \mathbb{R}_+^M$ be the set of all feasible product mixes given $Z_1 \times Z_2$. The production possibility set Y is assumed to be a closed, convex, and comprehensive set. Let $\mathcal{Y}$ be the class of all such production possibility sets. Denote by ∂Y the production possibility frontier of Y and let $F : \mathbb{R}_+^M \to \mathbb{R}$ be the corresponding transformation function:

$$Y = \left\{y \in \mathbb{R}_+^M | F(y) \leq 0\right\}, \text{ and}$$
$$\partial Y = \{y \in Y | F(y) = 0\}.$$

We denote by $x_i \in \mathbb{R}_+^M$ agent i's consumption vector, which again can be partitioned into $\left(x_i^1, x_i^2\right)$. A *distribution of* y is a pair of consumption vectors, one per agent, $x = (x_1, x_2) \in \mathbb{R}_+^M \times \mathbb{R}_+^M$ such that:

$$\begin{cases} x_{1k} + x_{2k} = y_k \text{ for any } k \in M_1 \text{ and} \\ x_{1l} = x_{2l} = y_l \quad \text{for any } l \in M_2. \end{cases}$$

For any $Y \in \mathcal{Y}$ and any product mix $y \in Y$, we denote by $X(y)$ the *set of distributions of* y and by $X(Y) = \bigcup_{y \in Y} X(y)$ the set of *feasible* distributions under Y. An *allocation* is a product mix-distribution pair $(y, x) \in Y \times X(y)$.

The preferences of each agent i are represented by a *utility function*, $u_i : \mathbb{R}_+^M \to \mathbb{R}$, which is strictly increasing, concave, differentiable, and satisfies $u_i(0) = 0$**.** We denote by $\mathcal{U}$ the class of such utility functions. A *utility profile* is a pair of utility functions, $(u_1, u_2) \in \mathcal{U}^2$, one per agent.

For any $(Y, u) \in \mathcal{Y} \times \mathcal{U}^2$, we denote by $U(Y, u) = \{\psi \in \mathbb{R}_+^2 | \psi = u(x)$ for some $x \in X(Y)\}$ the *utility possibility set* corresponding to (Y, u). It follows from our assumptions that $U(Y, u)$ is a closed, convex, and comprehensive set. We denote by $\partial U(Y, u)$ the Pareto frontier of $U(Y, u)$; i.e., $\partial U(Y, u) = \{\psi \in U(Y, u) | \psi' \geq \psi \implies \psi' \notin U(Y, u)\}$.[9]

[9] We adopt the following notation for vector inequalities:

- $x \geqq x'$ if $x_i \geq x_i'$ for all i;
- $x \geq x'$ if $x \geqq x'$ and $x \neq x'$;

Husband and wife settle on the distribution of produced goods in the form of a bargaining process. We denote by $d_i \geq 0$ agent i's stand-alone utility level and call $d = (d_1, d_2) \in \mathbb{R}^2_+$ the *disagreement point* of the bargaining process. For any $u \in \mathcal{U}^2$ and $d \in \mathbb{R}^2_+$, define $\mathcal{Y}(u,d) = \{Y \in \mathcal{Y} | d \in U(Y,u)\}$ the set of all production possibility sets for which d is achievable. Fixing $u \in U$ and $d \in \mathbb{R}^2_+$, we say that a pair $(U(Y,u), d)$ is a bargaining problem with associated u and d whenever $Y \in \mathcal{Y}(u,d)$. We denote by $\mathcal{B}(u,d)$ the class of bargaining problems associated with u and d. Finally, we define $\mathcal{B}(u) = \cup_{d \in \mathbb{R}^2_+} \mathcal{B}(u,d)$ as the set of all possible bargaining problems**.**

Note that the utility profile of spouses is assumed to be fixed. We assume that spouses know each other well and that their cardinal utility functions do not change between the production of goods and their distribution. For the question of whether a change in the feasibility of product mixes changes spouses' perception of what is the optimal distribution of utility gains, there is no need to contemplate the question of how the optimal distribution of utility gains changes if a spouse's utility function changes. Hence the focus on $\mathcal{B}(u)$ in both the definition of a bargaining solution and the definition of Independence of Irrelevant Alternatives below.

Definition 1 A bargaining solution is a function S defined on $\mathcal{B}(u)$, which associates with every bargaining problem $(U(Y,u), d) \in \mathcal{B}(u)$ a utility vector $S(U(Y,u),d) \in \partial U(Y,u)$ such that $S(U(Y,u),d) \geqq d$**.**

Observation 1 In order to implement $S(U(Y,u),d) \in \partial U(Y,u)$ agents are required to produce a particular y and hence choose action vector z.

Definition 2 A bargaining solution S satisfies *Independence of Irrelevant Alternatives (or IIA)* if, for any $\mathcal{B}(u,d) \subset \mathcal{B}(u)$ the following holds. For any $Y, Y' \in \mathcal{Y}(u,d)$ such that $U(Y',u) \subset U(Y,u)$ and $S(U(Y,u),d) \in U(Y',u)$

$$S(U(Y',u),d) = S(U(Y,u),d).$$

Observation 2 In the context of our model, before actions are chosen agents face a production possibility set Y and hence a utility possibility set $U(Y,u)$. Once agents have executed their actions, the production possibility set is reduced to a single product mix, that is $Y' = y(z)$, and spouses face $U(y(z),u)$. If agents choose an action based on $S(U(Y,u),d)$, then $U(Y',u) \subset U(Y,u)$ and $S(U(Y,u),d) \in U(Y',u)$.

Thus considering agents before and after executing their actions sets the stage for IIA.

Definition 3 A bargaining solution S belongs to the class of Generalized Utilitarian Bargaining Solutions (GUBS)—denoted by $\mathcal{G}$—if there exists a list of concave,

---

– $x > x'$ if $x_i > x'_i$ for all i.

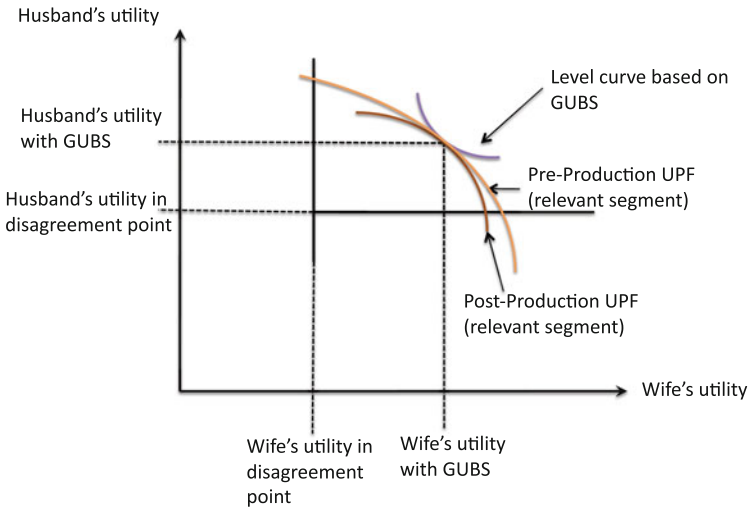


Fig. 1 Any GUBS satisfies IIA

strictly increasing, and continuous functions of $\mathbb{R}$, (γ_1, γ_2), such that

$$S(U(Y,u),d) \in \arg \max_{\psi \in \partial U(Y,u)} \gamma_1(\psi_1 - d_1) + \gamma_2(\psi_2 - d_2).$$

Note that $\mathcal{G}$ defines a broad family of bargaining solutions of which the Nash bargaining solution is a member with $\gamma_i(\psi_i - d_i) = \ln(\psi_i - d_i)$ for all $i = 1, 2$. The solution to the maximization problem

$$\max_{\psi \in \partial U(Y,u)} \gamma_1(\psi_1 - d_1) + \gamma_2(\psi_2 - d_2)$$

may not be unique in which case some tie breaking rule will have to be applied.[10] However, the solution to the maximization problem is unique if the Nash bargaining solution is applied or any other GUBS for which γ_i (.) is strictly concave (see e.g. [12]).

Observation 3 Any $S \in \mathcal{G}$ satisfies IIA.

This is a well known result.[11] Figure 1 illustrates the intuition.

[10] In particular, the class of weighted utilitarian bargaining solutions (WUBS), a subclass of GUBS, consists of bargaining solutions, W, characterized by a list of non-negative weights, $w = (w_1, w_2) \in \mathbb{R}^2_+$, with $\sum w_i = 1$, such that $W(U(Y,u),d) \in \arg\max_{\psi \in \partial U(Y,u)} \sum_{i \in N} w_i(\psi_i - d_i)$. In the case of WUBS the solution to the maximization problem may not be unique.

[11] See e.g. [12, p. 67].

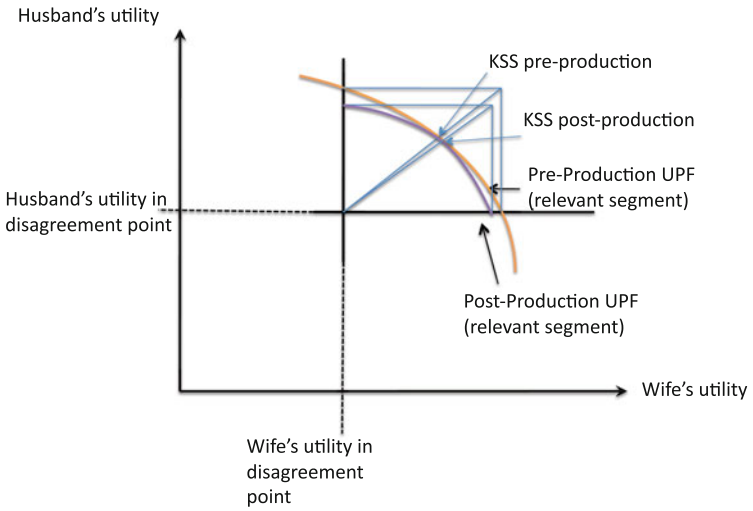


Fig. 2 KSS violates IIA

Next we turn to the Kalai-Smorodinsky solution. As an intermediate step, we define maximum utility gains for each spouse.

The maximum utility of spouse i is given by

$$\overline{\psi}_i = \max_{\psi \in \partial U(Y,u)} \psi_i \text{ s.t. } \psi_j \geq d_j$$

and hence the maximum utility gain is given by $\overline{\psi}_i - d_i$.

Definition 4 The Kalai-Smorodinsky solution, $KS(U(Y,u),d)$ selects the unique $\psi \in \partial U(Y,u)$ that equalizes relative gains between spouses

$$\frac{\psi_1 - d_1}{\overline{\psi}_1 - d_1} = \frac{\psi_2 - d_2}{\overline{\psi}_2 - d_2}.$$

Observation 4 $KS(U(Y,u),d)$ does not satisfy IIA.

This is obvious from the axioms characterizing KSS [6]. Figure 2 illustrates the intuition.

2.1 Discussion

The results presented in this section are well known in Axiomatic Bargaining. The contribution of this section lies in the application of axiomatic bargaining to family economics. IIA becomes immediately relevant because families represent

a miniature economy with production. Typically the utility possibility set before production takes place is larger than after production has occurred. Spouses run into a consistency problem if the distribution of goods that is optimal with the pre-production utility possibility set differs from the optimal distribution of goods with the post-production utility possibility set using the same bargaining rule. If the solution differs depending on to which set the bargaining rule is applied, there is the potential for one spouse arguing to distribute gains according to the application of the rule post production, but the other spouse might argue for the application to the pre-production utility possibility set. Bargaining rules that satisfy IIA help avoid this conflict between spouses. Another way the dilemma can be avoided is if there is no difference between the pre- and post-production utility possibility sets. The next section considers this case.

3 Equivalence of Pre- and Post-production UPS

In this section I define the concept of Almost Transferable Utility (ATU) introduced by Gugl and Leroux [4]. In order to do so, we first recall what is meant by a product mix being *efficient independent of distribution*, a necessary (and sufficient) condition for ATU to hold.

Definition 5 A product mix $\overline{y} \in Y$ is *efficient independent of distribution* if $\partial U(Y,u) \subset \{u(x)|x \in X(\bar{y})\}$.[12]

That is, any point on the utility possibility frontier is the result of various distributions of the same product mix.

Proposition 1 *If $y(z^*)$ is efficient independent of distribution, then $S(U(Y,u),d) = S(U(y(z^*),u),d)$.*

Proof For any bargaining solution S, $S(U(Y,u),d) \in \partial U(Y,u)$. Hence $S(U(Y,u),d)$ is associated with a particular efficient product mix y and action vector z^*. If $y(z^*)$ is efficient independent of distribution, then $\partial U(Y,u) \subset \{u(x)|x \in X(y(z^*))\}$. This implies $\partial U(y(z^*),u) = \partial U(Y,u)$.[13] Since pre- and post-production UPS coincide, any bargaining solution will result in the same distribution of utility gains whether it is applied to the pre- or post-production UPS, that is, $S(U(Y,u),d) = S(U(y(z^*),u),d)$.

If $y(z^*)$ is efficient independent of distribution, the issue that raises the question of whether or not IIA is a desirable property of the bargaining solution is entirely avoided. Thus we can state the following result.

[12] Bergstrom and Cornes [2] call this concept "independence of allocative efficiency from distribution."

[13] Whether other actions and hence other $y(z)$ would also lead to Pareto efficiency is irrelevant as the actions chosen are already allowing us to reach any point on the original UPF.

Corollary 1 *If $y(z^*)$ is efficient independent of distribution, for any GUBS and the KSS, $S(U(Y,u),d) = S(U(y(z^*),u),d)$.*

Whether a product mix is efficient independent of distribution, depends on the utility profile of spouses. Gugl and Leroux [4] show that if a product mix is to be efficient independent of distribution and the UPS is to be convex, the utility profile of spouses must satisfy almost transferable utility. The converse is also true; utility profiles satisfying ATU imply that an efficient product mix independent of distribution exists and that the UPS is convex. Hence Proposition 1 can be restated in terms of ATU.

Definition 6 The profile $u \in \mathcal{U}^2$ exhibits *Almost Transferable Utility (ATU)* if for any given $Y \in \mathcal{Y}$ there exists a pair of positive monotonic transformations f_i, such that $f = (f_1, f_2)$ with $f_i : \mathbb{R} \to \mathbb{R}$, and $\lambda \in \mathbb{R}_+$ such that the utility possibility frontier takes on the following form:

$$\partial U(Y,u) = \{\psi \in U(Y,u) : f_1(\psi_1) + f_2(\psi_2) = \lambda\}.$$

We denote by $\mathcal{ATU} \subset \mathcal{U}^2$ the class of utility profiles satisfying ATU. Because λ depends on the production possibility set Y and on the utility profile u, we denote it by $\lambda(Y,u)$.

Proposition 1 reformulated Let z^* denote the optimal action pair to implement $S(U(Y,u),d)$. If $u \in \mathcal{ATU}$, then $S(U(Y,u),d) = S(U(y(z^*),u),d)$.

Next I turn to the question of which utility profiles lead to ATU using two illustrative examples.

Example 1 Examples of utility profiles satisfying ATU.

a) (Inspired by Gugl and Leroux [4]) Let $0 < \alpha < 1$, $0 < \beta_i \leq 1$ and suppose $u_1 = \left(x_{11}^{\alpha} x_{21}^{1-\alpha}\right)^{\beta_1}$ and $u_2 = \left(x_{12}^{\alpha} x_{22}^{1-\alpha}\right)^{\beta_2}$. Then $\partial U(Y,u) = \{(\psi_1,\psi_2) \in \mathbb{R}_+^2 : \psi_1^{1/\beta_1} + \psi_2^{1/\beta_2} = \lambda(Y,u)\}$, where $\lambda(Y,u) = \max_{y \in Y}\left(y_1^{\alpha} y_2^{1-\alpha}\right)$ and $f_i(\psi_i) = \psi_i^{1/\beta_i}$.

b) (Gugl and Leroux [4]) Suppose the cardinal utility function of agent i over a private good (x_{1i}) and a public good (x_2) is given by $u_i = (x_{1i} + h_i(x_2))^{\delta_i}$, where h_i is a strictly concave function, $\lim_{x_i \to 0} h_i'(x_i) = \infty$ and $0 < \delta_i \leq 1$. Then the segment of the utility possibility frontier at which ATU holds is equal to the set

$$\left\{\psi \in \mathbb{R}_+^2 | \psi_i \in \left[h_i\left(y_2^*\right)^{\delta_i}, \left(y_1^* + h_i\left(y_2^*\right)\right)^{\delta_i}\right] \text{ for every } i \in N \text{ and } \psi_1^{1/\delta_1} + \psi_2^{1/\delta_2} = \lambda(Y,u)\right\} \text{ where } \left(y_1^*, y_2^*\right) = \max_{y \in Y} y_1 + h_1(y_2) + h_2(y_2), \lambda(Y,u) = \max_{y \in Y} y_1 + h_1(y_2) + h_2(y_2) \text{ and } f_i(\psi_i) = \psi_i^{1/\delta_i}.$$

Note that *Transferable Utility* is a special case of ATU with $f_i(\psi_i) = \psi_i$ [1]. To obtain TU in the examples above, we need $\beta_1 = \beta_2 = 1$ in a) and $\delta_1 = \delta_2 = 1$ in b). The efficient product mix is the same independent of the values of β_i in Example 1a)

and δ_i in Example 1b). Note that any profile of utility functions that leads to TU in these examples also satisfies ATU. Gugl and Leroux [4] demonstrate that "any utility profile satisfying TU can be transformed into many utility profiles satisfying Almost TU by using different concave transformations on the utility profile satisfying TU. Thus for any utility profile leading to ATU there is a utility profile leading to TU that has the same ordinal properties but different cardinal properties." [4, p. 137] Changing the cardinal properties of the individuals' utility functions leads to different distributions of the private goods under the GUBS and KSS, and hence the domain of ATU is significantly larger than the domain of TU.

4 Conclusion

The assumption that spouses face a fixed product mix, and that the only decision they have to make is how to distribute this product mix is unrealistic. However, introducing production to a bargaining model immediately sets the stage for IIA. IIA in family bargaining models is relevant if utility profiles do not lead to almost transferable utility. Almost TU is an important subdomain of all utility profiles, but it is still restrictive and hence IIA is a desirable axiom of family bargaining solutions. Requiring that bargaining solutions satisfy IIA rules out the Kalai-Smorodinsky solution, but the broad class of Generalized Utilitarian bargaining solutions satisfies this axiom. I focussed on bargaining within the family but the argument for IIA as a relevant property of bargaining solutions applies to other bargaining problems as well in which goods are produced or second-period renegotiation takes place.

In their two-period bargaining model with renegotiation Gugl and Welling [5] also find an important role of IIA. Imagine that we add a second period to our bargaining model and spouses anticipate renegotiation in the second period. In the first period, they still consider intertemporal utility gains in deciding which actions to take, but with the constraint that second-period utility gains will be renegotiated. This renegotiation puts again a constraint on the intertemporal UPS that is being considered in the bargaining problem in the first period, but the constraint is of a different nature than the restriction we considered above. Suppose public policy can enforce a division of second-period utility gains that would result in the same distribution as without renegotiation. In this case the restriction of feasible distributions in the second period would shrink the intertemporal utility possibility set but this set would share one point on the intertemporal utility possibility frontier. This common point is the optimal distribution of intertemporal utility gains when bargaining without renegotiation takes place. Given this relationship between the UPS with and without renegotiation, IIA leads to efficiency. This setup bears a resemblance to the bargaining problem analyzed in Sect. 2. Gugl and Welling [5] show that such a law (however unrealistic) would achieve efficiency if spouses are Nash-Bargainers. IIA guarantees efficiency with renegotiation. With KSS, however, the same policy applied to second-period distribution of utility gains does not guarantee efficiency.

While it is ultimately questionable whether such tailored policies are feasible (it may be possible for spouses to write prenuptial agreements that are tailored to their specific case, but not for a government), this is another illustration of the importance of IIA in family bargaining models and provides an overlooked justification for the prominence of the Nash Bargaining solution in family economics.

References

1. Bergstrom ThC (1989) A fresh look at the rotten kid theorem – and other household mysteries. J Polit Econ 97:1138–1159
2. Bergstrom ThC, Cornes RC (1983) Independence of allocative efficiency from distribution in the theory of public goods. Econometrica 51:1753–1766
3. Gugl E (2009) Income splitting, specialization, and intra-family distribution. Can J Econ 42:1050–1071
4. Gugl E, Leroux J (2011) Share the gain, share the pain? Almost Transferable Utility, changes in production possibilities, and bargaining solutions. Math Soc Sci 62:133–143
5. Gugl E, Welling L (2013) Prenuptial agreements and efficiency in marriage. Working paper, University of Victoria
6. Kalai E, Smorodinsky M (1975) Other solutions to Nash's bargaining problem. Econometrica 43:513–518
7. Konrad KA, Lommerud KE (2000) The bargaining family revisited. Can J Econ 33:471–487
8. Lundberg S (2002) Limits to specialization: family policy and economic efficiency. Working paper, University of Washington
9. Lundberg SJ, Pollak RA (1993) Separate spheres bargaining and the marriage market. J Polit Econ 101(6):988–1010
10. Manser M, Brown M (1980) Marriage and household decision-making: a bargaining analysis. Int Econ Rev 21(1):31–44
11. McElroy MB, Horney MJ (1981) Nash-bargained household decisions: toward a generalization of the theory of demand. Int Econ Rev 22(2):333–349
12. Moulin H (1988) Axioms of cooperative decision making. Cambridge University Press, Cambridge
13. Nash JF Jr (1950) The bargaining problem. Econometrica 18:155–162

Part III
Social Welfare and Equilibrium

Forced Trades in a Free Market

Marc Fleurbaey

Abstract A free trade is always Pareto-improving. But some "free trades" are actually forced in the sense that they reflect the trader's poverty rather than his or her preferences. We propose a rigorous concept of forced trade, and apply it to the ethical evaluation of Walrasian equilibria.

Keywords Competitive equilibrium • Constraint • Forced trade • Poverty • Preferences

1 Introduction

> Everyone but an idiot knows that the lower classes must be kept poor or they will never be industrious.
> Arthur Young, *The Farmer's Tour through the East of England*, 1771.

The contribution of the theory of social choice to the evaluation of market allocations has been for the most part limited to exploring the general possibilities for aggregating individual preferences into a social ordering. In this mainstream context, the Pareto principle is generally sacrosanct. However, scholars like Nick Baigent have explored beyond the boundaries of this domain, questioning the basic principles of consequentialism, the possibility of carving a space for rights in social evaluation, and examined how to reconcile the social choice approach with policy concerns about merit goods.[1] This paper takes inspiration from this critical tradition.

[1]See in particular the research published in Baigent [1, 2, 4], which provides an interesting supplement to his important contribution to mainstream social choice (in particular its topological branch).

M. Fleurbaey (✉)
Princeton University, Princeton, NJ, USA
e-mail: mfleurba@princeton.edu

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_14

A few years ago, TV news exhibited the sorrow of an English mother whose daughter died of the new form of Creutzfeld-Jacob disease. Her daughter, she explained, had been fed mostly with hamburgers for years because that was the only kind of food they could afford. At the same period, a French novel made a scandal because it describes sexual tourism as an ordinary business. Between affluent, cynical Westerners and poor people from developing countries, who only own their body, it is, as the author said, 'an ideal situation of exchange'.

More recurrently a debate is going on about labor market deregulation, and opposes those who point out the new possibilities of mutually beneficial trades that deregulation entails, and those who object that workers will have no choice but to accept badly paid jobs with low guarantee and unsafe working conditions. The project of an international market for pollution permits has also aroused a debate opposing arguments about efficiency to arguments having to do with the fact that poor countries would then be induced to sell permits just because they are, regrettably, not in a position to pollute themselves.

The common thread in all these stories is the following. Although voluntary trade is always mutually beneficial, some trades are less "voluntary" than others, and economic pressure may be such as to make some trades questionable. If two agents engage in a trade, but one of them accepts it only because of a relatively disadvantaged position, it looks like the Pareto-improvement obtained through that trade is a step in the wrong direction. In particular, the surplus obtained by the relatively advantaged agent seems questionable. This agent is only exploiting the relative disadvantage of his trade partner. This is not to say that the disadvantaged agent is not actually benefiting also from the trade. But his benefit is conditional on his relative disadvantage, which is what makes it worrisome.

This problem has long been recognized by the law, which stipulates, with a lot of variation across countries, that contracts accepted under conditions of economic duress have no legal validity. The notion of economic duress that is retained by the law has, however, tended to be quite restrictive in general. But state regulation has added many safeguards and prohibitions surrounding working hours, working conditions, minimum wages, organs and blood, surrogate mothers, prostitution, etc. which all have to do with the risk that without regulation many agents would accept the unacceptable just because they are poor. The issue is not only one of paternalism against dangerous preferences (those who like dangerous work, for instance) or of externalities (bad health entails many negative externalities), but, above all, that the poor must be protected against the consequences of poverty which have to do with their excessive willingness to enter bad contracts.

Be that as it may, economic theory has no formal concept for this problem. Because any voluntary trade is Pareto-improving, and a Pareto-improvement is close to being sacro-sanct in welfare economics, there is no way in which the currently available concepts may help discern a problem when an agent is only apparently free, and is actually forced to accept a trade by economic pressure. When markets are complete and competition is perfect, the only ethical problem that economic

theory acknowledges is the distribution of income, when it features inequalities or poverty.

Inequalities and poverty are commonly thought to be undesirable for a variety of reasons. On egalitarian grounds, it is just bad if some have less (resources, consumption, freedom, etc.) than others. In view of sufficiency principles, it is bad if some do not have enough. In the theory of fair allocation, recently surveyed by Thomson [25], it is bad if some agents envy others (in the sense that they would rather consume the others' bundles), or if they would rather have the average bundle than their own. In [19] theory of exploitation, it is bad if some people would be better-off under an egalitarian distribution of endowments. In [18] theory of justice, it is bad if the worse-off could be better-off. None of these approaches describes or analyzes the phenomenon we want to study here.[2]

The explanation for this gap in economic theory may lie in the difficulty to disentangle the various factors which give agents incentives to trade. Agents engage in trade essentially for a mixture of three causes: (1) different tastes; (2) qualitatively different endowments; (3) quantitatively unequal endowments. Voluntary trades induced by different tastes or qualitatively different endowments should not give any qualm to the ethical observer, but trades which stem from inequalities or poverty are more problematic. And the reason they are problematic is that in such trades, the disadvantaged agent would no longer accept the trade if his disadvantage was removed.

In this paper we attempt to devise a test for the fact that an agent would no longer accept a trade or part of a bundle if his endowment was higher. Starting with examples, we will progressively elaborate general concepts which will enable the economist to distinguish how much of an agent's market demand is due to economic pressure.

The availability of such concepts should permit a more complex view of the ethical properties of the market mechanism in welfare economics. A market economy with inequalities is not just an economy with unequal welfare, or with poor households who cannot buy enough consumption goods. It is also an economy where some agents are forced to accept the unacceptable, and, as a consequence, it is an economy where there is *too much* trade of some kind, such as bad jobs and junk food. An inegalitarian economy is therefore qualitatively different from an egalitarian one. Even the rich cannot have the same way of life in an economy without poverty, because they have no poor available, if only for domestic chores and cheap arts and crafts.

[2]There is, in philosophy and in informal economics, a large literature on freedom and coercion in actions and transactions (see [6–17, 20–24, 26, 27]). It opposes those who find constraint in standard economic transactions to those who find only freedom there. But, apart from Samuelson [21] who sees the price mechanism in general as a constraint device, none is really interested in the idea that constraint is pervasive even in a competitive market.

What we set up to do in this paper is to give a precise and rigorous definition to the notion of a forced trade. We will show in particular that one may distinguish an objective notion, having to do with budget constraints and survival, and a subjective notion which involves the agent's preferences.

The paper is organized as follows. The next two sections essentially provide more intuition, in particular contexts, and propose a heuristic analysis. Section 2 presents a simple model in which the idea that some agents are forced to work more than normal is formalized. Section 3 examines another simple model, with a focus on forced consumption of inferior goods by people with low income. Section 4 extends these preliminary analyses and proposes a more abstract approach in a very general setting. Section 5 concludes.

2 Forced Labor

Consider the following simple model. There are three goods, land, labor and corn. Corn is the numeraire, and prices of land and labor are denoted r and w, respectively. A constant-returns-to-scale technology transforms land (k) and labor (ℓ) into corn (c):

$$c = f(k, \ell).$$

The population has n individuals, whose initial endowment consists of land only. They can also work. Their supply of land is inelastic, but they have preferences over corn and labor. Individual i has initial endowment (and supply) of land k_i, yielding a budget constraint

$$c_i = rk_i + w\ell_i,$$

where c_i is consumption of corn and ℓ_i is labor, supposed to vary between 0 and 1 (see Fig. 1).

Let $c_i(w, rk_i)$ and $\ell_i(w, rk_i)$ denote individual i's demand of corn and supply of labor, respectively. We assume that these demand and supply functions are determined by maximization of preference satisfaction under individual budget constraint.

Let us assume that a Walrasian equilibrium exists. Such an equilibrium is necessarily Pareto-optimal, and this is often presented as the hallmark of a free market. But are agents really free to trade in such a context?

Obviously, individuals face a budget constraint, and the smaller the budget the less options they have. This is trivial, and one might think that the only meaningful exercise, in this respect, is measuring the size of the opportunity sets of agents. A

Fig. 1 Budget set

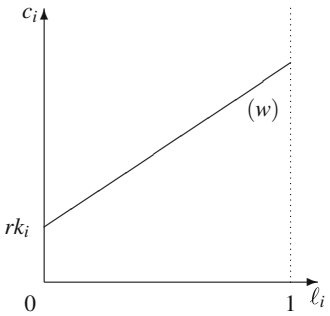
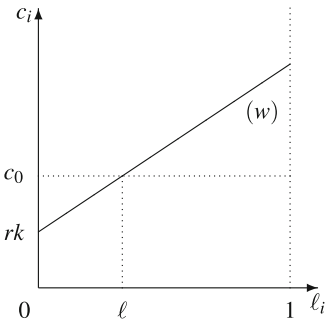


Fig. 2 Objective constraint



poor agent has less options than a rich agent, and this apparently raises a problem of inequality only. But, as suggested in the introduction, there is more to it. Poverty does not only reduce the available options, it also puts pressure on accepting some trades. A poor may accept a trade that a rich with similar preferences would refuse.

In order to make sense of this idea, one must first give a rigorous definition of the economic pressure due to poverty.

There is, first, an objective sense in which agents can be considered as constrained. Suppose that a decent (or subsistence) level of consumption is c_0. Then an agent is objectively forced to work if $rk_i < c_0$. More precisely, consider any $k > 0$ and $\ell > 0$ satisfying:

$$c_0 = rk + w\ell.$$

The agent *is objectively forced to work at least* ℓ*whenever his endowment is less than* k (see Fig. 2).

With this simple definition, one can ask the following questions. First, supposing that a given amount of labor $\ell^* \in (0, 1)$ is taken as a reference,one can identify the

agents who are objectively forced to work more than ℓ^*, and they are those whose endowment is less than

$$OLE(\ell^*) = \frac{c_0 - w\ell^*}{r},$$

which will be called the "objective liberation endowment", that is, the endowment which releases the agent from the obligation to work at least ℓ^* in order to attain c_0. The reference level ℓ^* can be normal legal duration of labor, or average labor, or a physiologically ideal amount of labor.

Second, supposing that a given amount of endowment $k^* > 0$ is taken as a reference, and that $rk^* + w \geq c_0$, one can notice that the agents with a lower endowment are forced to work at least

$$OFL(k^*) = \max\{0, \frac{c_0 - rk^*}{w}\},$$

which will be called "objectively forced labor". The reference k^* can for instance be derived from a poverty line, or simply be the average endowment.

Different amounts of minimal consumption c_0 can be considered for different agents, for instance as a function of their needs (the agents can be households of different sizes), in which case the functions OLE and OFL are agent-specific.

More interestingly, this objective definition of constrained supply of labor can be generalized so as to incorporate a subjective kind of constraint. When an agent supplies ℓ_i, it may not be because he is objectively forced to do so in order to reach some minimal consumption c_0, but still, it might be that with a higher endowment this agent would no longer be willing to supply that much labor.

Following this intuition, we will say that agent i is *subjectively forced to sell at least* ℓ when $\ell_i(w, rk_i) \geq \ell$ and there is k such that for all $k' > k$, $\ell_i(w, rk') < \ell$. Indeed, in such a case, the agent is willing to supply ℓ (or more), but would refuse to do so if his endowment was high enough.

Again, one can use this definition in various ways. Suppose that a reference ℓ^* is considered. Then it may be interesting to register agents who are subjectively forced to work at least ℓ^*. One can define the "subjective liberation endowment" as

$$SLE_i(\ell^*) = \max\{k \mid \ell_i(w, rk) \geq \ell^*\}.$$

Then, agent i is subjectively forced to sell at least ℓ^* when $\ell_i(w, rk_i) \geq \ell^*$ and $k_i \leq SLE_i(\ell^*) < +\infty$. When leisure is normal, the condition $k_i \leq SLEi(\ell^*) < +\infty$ is necessary and sufficient (see Fig. 3).

Notice that the higher $SLE_i(\ell^*)$, the less economic pressure the agent suffers. When $SLE_i(\ell^*) = +\infty$, the pressure just disappears, because the agent is willing to sell ℓ^* for indefinitely high incomes. When leisure is not a normal good, then $\ell_i(w, rk_i)$ may fluctuate around ℓ^* as k_i increases. We chose to define $SLE_i(\ell^*)$ as the highest endowment k such that $\ell_i(w, rk) = \ell^*$ and above which $\ell_i(w, rk) < \ell^*$. Indeed, it would seem strange to consider that the agent suffers a strong pressure

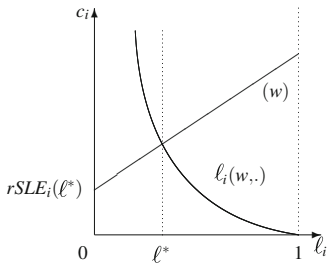


Fig. 3 Subjective liberation endowment. The curve $\ell_i(w,.)$ is the locus of pairs (c,ℓ) that are best in the budget $c = w\ell + rk$ for some k

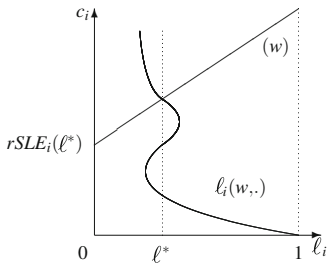


Fig. 4 *SLE* in a complex case

just because his labor supply is volatile around ℓ^*, while he is willing to sell ℓ^* for very high incomes. Taking the maximum threshold seems the only reasonable option (see Fig. 4).

The definition of *SLE* is a formal generalization of *OLE*, and both coincide when, over the relevant range, the agent works the minimum required to obtain c_0, i.e.,

$$\ell_i(w, rk) = \frac{c_0 - rk}{w}.$$

In the case when c_0 is a subsistence level which the agent seeks to attain in priority, one always has

$$\ell_i(w, rk) \geq \frac{c_0 - rk}{w}$$

and therefore, for any ℓ^*,

$$OLE(\ell^*) \leq SLE_i(\ell^*).$$

Symmetrically, if a reference k^* is given, it may be interesting to measure the amounts of labor that agents are subjectively forced to provide when their endowment is less than k^*. Subjectively forced labor is then defined as

$$SFL_i(k^*) = \max\{\ell \mid \forall k \leq k^*,\ \ell_i(w, rk) \geq \ell\}.$$

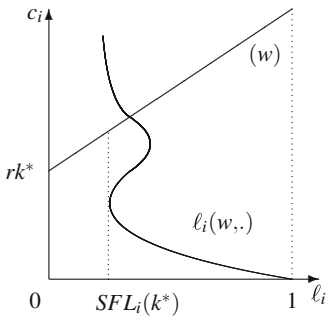


Fig. 5 Subjectively forced labor

When leisure is normal, one just has

$$SFL_i(k^*) = \ell_i(w, rk^*).$$

When leisure is not normal, one cannot take $\ell_i(w, rk^*)$ as forced labor due to endowment below k^* if there exists $k < k^*$ such that $\ell_i(w, rk) < \ell_i(w, rk^*)$. One should check that for all $k < k^*$ one has $\ell_i(w, rk)$ no less than the amount of forced labor, and this justifies the above definition (see Fig. 5).

Variants of these definitions can be imagined. For instance, suppose that k^* is an ideal amount of endowment, such as the per capita endowment in the economy. Then

$$\ell_i(w, rk_i) - \ell_i(w, rk^*)$$

is the additional amount of labor that agent i accepts because of an endowment lower than ideal.

The following proposition describes properties of the notions introduced here:

Proposition 1 *For all k^* and l^*:*

$$OLE(OFL(k^*)) = k^* \text{ when } OFL(k^*) > 0$$
$$OFL(OLE(\ell^*)) = \ell^*$$
$$SLE_i(SFL_i(k^*)) \geq k^* \text{ (with equality if leisure is strictly normal)}$$
$$SFL_i(SLE_i(\ell^*)) \leq \ell^* \text{ (with equality if leisure is normal).}$$

The simple proof is omitted.[3]

Up to now we have only provided definitions, and it remains to explain why these new notions can be ethically relevant by implying that some Walrasian trades are problematic. We can distinguish different cases.

[3]By "strictly normal", it is meant that the Engel curve is increasing.

Case 1. Suppose that for some reason it is regrettable that agent i has $k_i < k^*$, where k^* is some ideal endowment. For instance, k^* equals the average endowment, and equality of endowments would be the ideal situation. If one has

$$SLE_i(SFL_i(k_i)) < k^*,$$

one can say that agent i accepts to work the amount $SFL(k_i)$ (or more) only because his endowment is unduly low. If he had k^* or more, he would refuse to sell that much.

Why introduce $SFL_i(k_i)$ in the above condition, and not simply consider the condition $SLE_i(\ell_i) < k^*$? Because if $SFL_i(k_i) < \ell_i$, which may occur if leisure is not normal, then one may have

$$SLE_i(\ell_i) < k^* \leq SLE_i(SFL_i(k_i)),$$

which means that although the agent would not accept to work ℓ_i if his endowment was k^* or more, he would actually work less than ℓ_i (i.e., $SFL_i(k_i)$) for some endowment that is less than k_i. In such a case it would be strange to say that the situation is problematic, because the agent is not really forced to work ℓ_i —the agent can at most be considered forced to work as much as $SFL_i(k_i)$ (see Fig. 6).

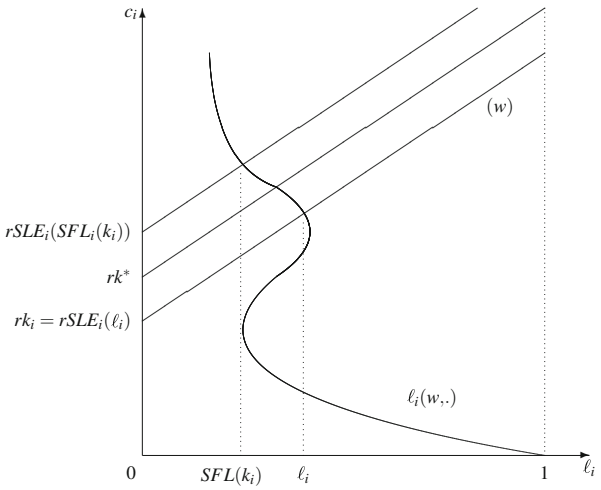


Fig. 6 A complex case

Case 2. Suppose that for some reason it is regrettable that agent i has $\ell_i > \ell^*$, where ℓ^* is some ideal amount of labor. For instance, ℓ^* is an amount of labor which is socially considered "decent". If one has

$$SFL_i(k_i) > \ell^*$$
$$SLE_i(\ell^*) < k^*$$

one can say that it is doubly regrettable that the agent works more than ℓ^*, and does so because of an unduly low endowment. Such a situation seems worse than the previous one.

Case 3. The situation is also worse than the first one if

$$SLE_i(OFL(k_i)) < k^*,$$

because in this case the agent is, currently, *objectively* forced to work an amount he would refuse to work if his endowment was k^*.

Case 4. The worst of all is when

$$OFL(k_i) > \ell^*$$
$$SLE_i(\ell^*) < k^*,$$

because the agent has no choice but to work more than ℓ^*, and would no longer accept it with a normal endowment.

As this discussion suggests, it is not in itself questionable if only one of $SFL_i(k_i) > \ell^*$ or $OFL(k_i) > \ell^*$ holds. After all, the agent might be a workaholic who would work more than ℓ^* no matter how rich he is. In such a case one cannot really say that his insufficient endowment is a relevant cause to his working that much. This is why it is essential to rely on SLE_i to check what the agent would choose if his endowment was sufficient.

The detection of questionable situations relies, here, on reference levels k^* and ℓ^*. It seems necessary to rely on such benchmarks. Otherwise one would face the following difficulty. Consider a case when

$$SLE_i(SFL_i(k_i)) < +\infty$$

but is extremely high, e.g., above the maximum endowment in an affluent population. This means that the agent would no longer accept to work that much only if his wealth was well above the currently observed endowments in this population. It seems hard to criticize this situation, because such high levels of wealth are just irrelevant. What is at stake here is the detection of situations where *disadvantage*, in terms of inequality or poverty, is at the root of the agent's supply or demand.

The selection of reference levels is not addressed in this paper, because the point of this analysis is to provide concepts that can be adapted to many different

normative views about what a decent endowment is or what a normal amount of labor is. Analysts focused on poverty and health may choose a lower k^* and a greater ℓ^* than analysts focused on inequality. The latter may want to take an average or a median value for these reference levels. Legal norms may also provide useful benchmarks. The legal hours (beyond which overtime pay is mandatory) may have been chosen for various reasons by the legislator, and independently of such reasons, it may be interesting to study if the workers who work overtime do it by choice or by constraint.

Another issue that this analysis raises is whether there is a link between forced trades and exploitation. Does the fact that some workers work more than a reference level under economic constraint mean that those who buy their labor take unfair advantage of the situation? In a Walrasian context with many agents, an employer who hires someone who works under economic constraint is not benefiting from this single worker's situation, because the market wage rate does not depend on one single agent. But when there are many individuals who are constrained to work more than they would under better circumstances, the effect on market wages is substantial and generally in the direction of lowering wages to the benefit of the employers. Therefore, while no single transaction between two agents can be identified as exploitative in isolation, it may be part of a general pattern in which a "class" of employers benefits from the constraints imposed on the "class" of workers.

One can, however, argue that the concepts introduced here are about forced individual *choices*, rather than about forced *transactions*. If the benchmark endowment k^* is defined on the basis of a poverty line and the whole economy is populated by poor individuals, they may be forced to work more than normal because of their poverty, without there being a class of exploiters. One can then recognize that something ethically unpleasant is taking place in this economy, although it not about unfair advantage or exploitation, but simply economic constraint. Moreover, notice that each of the four critical cases described above is compatible with the agent working in her own workshop, or even being a buyer, not a seller, of labor.

But these concepts can easily be applied to specific transactions, as the following simple example illustrates. Assume that, at the prevailing Walrasian equilibrium, every agent actually retains his own land, and works in priority over his own land. All agents use the same technique with factor ratio $\overline{\ell}/\overline{k}$, where $\overline{\ell}$ denotes the average labor and $\overline{k}$ the average endowment. If

$$k_i \frac{\overline{\ell}}{\overline{k}} < \ell_i,$$

the agent has an excess of labor that he can sell to other agents. If the inequality is the other way around, then the agent does not work enough to make use of his land and he has to hire other agents. Suppose, now, that one takes k^* to be the average endowment $\overline{k}$.If one wants to avoid forced *sale* of labor (due to inequality) at the

equilibrium, one has to check that there is no agent i such that

$$SLE_i(SFL_i(k_i)) < \overline{k}$$
$$k_i \frac{\overline{\ell}}{\overline{k}} < \ell_i.$$

If leisure is a strictly normal good, then $SLE_i(SFL_i(k_i)) = k_i$, and one only has to check that for every i,

$$\frac{k_i}{\overline{k}} \geq \min\{1, \frac{\ell_i}{\overline{\ell}}\},$$

i.e., that no agent with less-than-average endowment supplies labor on the market.

3 Forced Consumption

The previous model was also compatible with a different interpretation in terms of trade. It could describe an economy where all agents work on their own workshop, but trade land if their endowment does not fit their own quantity of labor. Then an agent will have to buy land if and only if

$$k_i < \ell_i \frac{\overline{k}}{\overline{\ell}},$$

which is the same condition as above for selling labor. In this new interpretation, one can talk about forced purchase of land instead of forced sale of labor.

But we will turn to another example of forced purchase. Suppose there are two consumption goods in the economy, hamburgers h and caviar c. For standard preferences in the population, the former is an inferior good, while the latter is a luxury good. Hamburger is the numeraire, and the price of caviar is denoted p.

We will assume that survival requires a minimal consumption

$$h_i + c_i \geq 1,$$

in which hamburger and caviar have a symmetric role, but we will also assume that caviar is more expensive: $p > 1$.

Individual i has an income I_i and his budget constraint is (see Fig. 7):

$$h_i + pc_i = I_i.$$

Let $h_i(p, I_i)$ and $c_i(p, I_i)$ denote the Marshallian demands of individual i.

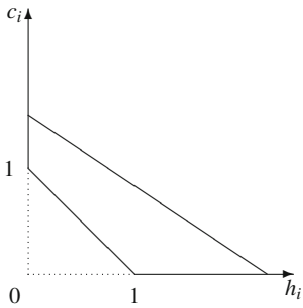


Fig. 7 Consumption set and budget set

Under these assumptions survival requires $I_i \geq 1$ and

$$h_i + \frac{I_i - h_i}{p} \geq 1,$$

or equivalently,

$$h_i \geq \frac{p - I_i}{p - 1}.$$

By analogy to the previous section, we can define an objectively forced consumption of hamburger for $1 \leq I_i \leq p$:

$$OFC(I_i) = \frac{p - I_i}{p - 1},$$

and an objective liberation income for $0 \leq h_i \leq 1$:

$$OLI(h_i) = p\,(1 - h_i) + h_i\,.$$

Figure 8 illustrates the computations.

Turning to subjective constraints, we can similarly define subjectively forced consumption as the amount which agent i would consume for any smaller income (see Fig. 9, where the curve $h_i(p, .)$ is the locus of pairs (c, h) that are best in the budget $h + pc = I$ for some $I \geq 1$):

$$SFC_i(I_i) = \max\{h \mid \forall I \in [1, I_i],\ h_i(p, I) \geq h\}$$

and a subjective liberation income as the income above which the agent i would no longer consume that much (see Fig. 10):

$$SLI_i(h_i) = \max\{I \mid h_i(p, I) \geq h_i\}.$$

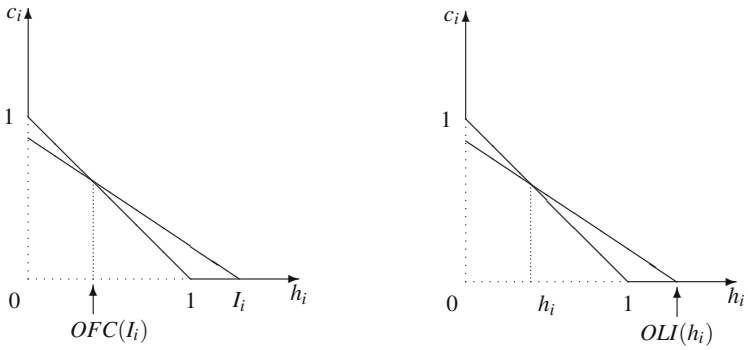


Fig. 8 Objective constraint

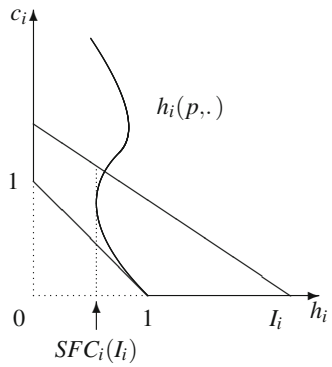


Fig. 9 Subjectively forced consumption

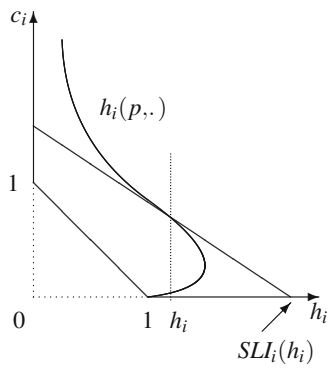


Fig. 10 Subjective liberation income

Such notions can be put to use in the following fashion. Suppose that there is some ideal income $I^* > 1$ such that it is regrettable if $I_i < I^*$, and some ideal consumption $h^* < 1$ such that it is regrettable (say, for health purposes) if $h_i > h^*$.

If agent i is such that

$$SLI_i(SFC_i(I_i)) < I^*,$$

one can say that this agent is subjectively forced to consume at least $SFC_i(I_i)$, because of an unduly low income. As a consequence, those agents who sell him that amount (or more) are unduly benefiting from his having $I_i < I^*$.

The situation is worse if

$$SLI_i(h^*) < I^*,$$
$$SFC_i(I_i) > h^*,$$

because it involves an excessive consumption by the h^* standard.

The situation is also worse if

$$SLI_i(OFC(I_i)) < I^*,$$

because the agent is objectively forced.

And the worst of all is

$$SLI_i(h^*) < I^*,$$
$$OFC(I_i) > h^*,$$

because the agent is objectively forced to overconsume hamburgers, while he would refuse to do so with ideal income.

The parallelism between this example and the previous one suggests that a generalization of these concepts is not out of reach. Such a generalization is attempted in the next section.

4 Forced Trade: A General Approach

4.1 Framework

Consider a standard Arrow-Debreu model with ℓ goods. We assume that the prevailing price vector $p \in \mathbb{R}^{\ell}_{++}$ is fixed throughout. An individual agent i has an endowment $\omega_i \in \mathbb{R}^{\ell}$ in goods.In case of production, we will assume that this

endowment contains the shares the agent owns in the m firms of the economy, that is:

$$\omega_i = \bar{\omega}_i + \sum_{j=1}^{m} \theta_{ij} y_j$$

where $\bar{\omega}_i$ denotes his personal endowment (as a household), θ_{ij} his share in firm j, and $y_j \in \mathbb{R}^\ell$ the production vector of firm j (with positive components for net outputs, and negative components for net inputs). Obviously, the production vectors depend on the prevailing prices but we will assume here that, like prices, the production plans of firms are fixed.

Agent i also has a consumption set $X_i \subset \mathbb{R}^\ell$, and a Walrasian demand correspondence for goods $x_i(p, \omega_i) \subset X_i$, derived from maximization of satisfaction over the budget set

$$B_i(p, \omega_i) = \{x \in X_i \mid px \leq p\omega_i\}.$$

We adopt the convention that for a bundle $x \in X$ and a good k, $x_k > 0$ means that good k is consumed, whereas $x_k < 0$ means that good k is a labor service provided by the agent. The demand set (or expansion path) of the agent is the set of all demanded bundles at all possible endowments:

$$D_i = \bigcup_{\omega \in \mathbb{R}^\ell} x_i(p, \omega).$$

(We adopt the convention that $x_i(p, \omega) = \emptyset$ whenever $B_i(p, \omega) = \emptyset$.)

We need the following notations. For a given bundle $x \in \mathbb{R}^\ell$, let

$$x^{\nearrow} = \left\{z \in \mathbb{R}^\ell \mid \forall k = 1, \ldots, \ell,\ x_k z_k \geq x_k^2\right\},$$

that is, $x^{\nearrow}$ is the set of bundles whose components have the same sign as components of x, and are larger (when $x_k = 0$, this puts no constraint on z_k). Concretely, $x^{\nearrow}$ is the set of bundles such that the agent consumes at least as much, and works at least as much, as in x. And, for any closed subset $A \subset \mathbb{R}^\ell$, let

$$A_{\text{inf}} = a \in \mathbb{R}^\ell \text{ such that:}$$
$$A \subset a^{\nearrow} \text{ and } \forall b \in a^{\nearrow},\ A \subset b^{\nearrow} \Rightarrow b = a.$$

In other words, A_{inf} is the maximal bundle a (in the "$\nearrow$" sense) such that $A \subset a^{\nearrow}$. Notice one always has $A \subset (A_{\text{inf}})^{\nearrow}$. Finally, let pA denote

$$pA = \{m \in \mathbb{R} \mid \exists a \in A,\ m = pa\}.$$

4.2 Forced Consumption and Trade

We are now equipped to give definitions which generalize the previous sections.

The agent is no longer objectively forced to demand a given bundle x or more, that is, to consume a bundle in $x^{\nearrow}$, when his income allows him to consume something in $X_i \setminus x^{\nearrow}$. One can then define the *objective liberation income* for x (see Fig. 11) as

$$OLI_i(x) = \inf p\left(X_i \setminus x^{\nearrow}\right).$$

Figure 11 illustrates two cases, with one in which x is not in X_i.

Conversely, with endowment ω_i the agent is objectively forced to consume at least any bundle x such that $B_i(p, \omega_i) \subset x^{\nearrow}$. Therefore, his *objectively forced consumption* is then:

$$OFC_i(\omega_i) = B_i(p, \omega_i)_{\inf}$$

and, similarly, his *objectively forced trade* is the maximal trade q such that for any $x \in B_i(p, \omega_i)$, $x - \omega_i \in q^{\nearrow}$, that is (see Fig. 12):

$$OFT_i(\omega_i) = [B_i(p, \omega_i) - \omega_i]_{\inf}.$$

This vector is illustrated in Fig. 12 as the arrow from ω_i to $B_i(p, \omega_i)_{\inf}$.

Notice that these concepts capture minimal constraints. One may also want to focus on $B_i(p, \omega_i)$ rather than just $B_i(p, \omega_i)_{\inf}$, in some cases, or concatenate some dimensions in order to talk about constraints over aggregate goods or services. For instance, consider a household with a couple without children. Suppose they have no unearned income, and assume that they need at least \$15,000 per year to live in a decent way. Both are able to work and earn \$30,000, but in different kinds of jobs: he is a nurse while she is a carpenter. According to the $B_i(p, \omega_i)_{\inf}$ concept, they

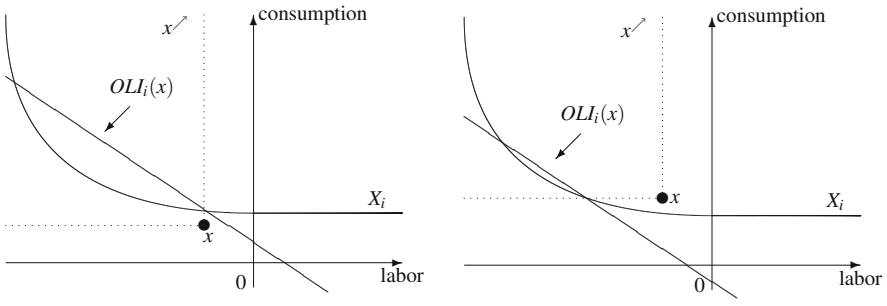


Fig. 11 Objective liberation income. The *continous curve* delineates X_i, the *dotted lines* delineate $x^{\nearrow}$

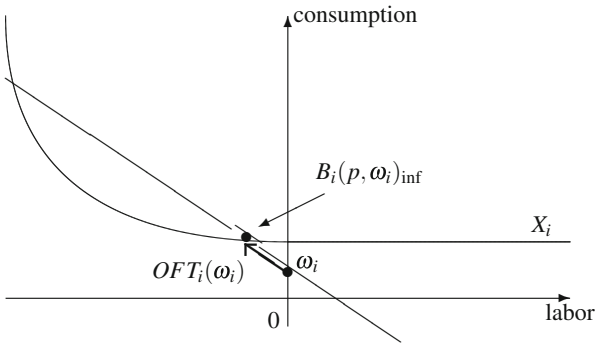


Fig. 12 Objectively forced consumption and trade

are not objectively forced to work, in the sense that they can live without providing any hour of nurse or any hour of carpenter to the market. But of course, aggregating nurse hours and carpenter hours gives a different picture: at least one of them must work half time.

The extension of these definitions to the idea of subjective constraints is essentially obtained by substituting D_i to X_i in the above definitions, if we assume that preferences are locally non-satiated in order to simplify the analysis (this allows us to have $x_i(p,\omega) = D_i \cap \{z \in X \mid pz = p\omega\}$). However, the definition of a subjective liberation income by the formula

$$SLI_i(x) = \inf p\left(D_i \setminus x^{\nearrow}\right)$$

is not satisfactory when the agent's demand is not normal, since it may happen that with some high income the agent is still willing to demand a bundle in $x^{\nearrow}$ again. As explained in the previous sections, the subjective liberation income must be such that, above it, the agent is no longer willing to consume in $x^{\nearrow}$. The appropriate generalization of the definitions of the previous sections is then (see Fig. 13):

$$SLI_i(x) = \sup p\left(D_i \cap x^{\nearrow}\right),$$

and this definition correctly yields $+\infty$ whenever the agent is willing to consume in $x^{\nearrow}$ for indefinitely high incomes. This definition can be extended again to accommodate any requirement that an area of X_i should be avoided (and in particular the case when $x_i(p,\omega_i)$ is not a singleton). For any $A \subset X_i$, one may define the subjective liberation income for A as the income above which the agent no longer accepts to consume in A :

$$SLI_i(A) = \sup p\left(D_i \cap A\right).$$

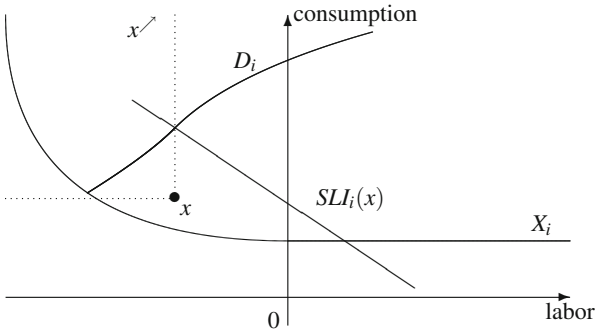


Fig. 13 Subjective liberation income. The *upward sloping curve* is D_i, the demand set defined by a given p and varying endowments

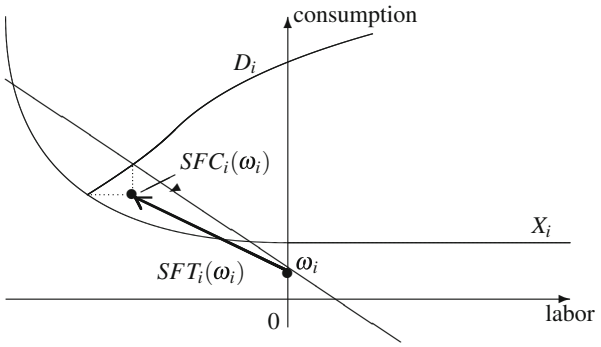


Fig. 14 Subjectively forced consumption and trade

With the other definitions the simple substitution of D_i to X_i provides the correct extension, and this can be denoted here as follows (see Fig. 14):

$$SFC_i(\omega_i) = [B_i(p, \omega_i) \cap D_i]_{\inf},$$
$$SFT_i(\omega_i) = [(B_i(p, \omega_i) \cap D_i) - \omega_i]_{\inf}.$$

There are general properties to be noted, about these various concepts.

Proposition 2

(i) *For all* $x \in \mathbb{R}^\ell$, $OLI_i(x) \leq SLI_i(x)$.
(ii) *For all* $x, q \in \mathbb{R}^\ell$, $x \in q^{\nearrow} \Rightarrow OLI_i(x) \leq OLI_i(q)$.
(iii) *For all* $A, B \subset \mathbb{R}^\ell$, $A \subset B \Rightarrow SLI_i(A) \leq SLI_i(B)$.
(iv) *For all* $\omega_i \in \mathbb{R}^\ell$, $SFC_i(\omega_i) \in OFC_i(\omega_i)^{\nearrow}$.
(v) *For all* $\omega_i \in \mathbb{R}^\ell$, $OLI_i(OFC_i(\omega_i)) \geq p\omega_i$.
(vi) *For all* $\omega_i \in \mathbb{R}^\ell$, $SLI_i(SFC_i(\omega_i)) \geq p\omega_i$.

Proof

(i) One has

$$\begin{aligned} OLI_i(x) &= \inf p\left(X_i \setminus x^{\nearrow}\right) \\ &\leq \inf p\left(D_i \setminus x^{\nearrow}\right) \\ &\leq \inf p\left(D_i \setminus x^{\nearrow}\right) \cap p\left(D_i \cap x^{\nearrow}\right) \\ &\leq \sup p\left(D_i \setminus x^{\nearrow}\right) \cap p\left(D_i \cap x^{\nearrow}\right) \\ &\leq \sup p\left(D_i \cap x^{\nearrow}\right) = SLI_i(x). \end{aligned}$$

(ii) One has

$$\begin{aligned} x \in q^{\nearrow} &\Rightarrow x^{\nearrow} \subset q^{\nearrow} \\ &\Rightarrow X_i \setminus q^{\nearrow} \subset X_i \setminus x^{\nearrow} \\ &\Rightarrow \inf p\left(X_i \setminus x^{\nearrow}\right) \leq \inf p\left(X_i \setminus q^{\nearrow}\right). \end{aligned}$$

(iii) The reasoning is the same as for (ii).

(iv) $A \subset B$ entails $A_{\inf} \in (B_{\inf})^{\nearrow}$, so that

$$SFC_i(\omega_i) = [B_i(p,\omega_i) \cap D_i]_{\inf} \in [B_i(p,\omega_i)_{\inf}]^{\nearrow}.$$

(v) $B_i(p,\omega_i) \subset [B_i(p,\omega_i)_{\inf}]^{\nearrow}$, so that

$$OLI_i(OFC_i(\omega_i)) = \inf p\left(X_i \setminus [B_i(p,\omega_i)_{\inf}]^{\nearrow}\right) \geq \inf p\left(X_i \setminus B_i(p,\omega_i)\right) = p\omega_i.$$

(vi) One has

$$\begin{aligned} SLI_i(SFC_i(\omega_i)) &= \sup p\left(X_i \cap [(B_i(p,\omega_i) \cap D_i)_{\inf}]^{\nearrow}\right) \\ &\geq \sup p\left(X_i \cap (B_i(p,\omega_i) \cap D_i)\right) \\ &= \sup p\left(B_i(p,\omega_i) \cap D_i\right) = p\omega_i. \end{aligned}$$

□

Let us now turn to the ethical discussion of situations where economic pressure is problematic. Following the same intuition as in the above examples, we can say that if

$$SLI_i(SFC_i(\omega_i)) < I^*,$$

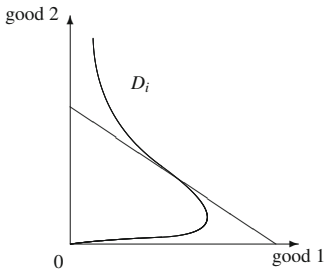


Fig. 15 Inferior good for high incomes

where I^* is an ideal income (such as the average level in the population), then there is a problem, because the agent is subjectively forced to consume bundles which he would no longer accept to consume if his income reached the reference I^*.

In some cases, it also makes sense to worry about the situation

$$SLI_i(x_i(p, \omega_i)) < I^*,$$

for instance in the case illustrated in Fig. 15, where the agent has

$$SLI_i(SFC_i(\omega_i)) = +\infty$$

but

$$SLI_i(x_i(p, \omega_i)) = p\omega_i.$$

In this example, $SFC_i(\omega_i)$ is just the bottom point of X and is not very relevant. Because good 1 is inferior for high incomes, one may say that the agent is forced to consume much of it because of his low income. The examples of the two previous sections did not provide similar cases because in those examples $SFC_i(\omega_i)$ was not much influenced by the shape of the agent's demand at very low incomes.

Similarly, suppose that there is a consumption subset $X^- \subset X_i$ such that it is considered problematic if an agent consumes $x \in X^-$ (for instance, $x \in X^-$ means that the agent suffers from malnutrition and overworks). Then, if one indeed observes $x \in X^-$, and moreover

$$SFC_i(\omega_i) \in X^-$$
$$SLI_i(X^-) < I^*,$$

then the situation is worse than above. (Notice that it implies $SLI_i(SFC_i(\omega_i)) < I^*$.)

Again, the situation is also rather bad if

$$SLI_i(OFC_i(\omega_i)) < I^*,$$

since the agent is objectively forced to accept consumptions he would reject with the ideal income.

Finally, the worst of all is when

$$OFC_i(\omega_i) \in X^-$$
$$SLI_i(X^-) < I^*.$$

4.3 Equilibria With and Without Forced Trade

The concepts defined above allow us to speak rigorously about agents who accept particular consumptions because of inadequate income and not by pure preference. We now proceed to examine whether it is possible to obtain Walrasian equilibria such that no agent suffers from undue economic pressure, and we will focus on equilibria where no agent suffers from

$$SLI_i(x_i(p, \omega_i)) < \bar{I},$$

where $\bar{I}$ is the average income in the population, or at least where no agent has

$$SLI_i(SFC_i(\omega_i)) < \bar{I}.$$

In other words, is it possible to make sure that no agent accepts certain consumption and work just because of a below-average income? Obviously, full equality of incomes in the population provides a sufficient condition for this to be achieved. But can one characterize the set of situations where any of the above requirement is satisfied?

An equilibrium is such that $SLI_i(x_i(p, \omega_i)) \geq \bar{I}$ for all i if and only if for every agent i such that $I_i < \bar{I}$, one has, by definition:

$$\sup p \left(D_i \cap x_i(p, \omega_i)^{\nearrow} \right) \geq \bar{I},$$

which is equivalent, since $D_i \cap x_i(p, \omega_i)^{\nearrow}$ is not empty and $x_i(p, \omega_i)$ is closed, to the condition that there exists ω' such that $p\omega' \geq \bar{I}$ and

$$x_i(p, \omega') \cap x_i(p, \omega_i)^{\nearrow} \neq \emptyset. \quad (1)$$

A *sufficient* condition for this to be obtained is that for ω' such that $p\omega' = \bar{I}$,

$$x_i(p, \omega') \subset x_i(p, \omega_i)^{\nearrow}.$$

And a *sufficient* condition for the latter to be obtained is that, for incomes over the range $[p\omega_i, \bar{I}]$, consumption goods (those k such that $x_{ik}(p, \omega_i) > 0$) must be normal, whereas labor services (those k such that $x_{ik}(p, \omega_i) < 0$) must be inferior.

One should not hope for more clearcut necessary and sufficient conditions than condition (1) here, because (1) can be satisfied in many different ways by complex demand correspondences.

If one focuses on the weaker condition that $SLI_i(SFC_i(\omega_i)) \geq \bar{I}$ for all i, one obtains the necessary and sufficient condition that for every agent i such that $I_i < \bar{I}$, there must exist ω' and ω'' such that $p\omega' \geq \bar{I}$, $p\omega'' \leq p\omega_i$ and

$$x_i(p, \omega') \cap x_i(p, \omega'')^{\nearrow} \neq \emptyset.$$

A sufficient condition for this to be obtained, in addition to the ones described above, is that for low endowments the demand x_i must be close enough to zero. As described in the previous section, in the case of consumption goods, this condition may be automatically satisfied because of the objective pressure of the budget constraint, and this new sufficient condition is then not very relevant, from the ethical standpoint.

Let us now examine the likelihood of observing undue economic pressure, in the sense of the above two conditions, in any market equilibrium. From the above discussion, it is easy to derive the conclusion that if all goods and services are normal, then it is very likely to observe unduly forced labor, but no forced consumption (of ordinary goods) will prevail. On the contrary, if many or most goods and services are inferior over the relevant range (that is, between the lowest and the average incomes), then one will not observe any forced labor, but forced consumption will be commonplace. The situation which is the least favorable is when labor services are normal, while consumption goods are inferior.

The latter situation is actually quite plausible for unpleasant kinds of works (dangerous or dirty chores) and for low quality consumption goods (industrial food of dubious quality, low quality clothes). As a consequence, one may safely conjecture that forced labor and forced consumption, in the sense defined here, do prevail over a large scale in most world economies.

5 Conclusion

The concepts developed here are meant to give a rigorous formulation to the widespread intuition that poor people are not only agents who consume too little, but also agents who work too much (in bad jobs) and consume too much (of bad quality goods). Some ideas for future research are proposed in this conclusion.

First, there is an obvious link between objective and subjective constraint, the latter generalizing the former, as explained in Sect. 2. More precisely, the objective constraint corresponds to the subjective constraint for individuals who seek to minimize the contemplated trade in priority. But another understanding of

the objective constraint, as defined in this paper, is that the agent is objectively constrained when he is subjectively constrained *for all* possible preferences. This formulation suggests a notion that would be intermediate between the objective and subjective notions, and would examine the size of the set of preferences for which the agent is constrained. The larger this set, the more objective is the constraint.

Moreover, one could focus on a subset that is centered on the agent's actual preferences. Intuitively, the idea would be that the agent is more constrained when it would require a greater change to his preferences in order to obtain a situation in which he is not subjectively constrained. This is an idea that needs topological concepts on preferences, as in topological social choice [3].

A related idea would be to extend the concepts proposed here in the direction of defining a *degree* of constraint. The definitions offered in this paper only seek to decide whether an individual is or is not constrained. But it would also make sense to seek a measure of economic constraint that would vary between 0 and 1. Two directions could be explored for this purpose. First, the relative size of the set of preferences for which the agent is constrained, or the minimal distance between his preferences and the preferences for which he would no longer be constrained, could be used to construct the index of constraint. Such a measure could focus on a particular trade and measure how much economic pressure the agent endures to accept this particular trade. Another possibility, in the direction of measuring the general economic constraint endured by the agent, would be to measure the size of the set of trades that the agent is forced to accept.

Finally, establishing a rigorous terminology to depict the ethical problems raised by economic pressure obviously suggests to examine remedies. Two general kinds of practical solutions are available. One is based on redistribution, and seeks to radically solve the problem by freeing the poor from the constraint of poverty itself. Another kind of policy consists in regulating the market and prohibiting the bad trades that poor people are likely to accept.

In first approximation, one may guess that the former is the most favorable to the target population, because the latter is likely to make them actually worse off according to their own preferences, at least in the short run. In some contexts, it appears, however, that prohibition may alter market prices so that, in the end, the poor actually benefit. An example dealing with child labor is provided by Basu [5]. The mixed results obtained on the impact of minimal wages on the labor market also suggest that prohibiting low-wage jobs does not necessarily hurt the potential low-wage earners.

The prohibition policy has, at any rate, often been chosen. There may be several reasons for that. For instance, bad trades which endanger health create negative externalities. Paternalistic views may insist on prohibiting certain patterns of consumption. But there might be another explanation, coming from political economy. Of the two policies described above, redistribution is the most effective and favorable to the poor, but is also the most costly to the rich. It may be much more acceptable to the rich to prohibit the most conspicuous and repugnant forms of bad trades without doing any redistribution. This is certainly costly to the rich as well, but probably much less than direct transfers.

When dealing with the issue of bad trades, one should be aware that the boundary between the acceptable and the unacceptable is really a matter of social convention. The work schedules and kinds of jobs that appeared normal in the last century now seem awful. Slavery seemed natural to Aristotle just as wage contracts seem natural to most of our contemporaries. Now, the concept of subjective liberation income may help to get more insight in the trend that affects the boundary of the acceptable. Just take a high income and examine what people would no longer be willing to accept if they had such income. This may give some indication about where the boundary of the acceptable will lie in the next centuries. What this device neglects is the potential impact of culture shifts. But at least some tendencies may be detected in that way.

Acknowledgements This paper has benefited from conversations with N. Yoshishara, from comments by N. Baigent, J. Perez-Castrillo and R. Veneziani, as well as reactions from audiences at workshops in Palermo and Cambridge.

References

1. Baigent N (1981) Decompositions of minimal libertarianism. Econ Lett 7:29–32
2. Baigent N (1981) Social choice and merit goods. Econ Lett 7:301–305
3. Baigent N (2011) Topological social choice. In: Arrow KJ, Sen AK, Suzumura K (eds) Handbook of social choice and welfare, vol 2. North-Holland, Amsterdam
4. Baigent N, Gaertner W (1996) Never choose the uniquely largest: a characterization. Econ Theory 8:239–249
5. Basu K (1999) Child labor: cause, consequence, and cure, with remarks on international labor standards. J Econ Lit 37:1083–1119
6. Buchanan JM (1977) Political equality and private property: the distributional paradox. In: Dworkin G, Bermant G, Brown PG (eds) Markets and morals. Wiley, New York
7. de Schweinitz K Jr (1979) The question of freedom in economics and economic organization. Ethics 89:336–353
8. Ellerman D (1992) Property and contract in economics. Blackwell, Oxford
9. Frankfurt H (1973) Coercion and moral responsibility. In: Honderich T (ed) Essays on freedom of action. Routledge, London
10. Friedman M (1962) Capitalism and freedom. University of Chicago Press, Chicago
11. Gibbard A (1985) What's morally special about free exchange? In: Paul EF, Miller FD, Paul J (eds) Ethics and economics. Blackwell, Oxford
12. Lyons D (1975) Welcome threats and coercive offers. Philosophy 50:425–436
13. Macpherson CB (1973) Democratic theory. University Press, Oxford
14. Nozick R (1969) Coercion. In: Morgenbesser S, Suppes P, White M (eds) Philosophy, science, and method. St. Martin's Press, New York
15. O'Neill O (1985) Between consenting adults. Philos Public Aff 14:252–277
16. Olsaretti S (1998) Freedom, force and choice: against the rights-based definition of voluntariness. J Pol Philos 6:53–78
17. Peter F (2004) Choice, consent, and the legitimacy of market transactions. Econ Philos 20:1–18
18. Rawls J (1971) A theory of justice. Harvard University Press, Cambridge
19. Roemer J (1982) A general theory of exploitation and class. Harvard University Press, Cambridge

20. Samuels WJ (1997) The concept of ‘coercion’ in economics. In: Samuels WJ, Medema SG, Schmid AA (eds) The economy as a process of valuation. Edward Elgar, Cheltenham
21. Samuelson P (1966) Modern economic realities and individualism. In: Stiglitz J (ed) Collected scientific papers. MIT Press, Cambridge
22. Satz D (2010) Why some things should not be for sale: the moral limits of markets. Oxford University Press, New York
23. Scanlon TM (1977) Liberty, contract, and contribution. In: Dworkin G, Bermant G, Brown PG (eds) Markets and morals. Wiley, New York
24. Sunstein CR (1989) Disrupting voluntary transactions. In: Chapman JW, Pennock JR (eds) Markets and justice. New York University Press, New York
25. Thomson W (2011) Fair allocation. In: Arrow KJ, Sen AK, Suzumura K (eds) Handbook of social choice and welfare, vol 2. North-Holland, Amsterdam
26. Trebilcock M (1993) The limits of freedom of contract. Harvard University Press, Cambridge
27. Zimmerman D (1981) Coercive wage offers. Philos Public Aff 10:121–145

Unequal Exchange, Assets, and Power: Recent Developments in Exploitation Theory

Roberto Veneziani and Naoki Yoshihara

Abstract This paper surveys and extends some recent contributions on the theory of exploitation as the unequal exchange of labour. A model of dynamic economies with heterogeneous optimising agents is presented which encompasses the models used in the literature as special cases. It is shown that the notion of exploitation is logically coherent and can be meaningfully analysed in such a general framework. It is then shown that the axiomatic approach of social choice theory can be adopted to explore the normative foundations of the notion of exploitation. Finally, it is argued that purely distributive approaches to exploitation are not entirely compelling and a notion of dominance, or unequal power is necessary.

Keywords Axiomatic social choice • Exploitation

1 Introduction

The notion of exploitation is prominent in the social sciences and in political philosophy. It is central in Marxist-based analyses of labour relations but it is also extensively discussed in liberal approaches, especially in the analysis of (possibly mutually beneficial) trades characterised by significant disparities in bargaining

R. Veneziani
School of Economics and Finance, Queen Mary University of London, Mile End Road, London E1 4NS, UK
e-mail: r.veneziani@qmul.ac.uk

N. Yoshihara (✉)
The Institute of Economic Research, Hitotsubashi University, Naka 2-1, Kunitachi, Tokyo 186-0004, Japan
e-mail: yosihara@ier.hit-u.ac.jp

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*, Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_15

power.[1] Yet, it has received relatively little attention in social choice theory and in normative economics. This is due partly to the traditional association of exploitation theory with the labour theory of value, whose logical flaws are assumed to carry over to the notion of exploitation, and to the fact that exploitation is usually analysed under fairly restrictive assumptions concerning technology, preferences, and endowments. But it is also due to the focus of social choice theory on distributive issues, and more specifically on the distribution of welfare, income, wealth, resources, or more recently, capabilities and opportunities.

John Roemer's classic work [23–25] has demonstrated that a coherent notion of exploitation can be provided independently of the labour theory of value. Moreover, he has proved that at least some of the key insights of exploitation theory hold outside of simple Leontief economies with homogeneous labour, subsistence consumption, and a polarised class structure. Somewhat paradoxically, however, the operation succeeded, but the patient died: the main conclusion of Roemer's work is that a concern for asset inequalities is the only sound legacy of exploitation theory, which reduces to a variant of liberal egalitarianism and is "a domicile that we need no longer maintain: it has provided a home for raising a vigorous family, who now must move on" [27, p. 67].

This paper surveys and extends recent work in exploitation theory and argues that the concept of exploitation is logically and theoretically sound, *and* provides interesting normative insights on the wrongs that characterise advanced capitalist economies, which go beyond the standard distributive focus of social choice theory and normative economics.

First, exploitation can be rigorously analysed in a rather general framework. Section 2 sets up a model of a dynamic economy with a convex technology, and heterogeneous optimising agents endowed with different amounts of physical and human capital. We discuss both the individual maximisation programme and the equilibrium notion—the concept of *Reproducible Solution* proposed by Roemer [22, 23],—and show that the structure of the economy is similar to standard growth models. Then we show that, contrary to the received view, the standard static models used in the literature are not ad hoc and can be interpreted as focusing on the steady state equilibria of the general model.

Second, unlike in the main theories of distributive justice, exploitation focuses on labour as the variable of normative interest. In the theory of *exploitation as an unequal exchange* (UE) *of labour*, exploitative relations are characterised by systematic differences between the amount of labour that individuals contribute to the economy, in some relevant sense, and the amount of labour they receive, in some relevant sense, via their income.

[1]The literature is too vast for a comprehensive list of references, but *recent* contributions include van Donselaar [32], Ypi [44], Fleurbaey [9], Steiner [31], Vrousalis [37].

We argue that the key normative insights of the notion of exploitation can be captured within the rigorous axiomatic framework of social choice theory. An axiomatic approach to exploitation was long overdue: outside of simple stylised economies, many definitions can be, and have in fact been proposed that incorporate different positive and normative intuitions. By adopting an axiomatic method, we start from first principles, thus explicitly discussing the intuitions underlying UE exploitation. Moreover, an axiomatic approach demonstrates that the notion of exploitation is not obscure or incoherent, and relies on some theoretically robust and normatively relevant intuitions that can be precisely stated in the rigorous language of normative economics.

Section 3 discusses some recent axiomatic analyses of exploitation theory. In particular, we analyse a characterisation of the class of UE exploitation-forms as indicators of capitalist relations of production that allow wealthy agents to appropriate social surplus generated from social labour as profits. Our characterisation result provides the necessary and sufficient condition for *coherent* definitions of exploitation, in that the basic property of exploitation and profits holds regardless of the complexity of the economic models. This characterisation leads us to conclude that among main approaches, an extension of the "New Interpretation" form of exploitation [4, 5, 10, 11] is the only coherent definition in this respect.

Another contribution of the paper is to argue that, unlike most of normative economics and social choice, the notion of exploitation suggests that the wrongs of capitalist economies go beyond inequalities in economic outcomes or opportunities. As Roemer [23, 25, 27] has forcefully argued, distributive injustices are at the core of exploitative relations and theories of exploitation based on dominance in the workplace or coercion in the labour market are unsatisfactory. However, at the philosophical level, purely distributive approaches—such as Roemer's—have too impoverished an informational basis to capture exploitative relations and to distinguish exploitation from other forms of injustice, or wrongs. Some notion of power, or dominance, or asymmetric relations between agents is an essential—definitional—part of exploitation, and this emphasis on the structure of the interaction between agents that allows someone to take (unfair) advantage of somebody else is an important contribution of exploitation theory that may correct the "distributive bias" of normative economics. In this respect, we take inspiration from some seminal contributions by Nick Baigent [1], which explore rights and more generally non-consequentialist principles in social choice.

Based on the general model set up in Sect. 2, Sect. 4 builds on and extends some recent contributions that analyse exploitation in a dynamic context. It is shown that inequalities in productive assets are not sufficient for exploitation to provide foundations to exploitation as a persistent phenomenon. Something else is necessary in order to generate persistent exploitation, and power or dominance are natural candidates for that role.

2 The General Model

This section sets up the model and the relevant equilibrium notion. Compared to the standard literature in exploitation theory and mathematical Marxian economics, our economies are general in at least three key dimensions. First, as in Roemer [22], we allow for a general convex cone production set, rather than the canonical Leontief or von Neumann technology. As is well known [17, 18], outside of the simple linear production model, many of the classical Marxian propositions do not necessarily hold: there is no obvious way of defining the labour value of each commodity; it is not clear that the two aggregate equalities between the sum of prices and the sum of values, and the sum of profits and the sum of surplus values can simultaneously hold; and so on. We aim to show that a logically consistent and theoretically rigorous notion of UE exploitation can instead be provided even in general production economies. This is essential in order to defend the normative relevance of UE in advanced capitalist economies.

Second, unlike in the standard literature, we do not focus on polarised, two-class economies in which capitalists save and accumulate while workers spend their wage revenue to buy a fixed subsistence bundle. Rather, we allow for agent heterogeneity concerning endowments of physical assets, as in Roemer [22, 23], and also for heterogeneous preferences and human capital. Further, rather than *assuming* individuals to belong to given classes, our general models allow one to analyse the class structures that endogenously emerge in the equilibrium of economies in which agents are allowed to save and thus class mobility is not ruled out.

In fact, third, we take account of the dynamic structure of the economy. On the one hand, as in Roemer [22, 23], we explicitly incorporate the time structure of production processes—whereby production takes time and outputs emerge only at the end of a given production period—and the fact that capital goods are reproducible. This is a major difference with standard neoclassical models which usually ignore the time structure of production, as in Walrasian general equilibrium theory, or treat capital as a primary factor, as in the Heckscher-Ohlin-Samuelson theory of international trade. The former feature implies that the role of capital scarcity in generating exploitation and classes cannot be analysed, whereas the latter feature yields a theory of profit that is analogous to the theory of rent.

On the other hand, unlike in the classic literature, including Roemer's seminal contributions, we explicitly model competitive resource allocations as involving a dynamic structure of economic interactions and assume that individuals face an intertemporal optimisation programme. We provide a definition of equilibrium in this dynamic setting that generalises Roemer's [22, 23] static notion of reproducible solution and show that the latter is a temporary equilibrium notion which can be interpreted as a one-period feature of our general equilibrium concept.

2.1 Technology

An economy comprises a set of agents $\mathcal{N} = \{1, .., N\}$. A sequence of nonoverlapping generations exist, each living for T periods, where T can be either finite or infinite, and indexed by the date of birth kT, $k = 0, 1, 2, \ldots$. Let $\mathbb{R}$ be the set of real numbers and let $\mathbb{R}_+, \mathbb{R}_-$ be, respectively, the set of nonnegative and nonpositive real numbers.

Production technology is freely available to all agents, who can operate any activity in the production set P, which has elements of the form $\alpha = (-\alpha_l, -\underline{\alpha}, \overline{\alpha})$ where $\alpha_l \in \mathbb{R}_+$ is the *effective* labour input; $\underline{\alpha} \in \mathbb{R}^n_+$ are the inputs of the produced goods; and $\overline{\alpha} \in \mathbb{R}^n_+$ are the outputs of the n goods. Thus, elements of P are vectors in $\mathbb{R}^{2n+1}$. The net output vector arising from α is denoted as $\hat{\alpha} \equiv \overline{\alpha} - \underline{\alpha}$. Let $\mathbf{0}$ denote the null vector. The following assumptions on P hold throughout the paper.[2]

Assumption 0 (A0) P is a closed convex cone in $\mathbb{R}^{2n+1}$ and $\mathbf{0} \in P$.
Assumption 1 (A1) For all $\alpha \in P, \overline{\alpha} \geq \mathbf{0} \Rightarrow \alpha_l > 0$.
Assumption 2 (A2) For all $c \in \mathbb{R}^n_+, \exists \alpha \in P : \hat{\alpha} \geqq c$.
Assumption 3 (A3) For all $\alpha \in P$ and all $\alpha' \in \mathbb{R}_- \times \mathbb{R}^n_- \times \mathbb{R}^n_+, [\alpha' \leqq \alpha \Rightarrow \alpha' \in P]$.

A1 implies that labour is indispensable to produce any output. **A2** states that any non-negative commodity vector is producible as net output. **A3** is a standard *free disposal* condition.

A0–**A3** are quite general and include the standard production technologies discussed in mathematical Marxian economics as special cases. For example, the Leontief technology with a $n \times n$ non-negative input matrix A and a $1 \times n$ positive vector of labour inputs L is represented by

$$P_{(A,L)} \equiv \left\{\alpha \in \mathbb{R}_- \times \mathbb{R}^n_- \times \mathbb{R}^n_+ \mid \exists x \in \mathbb{R}^n_+ : \alpha \leqq (-Lx, -Ax, x)\right\}.$$

Given P, the set of activities feasible with k units of effective labour is:

$$P(\alpha_l = k) \equiv \{(-\alpha_l, -\underline{\alpha}, \overline{\alpha}) \in P \mid \alpha_l = k\};$$

$\partial P \equiv \{\alpha \in P \mid \nexists \alpha' \in P : \alpha' > \alpha\}$ is the frontier of P; and for any $c \in \mathbb{R}^n_+$, the set of activities that produce at least c as net output is:

$$\phi(c) \equiv \{\alpha \in P \mid \hat{\alpha} \geqq c\}.$$

[2]For all $x, y \in \mathbb{R}^n$, $x \geqq y$ if and only if $x_i \geqq y_i$ $(i = 1, \ldots, n)$; $x \geq y$ if and only if $x \geqq y$ and $x \neq y$; $x > y$ if and only if $x_i > y_i$ $(i = 1, \ldots, n)$.

2.2 Agents

In the economy, agents produce, consume, and trade labour. On the production side, they can either sell their labour-power or hire workers to work on their capital, or they can be self-employed and work on their own assets.

In every period t, $(p_t, w_t) \in \mathbb{R}^{n+1}_+ \setminus \{\mathbf{0}\}$ denotes the $1 \times (n+1)$ price vector that prevails in competitive markets. Let $\Delta \equiv \left\{(p,w) \in \mathbb{R}^{n+1}_+ \mid \sum_{i=1}^{n} p_i + w = 1\right\}$.

For all $\nu \in \mathscr{N}$, let $\varsigma^\nu > 0$ be agent ν's skill level. Then, for all $\nu \in \mathscr{N}$, in every t: $\alpha_t^\nu = \left(-\alpha_{lt}^\nu, -\underline{\alpha}_t^\nu, \overline{\alpha}_t^\nu\right) \in P$ is the production process operated with ν's own capital and labour, where $\alpha_{lt}^\nu = \varsigma^\nu a_{lt}^\nu$ and a_{lt}^ν is the labour *time* expended by ν; $\beta_t^\nu = \left(-\beta_{lt}^\nu, -\underline{\beta}_t^\nu, \overline{\beta}_t^\nu\right) \in P$ is the production process operated by hiring others; $\gamma_t^\nu = \varsigma^\nu l_t^\nu$ is ν's effective labour supply, where l_t^ν is the labour *time* supplied by ν on the market. At any t, $\lambda_t^\nu = \left(a_{lt}^\nu + l_t^\nu\right)$ is the total amount of labour time expended by ν and $\Lambda_t^\nu = \alpha_{lt}^\nu + \gamma_t^\nu = \varsigma^\nu \lambda_t^\nu$ is the total amount of effective labour performed by ν, either as a self-employed producer or working for some other agent. Further, for all $\nu \in \mathscr{N}$, $s_t^\nu \in \mathbb{R}^n$ is the vector of net savings and $\omega_t^\nu \in \mathbb{R}^n_+$ is the vector of productive endowments, where ω_{kT}^ν denotes the endowments inherited when born in kT.

As in Roemer [22, 23], the time structure of production is explicitly considered and production activities are financed with current wealth. Agent ν's wealth, at the beginning of t, is given by $W_t^\nu = p_{t-1}\omega_t^\nu$: this is fixed at the end of $t-1$ given previous savings decisions s_{t-1} and market prices p_{t-1}. At the beginning of t, ν uses W_t^ν to purchase a vector of capital goods $\underline{\alpha}_t^\nu + \underline{\beta}_t^\nu$ at prices p_{t-1} and any wealth left can be used to purchase a vector of goods $\delta_t^\nu \in \mathbb{R}^n_+$ that can be sold on the market at the end of t.

On the consumption side, for each agent ν, $C \subseteq \mathbb{R}^n_+$ is the consumption set, $c_t^\nu \in C$ is the consumption vector at t, and total labour hours expended cannot exceed the endowment which is normalised to one. Agent ν's welfare is given by a monotonic function $u^\nu : C \times [0,1] \to \mathbb{R}_+$, which is increasing in consumption and decreasing in labour time.

For any t, let $\Omega_t = \left(\omega_t^1, \omega_t^2, \ldots, \omega_t^N\right)$; $E\left(P, \mathscr{N}, C, (u^\nu)_{\nu\in\mathscr{N}}, (\varsigma^\nu)_{\nu\in\mathscr{N}}, \Omega_{kT}\right)$ denotes the economy with technology P, agents $\mathscr{N}$, consumption set C, welfare functions $(u^\nu)_{\nu\in\mathscr{N}}$, skills $(\varsigma^\nu)_{\nu\in\mathscr{N}}$, and productive endowments Ω_{kT}. The universal class of all such convex cone economies is $\mathscr{E}$.

Let $c^\nu = \{c_t^\nu\}_{t=kT}^{(k+1)T-1}$ be ν's lifetime consumption plan; and likewise for $\alpha^\nu, \beta^\nu, \gamma^\nu, \delta^\nu, s^\nu$, and ω^ν. Let $(\mathbf{p}, \mathbf{w}) = \{(p_t, w_t)\}_{t=kT}^{(k+1)T-1}$ be the path of price vectors during the lifetime of a generation. Let $\xi^\nu = (\alpha^\nu, \beta^\nu, \gamma^\nu, \delta^\nu, c^\nu, s^\nu)$ denote a generic intertemporal plan for ν, with $\xi_t^\nu = \left(\alpha_t^\nu, \beta_t^\nu, \gamma_t^\nu, \delta_t^\nu, c_t^\nu, s_t^\nu\right)$ at any t. Let $0 < \rho \leq 1$ be the time preference factor. Given $(\mathbf{p}, \mathbf{w})$, each agent ν chooses ξ^ν to maximise welfare subject to the constraint that in every t: (1) income is sufficient for consumption and savings; (3) production activities, consumption choices and labour performed are feasible; and (4) the dynamics of capital is determined by net savings. Furthermore, (2) wealth must be sufficient for production plans and any

wealth not used productively is carried over to the end of the period. Finally, (5) reproducibility requires resources not to be depleted; in particular, generation k is constrained to bequeath at least as many resources as they inherited. Formally:

$$MP^{\nu}\colon V(\omega_{kT}^{\nu}) = \max_{\xi^{\nu}} \sum_{t=kT}^{(k+1)T-1} \rho^{t} u^{\nu}\left(c_t^{\nu}, \lambda_t^{\nu}\right),$$

subject to (for all $t = kT, \ldots, (k+1)\,T-1$):

$$\left[p_t \overline{\alpha}_t^{\nu}\right] + \left[p_t \overline{\beta}_t^{\nu} - w_t \beta_{lt}^{\nu}\right] + w_t \gamma_t^{\nu} + p_t \delta_t^{\nu} \geqq p_t c_t^{\nu} + p_t \omega_{t+1}^{\nu}, \tag{1}$$

$$p_{t-1}\delta_t^{\nu} + p_{t-1}\left(\underline{\alpha}_t^{\nu} + \underline{\beta}_t^{\nu}\right) = p_{t-1}\omega_t^{\nu}, \tag{2}$$

$$\alpha_t^{\nu}, \beta_t^{\nu} \in P,\ \left(c_t^{\nu}, \lambda_t^{\nu}\right) \in C \times [0,1], \tag{3}$$

$$\omega_{t+1}^{\nu} = \omega_t^{\nu} + s_t^{\nu}, \tag{4}$$

$$\omega_{(k+1)T}^{\nu} \geqq \omega_{kT}^{\nu}. \tag{5}$$

MP^{ν} generalises similar programmes in Roemer [22, 23]. As in standard microeconomics, agents are not assumed to be "agents of capital" or to produce for production's own sake: they are endowed with general preferences over consumption and leisure. However, following Roemer [22, 23], and unlike in the standard approach, MP^{ν} explicitly incorporates the simultaneous role of economic actors as consumers and producers—so that no separate consideration of firms is necessary,—*and* the time structure of the production process. Thus, at the beginning of each t, agent ν supplies γ_t^{ν} on the labour market and uses her wealth W_t^{ν} to purchase goods $\underline{\alpha}_t^{\nu} + \underline{\beta}_t^{\nu} + \delta_t^{\nu}$ at prices p_{t-1}. The capital goods $\underline{\alpha}_t^{\nu} + \underline{\beta}_t^{\nu}$ are used to activate production by employing β_{lt}^{ν} units of labour, whereas δ_t^{ν} are carried over to the end of the period. Production then takes place and outputs appear at the end of t, when ν's proceedings from production are $p_t\left(\overline{\alpha}_t^{\nu} + \overline{\beta}_t^{\nu}\right)$ and wage earnings are $w_t\gamma_t^{\nu}$. Therefore, gross revenue at t is $p_t\left(\overline{\alpha}_t^{\nu} + \overline{\beta}_t^{\nu}\right) + w_t\gamma_t^{\nu} + p_t\delta_t^{\nu}$ which is used to pay $w_t\beta_{lt}^{\nu}$ to employees, and to purchase—at the current prices p_t—consumption goods c_t^{ν} and capital goods $\omega_{t+1}^{\nu} = \omega_t^{\nu} + s_t^{\nu}$ for next period's production.

Agents need to lay out in advance the capital necessary for production and can do so only by using their own wealth, which may be deemed restrictive. Two points should be noted here. First, as in Roemer [23, 25], this assumption rules out intertemporal credit markets and intertemporal trade *between agents*. Due to the possibility of saving, however, the model allows for intertemporal trade-offs in the allocation of labour and consumption goods *during an agent's life*, consistently with a dynamic setting in which agents' lives are divided into more than one period and this significantly generalises Roemer's models. Second, a credit market may be introduced but it would not change the main results (see Roemer [22, chapter 3], [23]).

Finally, our conclusions are robust to alternative specifications of the individual optimisation programme. All of the main insights continue to hold if MP^{ν} is reformulated by focusing on end-of-period prices p_t in (2), which generalises Veneziani [33]; or by letting the length of the production period tend to zero, so as to move to a continuous time setting, as in Veneziani [34].

2.3 Equilibrium

Let $c_t = \sum_{\nu=1}^{N} c_t^{\nu}$; and likewise for all other variables. For the sake of simplicity, let "all t" stand for "all $t = kT, \ldots, (k+1)T - 1$". Let $\mathscr{O}^{\nu}(\mathbf{p}, \mathbf{w}) \equiv \{\xi^{\nu} \text{ solves } MP^{\nu} \text{ at } (\mathbf{p}, \mathbf{w})\}$. The equilibrium concept can now be defined.

Definition 1 A *reproducible solution* (RS) for $E\,(P, \mathscr{N}, C, (u^{\nu})_{\nu \in \mathscr{N}}, (\varsigma^{\nu})_{\nu \in \mathscr{N}}, \Omega_{kT})$ is a price vector $(\mathbf{p}, \mathbf{w})$ and an associated set of actions such that :

(i) $\xi^{\nu} \in \mathscr{O}^{\nu}(\mathbf{p}, \mathbf{w})$, all ν;
(ii) $\hat{\alpha}_t + \hat{\beta}_t \geqq c_t + s_t$, all t;
(iii) $\underline{\alpha}_t + \underline{\beta}_t + \delta_t \leqq \omega_t$, all t;
(iv) $\beta_{lt} = \gamma_t$, all t;
(v) $\omega_{(k+1)T} \geqq \omega_{kT}$.

The equilibrium notion is standard. Condition (i) requires that every agent optimises. Conditions (ii) and (iii) are aggregate excess demand requirements. The former states that in every t there must be enough resources for consumption and saving plans, and it is equivalent to: $\overline{\alpha}_t + \overline{\beta}_t + \left(\omega_t - \underline{\alpha}_t - \underline{\beta}_t\right) \geqq c_t + (\omega_t + s_t)$, which states that, at the end of period t, the aggregate supply of resources available be at least as big as the aggregate demand for consumption and investment goods. The latter states that demand should not exceed supply in the produced inputs market and in every t there must be enough resources for production plans. Condition (iv) imposes labour market clearing in every t.

Condition (v) is the intertemporal *reproducibility condition*, which requires that every generation leave to the following at least as many resources as they have inherited. This significantly relaxes the analogous reproducibility condition implicit in Roemer's [22, 23] static models without savings in which $\omega_{t+1} \geqq \omega_t$ automatically follows from conditions (ii) and (iii). In a finite horizon model, condition (v) can be seen as a simple fairness and sustainability condition analogous to the constraints often imposed in optimal Ramsey growth problems (see, for example, Morishima [16, Chapter 13]). Formally, this condition is consistent with the transversality condition which is necessary in an infinite horizon model.

In what follows, we devote special attention to the subset of stationary equilibria in which prices and actions remain constant over time:

Definition 2 A *stationary reproducible solution* (SRS) for $E(P, \mathscr{N}, C, (u^{\nu})_{\nu \in \mathscr{N}}, (\varsigma^{\nu})_{\nu \in \mathscr{N}}, \cdot)$ is a price vector $(\mathbf{p}, \mathbf{w})$, an associated set of actions $(\xi^{\nu})_{\nu \in \mathscr{N}}$, and a

profile of capital stocks $\Omega^* = \left(\omega^{*1}, \omega^{*2}, \ldots, \omega^{*N}\right)$ such that $\langle (\mathbf{p}, \mathbf{w}), (\xi^\nu)_{\nu \in \mathscr{N}} \rangle$ is a RS for $E(P, \mathscr{N}, C, (u^\nu)_{\nu \in \mathscr{N}}, (\varsigma^\nu)_{\nu \in \mathscr{N}}, \Omega^*)$ with:

(1) $(p_t, w_t) = (p_{t+1}, w_{t+1})$, all t;
(2) for any $\nu \in \mathscr{N}$, $\xi^\nu \in \mathscr{O}^\nu(\mathbf{p}, \mathbf{w})$ is such that $\xi_t^\nu = \xi_{t+1}^\nu$ and $s_t^\nu = \mathbf{0}$, all t.

In order to analyse the existence and properties of SRSs, it suffices to consider a stationary price vector $(\mathbf{p}, \mathbf{w})$ with $(p_t, w_t) = (p_{t+1}, w_{t+1}) = (p, w)$ for all t. In this case, programme MP^ν reduces to the following:

$$MP^\nu\colon V(\omega_{kT}^\nu) = \max_{\xi^\nu} \sum_{t=kT}^{(k+1)T-1} \rho^t u^\nu \left(c_t^\nu, \lambda_t^\nu\right),$$

subject to (for all t):

$$\left[p\left(\overline{\alpha}_t^\nu - \underline{\alpha}_t^\nu\right)\right] + \left[p\left(\overline{\beta}_t^\nu - \underline{\beta}_t^\nu\right) - w\beta_{lt}^\nu\right] + w\gamma_t^\nu \geqq pc_t^\nu + ps_t^\nu$$

$$p\left(\underline{\alpha}_t^\nu + \underline{\beta}_t^\nu\right) \leqq p\omega_t^\nu,$$

$$\alpha_t^\nu, \beta_t^\nu \in P,\ \left(c_t^\nu, \lambda_t^\nu\right) \in C \times [0,1],$$

$$\omega_{t+1}^\nu = \omega_t^\nu + s_t^\nu,$$

$$\omega_{(k+1)T}^\nu \geqq \omega_{kT}^\nu.$$

Further, noting that at a SRS, $\max_{\alpha_t' \in P} \frac{p\hat{\alpha}_t' - w\alpha_{lt}'}{p\underline{\alpha}_t'} = \frac{1-\rho}{\rho}$ all t, the set $\Delta(\rho) \equiv \left\{(p', w') \in \Delta \mid p'\left(\overline{\alpha} - \rho^{-1}\underline{\alpha}\right) - w'\alpha_l \leqq 0 \text{ for all } \alpha \in P\right\}$ is compact and convex. Then, for any given $(p, w) \in \Delta(\rho)$, the individual optimisation programme can be further reduced to the following:

$$MP^\nu\colon \max_{\alpha^\nu, \beta^\nu \in P,\ (c^\nu, \lambda^\nu) \in C \times [0,1]} u^\nu(c^\nu, \lambda^\nu)$$

subject to

$$[p(\overline{\alpha}^\nu - \underline{\alpha}^\nu)] + \left[p\left(\overline{\beta}^\nu - \underline{\beta}^\nu\right) - w\beta_l^\nu\right] + w\gamma^\nu \geqq pc^\nu$$

$$p\left(\underline{\alpha}^\nu + \underline{\beta}^\nu\right) \leqq p\omega^\nu.$$

The set of solutions of the reduced programme is denoted by $\mathscr{O}^\nu(p, w)$.

3 UE Exploitation: An Axiomatic Approach

In the UE approach, exploitative relations are characterised by systematic differences between the labour that agents contribute to the economy and the labour "received" by them, which is given by the amount of labour contained, or embodied, in some relevant consumption bundle(s). Therefore, in order to define exploitation status, it is necessary *both* to select the relevant bundle(s) *and* to identify their labour content. In economies with heterogeneous optimising agents and a general technology, neither choice is obvious, and various definitions have, in fact, been proposed.

The question, then, is which approach best captures the key insights of UE exploitation theory among those proposed, but also in the space of all conceivable definitions. In the literature, the proposal of alternative definitions has sometimes appeared as a painful process of adjustment of the theory to anomalies and counterexamples. In order to answer the question, and discriminate among a potentially infinite number of definitions, the axiomatic method pioneered by Yoshihara [40] seems more promising. An axiomatic approach suggests to start from first principles, thus explicitly identifying the class of suitable exploitation forms.

In his paper, Yoshihara [40] focuses on the *Class-Exploitation Corresponding Principle* (**CECP**; see Roemer [23]), which states that in equilibrium class membership and exploitation status emerge endogenously: the wealthy can rationally choose to belong to the capitalist class among other available options and become an exploiter, while the poor have no other option than being in the working class and are exploited. From this perspective, UE exploitative relations are relevant because they reflect unequal opportunities of life options, due to asset inequalities.

Under the classic definition by Okishio [19] and Morishima [18], **CECP** is proved as a formal theorem in simple Leontief production economies with rational agents [23, 33], but it does not hold in more general production economies [23, 40]. In contrast, Yoshihara [40] formulates **CECP** as an axiom capturing a key insight of UE exploitation theory on a generic feature of capitalist economy, and introduces a *domain axiom* that defines the class of admissible exploitation forms. Then, he derives a necessary and sufficient condition to identify the UE definitions that satisfy the domain axiom, and under which **CECP** holds in any general convex production economy [40, Theorem 2]. This condition allows us to test which UE definition within the appropriate domain preserves **CECP** in general. Interestingly, among the main definitions in the literature, an extension of the "New Interpretation" form of exploitation [4, 5, 10, 11] is the only one that passes the test [40, Corollaries 1–4].

In this paper, we focus more specifically on exploitation, rather than class. This section discusses a recent axiomatic analysis of UE exploitation theory based on Veneziani and Yoshihara [36]. An axiom called the *Profit-Exploitation Corresponding Principle* (**PECP**), is presented which states that in equilibrium, the existence of positive profits corresponds to the social condition that every employed propertyless agent is exploited. This axiom is consistent with the traditional Marxian view that profits represent capitalist relations of production in which capitalists

appropriate social surplus produced from the social labour of (propertyless) workers. But the nexus between profits, asset inequalities and the distribution of labour is relevant beyond Marxian theory. We then characterise the class of UE exploitation-forms which satisfy **PECP** and a weak domain axiom.

In what follows, we focus on stationary RSs and examine UE exploitation and profits associated with the one-period allocations generated at a SRS. For the sake of notational simplicity, we denote SRSs simply by (p, w) and any general convex economy as described in Sect. 2 by E.

3.1 The Main Definitions

In this subsection, we introduce the main definitions of UE exploitation in the literature, suitably extended to economies with heterogeneous skills. Given any definition of exploitation, let $\mathcal{N}^{ter} \subseteq \mathcal{N}$ and $\mathcal{N}^{ted} \subseteq \mathcal{N}$ denote, respectively, the set of exploiters and the set of exploited agents at a given allocation, where $\mathcal{N}^{ter} \cap \mathcal{N}^{ted} = \varnothing$.

The classic and perhaps best known definition was provided by Okishio [19] in a simple Leontief economy, and was later generalised to the von Neumann economy by Morishima [18]. Formally, for all $c \in \mathbb{R}^n_+$, the minimum amount of (effective) labour necessary to produce c as net output is:

$$l.v.(c) \equiv \min \{\alpha_l \mid \alpha = (-\alpha_l, -\underline{\alpha}, \overline{\alpha}) \in \phi(c)\}.$$

By **A0**∼**A2**, $l.v.(c)$ is well-defined and is positive whenever $c \neq \mathbf{0}$ [22]. Then:

Definition 3 (Morishima [18]) Consider any $E \in \mathscr{E}$. For any $\nu \in \mathcal{N}$, who supplies Λ^ν and consumes $c^\nu \in \mathbb{R}^n_+$, $\nu \in \mathcal{N}^{ted}$ if and only if $\Lambda^\nu > l.v.(c^\nu)$ and $\nu \in \mathcal{N}^{ter}$ if and only if $\Lambda^\nu < l.v.(c^\nu)$.

Definition 3 is consistent with classical Marxian theory, in that UE exploitation is defined based upon the labour value of labour power, which is defined independently of price information. However, as argued by Roemer [23], a definition of exploitation independent of price information gives rise to counterintuitive results. Thus, a number of alternative definitions have been proposed, in which price information plays a crucial role.

Consider Roemer's [23, chapter 5] definition. Given a price vector (p, w), the set of activities that yield the maximum profit rate is:

$$P^\pi(p, w) \equiv \left\{ \alpha \in \arg\max_{\alpha' \in P} \frac{p\hat{\alpha}' - w\alpha'_l}{p\underline{\alpha}'} \right\},$$

and the set of profit-rate-maximising activities that produce at least $c \in \mathbb{R}^n_+$ as net output is:

$$\phi(c; p, w) \equiv \{\alpha \in P^\pi(p, w) \mid \hat{\alpha} \geqq c\}.$$

For all $c \in \mathbb{R}^n_+$, the minimum amount of (effective) labour necessary to produce c as net output among profit-rate-maximising activities is:

$$l.v.(c; p, w) \equiv \min\{\alpha_l \mid \alpha = (-\alpha_l, -\underline{\alpha}, \overline{\alpha}) \in \phi(c; p, w)\}.$$

Again, $l.v.(c; p, w)$ is well defined at SRSs and is positive for all $c \neq \mathbf{0}$. Then:

Definition 4 (Roemer [23]) Consider any $E \in \mathscr{E}$. Let (p, w) be a SRS for E. For any $\nu \in \mathscr{N}$, who supplies Λ^ν and consumes c^ν, $\nu \in \mathscr{N}^{ted}$ if and only if $\Lambda^\nu > l.v.(c^\nu; p, w)$ and $\nu \in \mathscr{N}^{ter}$ if and only if $\Lambda^\nu < l.v.(c^\nu; p, w)$.

Finally, we analyse a definition recently proposed by Yoshihara and Veneziani [41, 42] and Yoshihara [40]. For any $p \in \mathbb{R}^n_+$ and $c \in \mathbb{R}^n_+$, let $\mathscr{B}(p, c) \equiv \{x \in \mathbb{R}^n_+ \mid px = pc\}$ be the set of bundles that cost exactly as much as c at prices p. Let $\alpha^{p,w} \equiv \sum_{\nu=1}^N (\alpha^\nu + \beta^\nu)$ denote the aggregate equilibrium production activity at a SRS (p, w) for E.

Definition 5 Consider any $E \in \mathscr{E}$. Let (p, w) be a SRS for E with aggregate production activity $\alpha^{p,w}$. For all $c \in \mathbb{R}^n_+$ with $pc \leqq p\hat{\alpha}^{p,w}$, let $\tau^c \in [0, 1]$ be such that $\tau^c \hat{\alpha}^{p,w} \in \mathscr{B}(p, c)$. The *labour embodied in* c at $\alpha^{p,w}$ is $\tau^c \alpha_l^{p,w}$.

As in Roemer's [23] approach, in Definition 5 the labour content of a bundle can be identified only if the price vector is known. Yet social relations play a more central role, because the definition of labour content requires a prior knowledge of the social reproduction point, and labour content is explicitly linked to the redistribution of total social labour, which corresponds to the total labour content of national income. Then:

Definition 6 Consider any $E \in \mathscr{E}$. Let (p, w) be a SRS for E with aggregate production activity $\alpha^{p,w}$. For any $\nu \in \mathscr{N}$, who supplies Λ^ν and consumes c^ν, let τ^{c^ν} be defined as in Definition 5. Then $\nu \in \mathscr{N}^{ted}$ if and only if $\Lambda^\nu > \tau^{c^\nu} \alpha_l^{p,w}$ and $\nu \in \mathscr{N}^{ter}$ if and only if $\Lambda^\nu < \tau^{c^\nu} \alpha_l^{p,w}$.

Definition 6 is conceptually related to the "New Interpretation" (NI) developed by Duménil [4, 5] and Foley [10, 11]: for all $\nu \in \mathscr{N}$, τ^{c^ν} represents ν's share of national income, and so $\tau^{c^\nu} \alpha_l^{p,w}$ is the share of social labour that ν receives by earning income barely sufficient to buy c^ν. Then, as in the NI, the notion of exploitation is related to the production and distribution of national income and social labour.

3.2 Labour Exploitation

In this section, a general domain condition is presented which captures the core insights of UE exploitation theory shared by all of the main approaches.

Let $\mathscr{W} \equiv \{\nu \in \mathscr{N} \mid \omega^{\nu} = \mathbf{0}\}$. The set $\mathscr{W}$ is of focal interest in exploitation theory: if *any* agents are exploited, then those with no initial endowments should be among them, if they work at all. It is therefore opportune, from an axiomatic viewpoint, to focus on $\mathscr{W}$ in order to identify some minimum requirements that all UE definitions should satisfy.

Let $B(p, w\Lambda) \equiv \left\{c \in \mathbb{R}^n_+ \mid pc = w\Lambda\right\}$ be the set of consumption bundles that can be *just* afforded, at prices p, by an agent in $\mathscr{W}$, who supplies Λ units of labour at a wage w. We can now introduce the domain axiom.

Labour Exploitation (LE): *Consider any $E \in \mathscr{E}$. Let (p, w) be a SRS for E. Given any definition of exploitation, the set $\mathscr{N}^{ted} \subseteq \mathscr{N}$ should have the following property at (p, w): there exists a profile $(c_e^1, \ldots, c_e^{|\mathscr{W}|})$ such that for any $\nu \in \mathscr{W}$, $c_e^{\nu} \in B(p, w\Lambda^{\nu})$ and for some $\alpha^{c_e^{\nu}} \in \phi\left(c_e^{\nu}\right) \cap \partial P$ with $\hat{\alpha}^{c_e^{\nu}} \succ c_e^{\nu}$:*

$$\nu \in \mathscr{N}^{ted} \Leftrightarrow \alpha_l^{c_e^{\nu}} < \Lambda^{\nu}.$$

LE requires that, at any SRS, the exploitation status of each propertyless worker $\nu \in \mathscr{W}$ be characterised by identifying a nonnegative vector c_e^{ν}, that may be defined an *exploitation reference bundle* (ERB). The ERB must be technically feasible and on ν's budget line, and it identifies the amount of labour that ν receives, $\alpha_l^{c_e^{\nu}}$. Thus, if $\nu \in \mathscr{W}$ supplies Λ^{ν}, and Λ^{ν} is more than $\alpha_l^{c_e^{\nu}}$, then ν is regarded as contributing more labour than ν receives. According to **LE**, all such agents belong to $\mathscr{N}^{ted}$.

In UE theory, the exploitation status of agent ν is determined by the difference between the amount of labour that ν "contributes" to the economy, and the amount she "receives". As a domain condition for the admissible class of exploitation-forms, **LE** provides some minimal, key restrictions on the definition of the amount of labour that a theoretically relevant subset of agents contributes *and* the amount they receive.

According to **LE**, the former quantity is given by the *effective* labour, Λ^{ν}, rather than the labour *time*, λ^{ν}, performed by the agent. This is because, as a domain condition for UE exploitation, **LE** aims to capture the key intuitions common to all of the main approaches: not only is a focus on effective labour the natural extension of all of the classic definitions in the Okishio-Morishima-Roemer tradition, it is also the standard approach in the literature on exploitation in economies with heterogeneous labour and skills (see, e.g., [6, 13]).[3] Moreover, by focusing on Λ^{ν}, **LE** incorporates the key normative intuition of what may be called the "*contribution*

[3]For a slightly different, but related approach based on the notion of "abstract labour", see Fleurbaey [8, section 8.5].

view" of exploitation theory: a UE exploitation-free allocation coincides with the *proportional solution*, a well-known fair allocation rule whereby every agent's income is proportional to her contribution to the economy [29]. Proportionality is a strongly justified normative principle, whose philosophical foundations can be traced back to Aristotle [14] and which can be justified in terms of the Kantian categorical imperative [28].

LE imposes even weaker restrictions on the amount of labour received by the agents in $\mathscr{W}$. First, the amount of labour that $\nu \in \mathscr{W}$ receives depends on her income, or more precisely, it is determined in equilibrium by some reference bundle that ν can purchase. In the standard approaches, the ERB corresponds to the bundle actually chosen by the agent. In Definitions 3 and 4, for example, $c_e^\nu \equiv c^\nu \in B\left(p, w\Lambda^\nu\right)$. Indeed, as noted by an anonymous referee, it may be argued that **LE** should explicitly require that $c_e^\nu = c^\nu$, for UE exploitation status should be defined based only on the information emerging from the *actual* exchange process. But this subjectivist view is not uncontroversial. Following the standard Marxian approach, for example, one may insist that exploitation status depend on *productive* decisions, and not on possibly arbitrary *consumption* decisions. From this viewpoint, agents who are identical in all characteristics except their consumption choices should have the same exploitation status.

At any rate, we need not adjudicate this issue. For our aim is to provide a *weak* domain condition that is shared by all of the main approaches. Therefore, **LE** does not rule out the possibility that $c_e^\nu = c^\nu$, but it does not impose it as a requirement and it only requires that the ERB be *potentially* affordable. Thus, in Definition 6, given any (p, w) with aggregate production activity $\alpha^{p,w}$, $c_e^\nu \equiv \tau^{c^\nu} \cdot \hat{\alpha}^{p,w} \in B\left(p, w\Lambda^\nu\right)$, where $\tau^{c^\nu} = \frac{pc^\nu}{p\hat{\alpha}^{p,w}}$.

Second, the amount of labour associated with the ERB—and thus "received" by an agent—is related to production conditions: **LE** states that the ERB be technologically feasible as net output, and its labour content is the amount of labour socially necessary to produce it. Observe that **LE** requires that the amount of labour associated with each ERB be uniquely determined with reference to production conditions, but it does not specify how such amount should be chosen, and there may be many (efficient) ways of producing c_e^ν, and thus of determining $\alpha_l^{c_e^\nu}$. In Definition 3, $\alpha^{c_e^\nu} \in \arg\min\left\{\alpha_l \mid \alpha \in \phi\left(c_e^\nu\right)\right\}$; in Definition 4, $\alpha^{c_e^\nu} \in \arg\min\left\{\alpha_l \mid \alpha \in \phi\left(c_e^\nu; p, w\right)\right\}$; and in Definition 6, $\alpha^{c_e^\nu} \equiv \tau^{c^\nu}\alpha^{p,w}$, where $\tau^{c^\nu} = \frac{pc^\nu}{p\hat{\alpha}^{p,w}}$.

Finally, note that **LE** does *not* provide comprehensive conditions for the determination of exploitation status: it only focuses on a subset of agents and it imposes no restrictions on the set of exploiters $\mathscr{N}^{ter}$.[4]

In summary, **LE** represents an appropriate domain condition in exploitation theory: it is formally weak and incorporates some widely shared views on UE

[4]It is worth noting in passing that the vector c_e^ν in **LE** may be a function of (p, w) and that once c_e^ν is identified, the existence of $\alpha^{c_e^\nu}$ is guaranteed by **A2** and **A3**.

exploitation. Thus, although it is not trivial and not all definitions in the literature satisfy it, all of the major approaches do.[5] The next question, then, is how to discriminate among the various definitions satisfying **LE**.

3.3 The Profit-Exploitation Correspondence Principle

A key tenet of UE exploitation theory is the idea that profits are one of the main determinants of the existence of exploitation, and of inequalities in well-being freedom: profits represent the way in which capitalists appropriate social surplus and social labour. Therefore a general correspondence should exist between positive profits and the exploitation of at least the poorest segments of the working class. This is formalised in the next axiom.

Profit-Exploitation Correspondence Principle (PECP): *For any* $E \in \mathscr{E}$ *and any SRS for* E, (p, w), *with aggregate production activity* $\alpha^{p,w}$*:*

$$\left[p\hat{\alpha}^{p,w} - w\alpha_l^{p,w} > 0 \Leftrightarrow \mathscr{N}^{ted} \supseteq \mathscr{W}_+\right],$$

whenever $\mathscr{W}_+ \equiv \{v \in \mathscr{W} \mid \Lambda^v > 0\} \neq \varnothing$.

Observe that **PECP** is formulated without specifying *any* definition of exploitation: *whatever* the definition adopted, propertyless agents should be exploited if and only if profits are positive in equilibrium. The axiom is weak in that it only focuses on a subset of $\mathscr{N}$ and it is silent on the set of exploiters $\mathscr{N}^{ter}$. Further, **PECP** is fairly general, because it both applies to economies with a complex class structure, and allows for the possibility that propertyless workers in $\mathscr{W}_+$ are a *strict* subset of $\mathscr{N}^{ted}$. Note that the axiom focuses only on propertyless workers who perform some labour: this is theoretically appropriate, since the exploitation status of agents who do not engage in any economic activities is unclear. Finally, **PECP** allows for fairly general assumptions on agents and technology, including heterogeneous preferences and skills, a convex technology, and so on.

It may be objected that **PECP** should not be considered as a postulate. In mathematical Marxian economics, and in Marx's own work, the equivalence between positive profits and the existence of (aggregate) exploitation has been traditionally derived as a theoretical *result*, as in the literature on the *Fundamental Marxian Theorem* (**FMT**; see Okishio [19] and Morishima [18]). As such, the link between exploitation and profits should hold under some conditions but not others, which seems prima facie inconsistent with the logical status of a postulate.[6]

[5] Based on Flaschel's [7] notion of *actual labour values*, another definition can be derived which satisfies **LE**. Instead, the subjectivist notion of labour exploitation based on workers' preferences proposed by Matsuo [15] does not satisfy **LE**.

[6] We are grateful to two anonymous referees for bringing this issue to our attention.

This objection is not entirely compelling. Although the axiomatic approach has not been used explicitly in mathematical Marxian economics, the **FMT** has been de facto, albeit implicitly, considered as a key axiom in exploitation theory (and the same holds for the **CECP**). The central relevance of the **FMT** is suggested by its very name and it has been widely considered as "the core of [Marx's] economic theory" [18, p. 622] such that alternative definitions have been proposed and compared in the literature based on whether they preserved it (and the **CECP**). Roemer's interpretation of the **CECP** can indeed be extended to the **FMT**: although it is formally proved as a theorem, it defines the core of Marx's theory and thus "its epistemological status is as a postulate. We seek to construct models that allow us to prove it" [24, p. 270].

To consider **PECP** as a postulate is therefore consistent with the central theoretical role assigned to the relation between exploitation and profits in the literature. Indeed, if an impossibility result followed from the imposition of **PECP**, this would arguably raise serious questions about some of the key intuitions of UE exploitation theory. And this is particularly relevant given that the **PECP** is significantly weaker than the **FMT** in that it imposes no constraints in equilibria where $\mathscr{W}_+ = \varnothing$ and when equilibrium profits are zero it only requires that *some* propertyless agents not be exploited.

Theorem 1, however, characterises the non-empty class of exploitation-forms that satisfy **LE** and such that **PECP** holds.

Theorem 1 (Veneziani and Yoshihara [36]) *For any definition of labour exploitation satisfying **LE**, the following statements are equivalent for any $E \in \mathscr{E}$ and for any SRS (p, w) with aggregate production activity $\alpha^{p,w}$:*

*(1) **PECP** holds under this definition;*
(2) if $\pi^{\max} > 0$, then for each $\nu \in \mathscr{W}_+$, there exists $\alpha_\pi^\nu \in P\,(\alpha_l = \Lambda^\nu) \cap \partial P$ such that $\hat{\alpha}_\pi^\nu \in \mathbb{R}^n_+$, $p\hat{\alpha}_\pi^\nu > w\Lambda^\nu$, and $\left(\alpha_{\pi l}^\nu, \underline{\alpha}_\pi^\nu, \overline{\alpha}_\pi^\nu\right) \geqq \eta^\nu \left(\alpha_l^{c_e^\nu}, \underline{\alpha}^{c_e^\nu}, \overline{\alpha}^{c_e^\nu}\right)$ for some $\eta^\nu > 1$.

Theorem 1 can be interpreted as follows. **PECP** states that propertyless workers are exploited if and only if equilibrium profits are positive. According to **LE**, the exploitation status of propertyless agents is determined by identifying a profile of (affordable) reference bundles which must be producible with less than Λ^ν units of labour for all exploited workers. By Theorem 1, in every convex economy, **PECP** holds if and only if the existence of positive profits in equilibrium is also determined by identifying a profile of reference bundles $\left(\hat{\alpha}_\pi^\nu\right)_{\nu \in \mathscr{W}_+}$. According to condition (2), for all $\nu \in \mathscr{W}_+$, these reference bundles must be producible with a technically efficient process using Λ^ν units of labour, and must be such that they are not affordable by ν and dominate the ERBs if the maximum profit rate is positive.

Theorem 1 does not identify a unique definition that meets **PECP**, but rather a class of definitions satisfying condition (2). Yet Veneziani and Yoshihara [36, Corollary 1] show that it has surprising implications concerning the main approaches in exploitation theory. For there are economies in which for all $\nu \in \mathscr{W}_+$, condition (2)

is never satisfied, if $\alpha^{c_e^v}$ is given by Definition 3 or 4 and so the **PECP** does not hold. In contrast, Definition 6 satisfies condition (2), and thus **PECP** holds *for all* $E \in \mathscr{E}$ *and all* SRS (p, w).

Methodologically, Theorem 1 suggests that an axiomatic analysis provides interesting insights and has relevant implications. This conclusion is far from trivial: as noted by an anonymous referee, one may doubt that the definition of exploitation requires an axiomatic analysis and argue that it would be more interesting to use axioms to justify a measure of the *degree* of exploitation. Yet, as shown above, as soon as the simplest polarised, two-class economies with restrictive assumptions on preferences and technology are abandoned, different definitions of UE exploitation have very different properties and incorporate different normative and positive insights. An axiomatic analysis of the *definition* of exploitation is therefore useful, if not necessary, in order to adjudicate the possible alternative views, before one can actually tackle the issue of the *degree* of exploitation.

Theorem 1 provides a demarcation line (condition 2) by which one can test which of infinitely many potential definitions preserves the relation between exploitation and profits in capitalist economies. Indeed, it characterises the class of definitions which are *coherent*, in the sense of preserving such relation regardless of the complexity of economic models. Note that Theorem 1 immediately implies that in simple Leontief economies with homogeneous agents, any definition of UE exploitation within the domain given by **LE** satisfies **PECP**. Yet in more complex economies, many of the definitions proposed in the literature violate **PECP**. Rather than concluding that there exists no general relation between exploitation and profits in capitalist economies, it seems more apt to consider these definitions as *incoherent*, or at least as not being robust. For both the Leontief model and the more general economies analysed in this paper represent competitive market economies with differential ownership of productive assets, and the differences between the two do not reflect different stages of development of capitalist societies, or differences in degrees of income disparity and social productivity, or in uneven power relations in the capitalist production process. Hence, there is no reason why the basic implications of exploitative social relations should vary according to the technical complexity of the economic models.

Substantively, the above arguments, and other recent axiomatic analyses, provide significant support to Definition 6, as the appropriate definition of UE exploitation. Theorem 1 proves that, unlike the main competing definitions, the NI preserves one of the key insights of classic exploitation theory.

Actually, not only does Theorem 1 establish that the set of definitions that preserve the **PECP** is not empty: if Definition 6 is adopted, the existence of profits is synonymous with the exploitation of *labour*. A traditional objection moved against the **FMT** in the Okishio-Morishima-Roemer approach is that the existence of profits is equivalent to the productiveness of the economy, which in turn is equivalent to the "exploitation" of any commodity (see the *Generalised Commodity Exploitation Theorem*, Roemer [23]), which raises doubts on the significance of the **FMT**. Yoshihara and Veneziani [43] have proved that the NI captures exploitation as

the unequal exchange of *labour*: unlike all other approaches, the existence of UE exploitation is not synonymous with the existence of any commodity exploitation.

Indeed, the NI may provide the foundations for a *general* theoretical framework that can deal with many unresolved issues in exploitation theory. Definition 6 can be easily extended to the general economies described in Sect. 2 and, as Veneziani and Yoshihara [36, Theorem 2] have shown, it is possible to determine the exploitation status of all agents and the whole exploitation structure of such general economies in equilibrium. Moreover, a robust relationship between profits and exploitation can be proved even at disequilibrium allocations [36, Theorem 3].

Definition 6 has a clear empirical content, for it is firmly anchored to the actual data of the economy and, unlike Definitions 3 and 4, it does not require information about all conceivable production techniques: only actual production decisions and the social allocation of labour, income and production activities matter. Indeed, as [42] have shown, it also satisfies a property of *Minimal objectivism* in that it does not rely on information about agents' subjective preferences and possibly arbitrary consumption decisions.

Perhaps more importantly, from a normative perspective, Definition 6 conceptualises exploitation as a *social* relation. Not only is the notion of exploitation related to the production and distribution of national income and social labour, as noted above. It can be proved that, unlike the main definitions in the literature, the NI identifies the existence of *exploitative relations*, in that some agents are exploited if and only if there is someone exploiting them [36, 41]).

Finally, Definition 6 identifies exploitative relations as characterised by inequalities in individual income/labour ratios—an important normative intuition of the UE approach which provides an interesting conceptual link with liberal egalitarian approaches.

4 The Dynamics of Exploitation

The previous section provides an axiomatic analysis of the *distributive* aspects of UE exploitation. Yet, a fundamental and contentious question in exploitation theory concerns precisely the role of distributive issues, on the one hand, and of relations of coercion, force, or power, on the other hand. At the most general level, A exploits B if and only if A takes unfair advantage of B. But do exploitative relations mainly, or uniquely, involve some (wrongful) characteristic of the structure of the interaction between A and B (such as asymmetric relations of power, force, coercion, etc.)? Or is exploitation mainly, or uniquely, concerned with some form of (wrongful) inequality (in asset ownership, labour exchanged, income, etc.)?

A path-breaking answer to these questions is provided by John Roemer's seminal theory [23, 25]. Roemer's key conclusion is that all relevant moral information is conveyed by the analysis of Differential Ownership of Productive Assets (henceforth, DOPA) and the resulting welfare inequalities. Notions of power or dominance are not relevant. On the one hand, Roemer rejects all approaches

based on domination at the point of production or coercion in the labour market. As Roemer put it, "Capitalism's necessary coercions are economic: ... it can substantially rid itself ... of extra-economic coercions, such as domination in the workplace ... Such a capitalism might be kinder and gentler, as they say, but it would not be socialism" [26, p. 386]. On the other hand, he proves that the labour market is not "intrinsically necessary for bringing about the Marxian phenomena of exploitation and class ... competitive markets and [DOPA] are the institutional culprits in producing exploitation and class" [23, p. 93].

Consequently, Roemer has developed an alternative game theoretic approach that focuses on property relations, which aims to generalise Marxian exploitation "in terms of the institutional variation permitted" [24, p. 256] and to capture its essential normative content, which is interpreted as requiring an egalitarian distribution of resources in the external world.

Roemer is effective in criticising approaches that focus on domination and direct coercion, and in stressing the relevance of distributive issues. It is however unclear that weaker forms of asymmetric relations between agents can, or indeed should be ruled out. Concerning Roemer's philosophical argument, Veneziani [34] has shown that purely distributive approaches to exploitation have too impoverished an informational basis to capture exploitative relations and to distinguish exploitation from other forms of injustice, or wrongs. Roemer's own game theoretic approach somewhat paradoxically casts doubts on the idea that relations of power *should* be ruled out.[7]

Perhaps more importantly, it is unclear that Roemer's formal argument convincingly establishes that exploitation *can* be reduced to a focus on DOPA. For "The economic problem for Marx, in examining capitalism, was to explain the *persistent* accumulation of wealth by one class and the *persistent* impoverishment of another, in a system characterized by voluntary trade" [23, p. 6], italics added). Roemer's models, however, are essentially static in that there are no intertemporal trade-offs, and so they are not suitable for analysing the persistence of exploitation in a capitalist economy.

In this section, based on the general model set up in Sect. 2 above, we survey and extend some recent contributions that analyse exploitation in an intertemporal context [33, 34]. A dynamic generalisation of Roemer's [23] subsistence economies is analysed in order to assess the relevance of DOPA, focusing on its role in generating exploitation as a persistent feature of a competitive economy with savings and a variable distribution of productive assets. We analyse subsistence economies because this allows us to examine the role of DOPA in a context where capital scarcity persists. In fact, the results obtained in Roemer's static economies depend on differential ownership of *scarce* productive assets [30] and it is not too surprising that exploitation may disappear when accumulation is allowed [3]. Moreover, Roemer's main conclusions do not depend on accumulation. On the

[7] In later writings, Roemer himself has acknowledged the limits of purely distributive definitions. See, for example, Roemer [26] and, for a discussion, Veneziani [34].

contrary, one of his key results is precisely that "exploitation emerges logically prior to accumulation" [24, p. 264].

4.1 Subsistence Economies: Equilibrium

A subsistence economy is a special case of the economies analysed in Sect. 2 in which agents are endowed with identical preferences and skills, and wish to minimise labour time subject to earning enough to purchase a given subsistence bundle. Formally, a convex economy $E\left(P, \mathscr{N}, C, (u^\nu)_{\nu\in\mathscr{N}}, (\varsigma^\nu)_{\nu\in\mathscr{N}}, \Omega_{kT}\right)$ is a *subsistence economy* if for all $\nu \in \mathscr{N}$: (i) there exists a bundle $b \in \mathbb{R}^n_+ \setminus \{\mathbf{0}\}$ which must be consumed in order to survive in each period, so that $C = C_b \equiv \left\{c \in \mathbb{R}^n_+ \mid c \geqq b\right\}$; (ii) $u^\nu(c,\lambda) = u_b(c,\lambda) \equiv 1-\lambda$, for all $(c,\lambda) \in C_b \times [0,1]$, and (iii) $\varsigma^\nu = 1$. Denote a subsistence economy by a list $E\left(P, \mathscr{N}, C_b, u_b, 1, \Omega_{kT}\right)$ or, as a shorthand notation, $E_b(\Omega_{kT})$.

In any $E_b(\Omega_{kT})$, the individual optimisation programme MP^ν is a special case of the general programme in Sect. 2.2, where the objective function is $\sum_{t=kT}^{(k+1)T-1} -\rho^t \Lambda^\nu_t$ and in each period $c^\nu_t = b$, without loss of generality. Accordingly, Definition 1 is slightly revised.

Definition 7 A *reproducible solution* (RS) for $E_b(\Omega_{kT})$ is a price vector $(\mathbf{p}, \mathbf{w})$ and an associated set of actions such that :

(i) $\xi^\nu \in \mathscr{O}^\nu(\mathbf{p}, \mathbf{w})$, all ν;
(ii) $\hat{\alpha}_t + \hat{\beta}_t \geqq Nb + s_t$, all t;
(iii) $\underline{\alpha}_t + \underline{\beta}_t + \delta_t \leqq \omega_t$, all t;
(iv) $\beta_{lt} = \overline{\gamma}_t$, all t;
(v) $\omega_{(k+1)T} \geqq \omega_{kT}$.

In what follows, unless otherwise stated, only *non-trivial* RS's are considered in which *some* production takes place in every period. Given any $(\mathbf{p}, \mathbf{w})$, let $\pi^{\max}_t \equiv \max_{\alpha\in P} \frac{p_t\overline{\alpha} - p_{t-1}\underline{\alpha} - w_t\alpha_l}{p_{t-1}\underline{\alpha}}$: at a non-trivial RS, it must be $\pi^{\max}_t \geqq \max_i \frac{p_{it}-p_{it-1}}{p_{it-1}}$, all t. For if $\pi^{\max}_t < \max_i \frac{p_{it}-p_{it-1}}{p_{it-1}}$, then at the solution to MP^ν, $\alpha^\nu_t + \beta^\nu_t = \mathbf{0}$ holds for all $\nu \in \mathscr{N}$. Therefore, for all $\nu \in \mathscr{N}$ who work in equilibrium, we can set $\delta^\nu_t = 0$ without loss of generality.

In order to avoid uninteresting technicalities, following Roemer [23, 25] we assume that agents who can reproduce themselves without working use the amount of wealth strictly necessary to obtain their subsistence bundle b.

Non Benevolent Capitalists (NBC): *If agent ν has a solution to MP^ν with $\Lambda^\nu_t = 0$, all t, then ν chooses ξ^ν to satisfy* $(1+\pi^{\max})\, p_{t-1}\left(\underline{\alpha}^\nu_t + \underline{\beta}^\nu_t\right) = p_t b + p_t\omega^\nu_{t+1}$ *at each t.*

Lemma 1 states that at a RS, at all t, the revenues constraint binds for all agents and, if the equilibrium profit rate is positive, the wealth constraint binds, for all ν who work.

Lemma 1 *Let* $(\mathbf{p}, \mathbf{w})$ *be a RS for* $E_b(\Omega_{kT})$. *Then:*

(i) under ***NBC****,* $\left[p_t \overline{\alpha}_t^{\nu}\right] + \left[p_t \overline{\beta}_t^{\nu} - w_t \beta_{lt}^{\nu}\right] + w_t \gamma_t^{\nu} = p_t b + p_t \omega_{t+1}^{\nu}$, *all* t, ν;

(ii) if $\pi_t^{\max} > 0$ *all* t, *and* $\sum_{t=kT}^{(k+1)T-1} \Lambda_t^{\nu} > 0$ *all* $\xi^{\nu} \in \mathscr{O}^{\nu}(\mathbf{p}, \mathbf{w})$, *then* $p_{t-1}\left(\underline{\alpha}_t^{\nu} + \underline{\beta}_t^{\nu}\right) = p_{t-1}\omega_t^{\nu}$, *all* t.

Lemma 2 derives some properties of equilibrium prices in every t.

Lemma 2 *Let* $(\mathbf{p}, \mathbf{w})$ *be a RS for* $E_b(\Omega_{kT})$. *Then, for all* t, *(i)* $p_t \overline{\alpha} - p_{t-1}\underline{\alpha} - w_t \alpha_l \geqq 0$, *for some* $\alpha \in P \setminus \{\mathbf{0}\}$; *(ii)* $p_t \geq \mathbf{0}$ *with* $p_t b > 0$; *and (iii)* $w_t > 0$.

The proofs of Lemmas 1 and 2 are straightforward and therefore omitted. Proposition 1 derives labour expended by each agent in each period.

Proposition 1 *Let* $(\mathbf{p}, \mathbf{w})$ *be a RS for* $E_b(\Omega_{kT})$. *Then, for all* t, ν, $\Lambda_t^{\nu} = \max\{0, \frac{p_t b - \left(1+\pi_t^{\max}\right) p_{t-1}\omega_t^{*\nu} + p_t \omega_{t+1}^{*\nu}}{w_t}\}$.

Proof If $\sum_{t=kT}^{(k+1)T-1} \Lambda_t^{\nu} > 0$ for all $\xi^{\nu} \in \mathscr{O}^{\nu}(\mathbf{p}, \mathbf{w})$, the result immediately follows from Lemma 1. If there is a ξ^{ν} such that $\sum_{t=kT}^{(k+1)T-1} \Lambda_t^{\nu} = 0$, then $\Lambda_t^{\nu} = 0$ and, by Lemma 1(i), $p_t b - \left(1 + \pi_t^{\max}\right) p_{t-1}\omega_t^{*\nu} + p_t \omega_{t+1}^{*\nu} \leqq 0$ all t. ■

Proposition 2 describes a dynamic property of equilibrium prices.

Proposition 2 *Let* $(\mathbf{p}, \mathbf{w})$ *be a RS for* $E_b(\Omega_{kT})$ *such that at all* t *there is some* $\nu \in \mathscr{N}$ *such that* $p_{t-1}\omega_t^{\nu} > 0$ *and* $\Lambda_t^{\nu} \in (0, 1)$. *Then,* $\frac{1}{w_t} = \rho(1 + \pi_{t+1}^{\max})\frac{1}{w_{t+1}}$, *all* t.

Proof By the convexity of MP^{ν}, we can consider solutions with $\alpha_t^{\nu} = \mathbf{0}$, for all t and all $\nu \in \mathscr{N}$ without loss of generality.

Take any t and consider $\nu \in N$ such that $p_{t-1}\omega_t^{\nu} > 0$ and $\Lambda_t^{\nu} \in (0, 1)$. By Proposition 1

$$\Lambda_t^{*\nu} = \gamma_t^{*\nu} = \frac{p_t b - \pi_t^{\max} p_{t-1}\underline{\beta}_t^{*\nu} + p_t s_t^{*\nu} + (p_t - p_{t-1})\,\omega_t^{*\nu}}{w_t}.$$

Then,

$$u_t^{\nu} + \rho u_{t+1}^{\nu} = -[\Lambda_t^{*\nu} + \rho\Lambda_{t+1}^{*\nu}] = \frac{-p_t b + \pi_t^{\max} p_{t-1}\underline{\beta}_t^{*\nu} - p_t s_t^{*\nu} - (p_t - p_{t-1})\omega_t^{*\nu}}{w_t}$$
$$+\rho\frac{-p_{t+1}b + \pi_{t+1}^{\max} p_t \underline{\beta}_{t+1}^{*\nu} - p_{t+1}s_{t+1}^{*\nu} - (p_{t+1} - p_t)\,\omega_{t+1}^{*\nu}}{w_{t+1}}.$$

Consider a one-period perturbation $s_t'^{\nu} = s_t^{*\nu} + \Delta_t^{\nu}, s_{t+1}'^{\nu} = s_{t+1}^{*\nu} + \Delta_{t+1}^{\nu}$, such that $\Delta_t^{\nu} = -\Delta_{t+1}^{\nu}$. In the perturbed path,

$$u_t'^{\nu} + \rho u_{t+1}'^{\nu} = -\left[\Lambda_t^{*\nu} + \rho\Lambda_{t+1}^{*\nu}\right] - \frac{p_t\Delta_t^{\nu}}{w_t} + \rho\frac{\pi_{t+1}^{\max}p_t\Delta_t^{\nu} - p_t\Delta_{t+1}^{\nu}}{w_{t+1}}$$

$$= u_t^{\nu} + \rho u_{t+1}^{\nu} - \left[\frac{1}{w_t} - \rho\frac{\left(1+\pi_{t+1}^{\max}\right)}{w_{t+1}}\right]p_t\Delta_t^{\nu}.$$

Note that at a non-trivial RS it must be $p_{it} > 0$, some i for all t. Therefore, if $\frac{1}{w_t} < \rho\frac{\left(1+\pi_{t+1}^{\max}\right)}{w_{t+1}}$, then there is a sufficiently small $\Delta_t^{\nu} \geq \mathbf{0}$ that is feasible and generates $u_t'^{\nu} + \rho u_{t+1}'^{\nu} > u_t^{\nu} + \rho u_{t+1}^{\nu}$, contradicting optimality. A similar argument holds for $\frac{1}{w_t} < \rho\frac{\left(1+\pi_{t+1}^{\max}\right)}{w_{t+1}}$. Hence, $\frac{1}{w_t} = \rho\frac{\left(1+\pi_{t+1}^{\max}\right)}{w_{t+1}}$. ■

In what follows, RSs with *stationary* capital are of focal interest. As argued below, in equilibria with savings, some of the basic insights of Roemer's analysis do not hold. Moreover given the absence of population growth and technical progress, a RS with stationary capital implies that aggregate capital at the beginning of each generation's life is already optimal in terms of the "golden rule". For a sufficiently large T, if the initial aggregate capital stock was not at the optimal level, then agents would accumulate up to the optimal level as soon as possible and spend most of their lives with this optimal level of capital stock in order to minimise labour, as in the so-called *Turnpike Theorem* (see Morishima [16]). Therefore, we focus on RSs with stationary capital, and persistent capital scarcity, by assuming that aggregate capital is already at the optimal level in the initial period. Formally:

Definition 8 An *interior reproducible solution* (IRS) for $E_b(\Omega_{kT})$ is a RS $\langle(\mathbf{p}, \mathbf{w}), (\xi^{\nu})_{\nu\in N}\rangle$ such that $s_t^{\nu} = \mathbf{0}$ for all $\nu \in \mathscr{N}$ at every t.

4.2 Two Views of UE Exploitation in Dynamic Contexts

In what follows, we abstract from unnecessary technicalities in our dynamic analysis and assume that there is only one consumption good, and the technology is of a simple Leontief type. Formally, the subsistence bundle is $b > 0$, and the production set is $P = P_{(A,L)}$ for a Leontief technology (A, L), with $A \in (0, 1)$, and $L > 0$. We adopt the standard notation: in every t, x_t^{ν} represents ν's activity level as a self-employed producer, and y_t^{ν} is the activity level that ν hires others to operate. Thus, for any $\left(x_t^{\nu}, y_t^{\nu}, \gamma_t^{\nu}\right)$, there is $\left(\alpha_t^{\nu}, \beta_t^{\nu}, \gamma_t^{\nu}\right) \in P_{(A,L)} \times P_{(A,L)} \times [0, 1]$ such that $\alpha_t^{\nu} = \left(-Lx_t^{\nu}, -Ax_t^{\nu}, x_t^{\nu}\right)$, and $\beta_t^{\nu} = \left(-Ly_t^{\nu}, -Ay_t^{\nu}, y_t^{\nu}\right)$, and a similar notation holds at the aggregate level. Hence, we use the notation $\left(x_t^{\nu}, y_t^{\nu}, \gamma_t^{\nu}\right)$ and $x_t + y_t$ to denote, respectively, individual plans and the aggregate production activity.

Let $\upsilon = L(1-A)^{-1}$. In one-good economies with a linear production technology, the labour content of an amount c of the good is simply υc in all of the main approaches. Let $\Delta^{\nu} = \sum_{t=kT}^{(k+1)T-1} \left(\Lambda_t^{\nu} - \upsilon b\right)$. Unlike in the static model, there are two different criteria to define the exploitation status of an agent, focusing on the amount of labour performed either in each period, or during her whole life.

Definition 9 Agent ν is exploited within period t, or WP_t exploited, iff $\Lambda_t^{\nu} > \upsilon b$; and a WP_t exploiter iff $\Lambda_t^{\nu} < \upsilon b$. Similarly, ν is exploited during her whole life, or *WL* exploited, iff $\Delta^{\nu} > 0$; and a *WL* exploiter iff $\Delta^{\nu} < 0$.

The *WP* and *WL* definitions incorporate different normative concerns. The *WL* definition captures the intuition that, from an *individual's* viewpoint, to be exploited in every period is certainly worse than being exploited only in some periods. This criterion may lead us to conclude that, from a perspective of individual well-being, an exploitative economy with social mobility is better than an exploitative economy without it.

Marx's idea, however, is more radical: the existence of exploitation is morally relevant per se, and exploitation may be considered as a property of the economy as a whole, not just of individuals. The *WP* definition captures this intuition: the existence of WP_t *exploited agents* and WP_t *exploiters* implies the existence of exploitative social relations as a property of the whole economy. Further, an analysis based on the *WL* definition can only *partly* capture the exploitative structure of the economy, because it may lead to the conclusion that there would be no exploitation in "*changing places capitalism*," that is in a capitalist economy with significant social mobility, where *WP* exploitation exists in every period but the agents' status changes over time so as to equalise lifetime labour hours, which is rather counterintuitive from a Marxian perspective. Hence, although both criteria convey normatively relevant information, we focus mainly on the *WP* definition, which is also more suitable to analyse the dynamics of exploitation.

If agents save, it may be difficult to extend Roemer's asset-based theory of exploitation to the dynamic context: given the optimality of $\sum_{t=kT}^{(k+1)T-1} s_t^{\nu} = 0$ for all ν, and the linearity of MP^{ν}, an agent can be a WP_t exploiter while being WP_{t+j} exploited, for some $j \neq 0$, depending on the path of savings (and only indirectly on ω_{kT}^{ν}). Such changes in *WP* status, however, do not necessarily convey morally relevant information: the fact that at a non-interior RS a relatively wealthy agent might optimally work more than υb in t, in order to accumulate assets and minimise labour in $t + j$, does not raise serious moral concerns. Actually, it is not difficult to show that if $s_t \neq 0$ then there is no conceptual equivalence between *WP* exploitative and inegalitarian solutions: only at an IRS, if an agent works less than υb, there must be another agent working more than υb.[8]

[8]This argument does not apply to the *WL* definition: the existence of a general monotonic relationship between initial wealth and *WL* exploitation at a RS where agents save is an interesting issue for further research.

Given our focus on IRSs, in what follows we consider $kT = 0$ without loss of generality. By Lemma 2, at an IRS, we can set $p_t = p$, all t, and consider equilibrium price vectors of the form $(p, \mathbf{w})$. Finally, in the one-good economy, at any t the profit rate is denoted more simply as π_t.

For any $(p, \mathbf{w})$, let $W_t^* \equiv (pb - w_t \upsilon b)/\pi_t$. Proposition 3 proves that at an IRS, the *WL* and *WP* definitions are equivalent and it extends Roemer's asset-based theory of exploitation to the dynamic context.

Proposition 3 *Let $(p, \mathbf{w})$ be an IRS for $E(\Omega_0)$ with $\pi_0 > 0$. Then:*

1. *$\Delta^\nu > 0$ and $\Lambda_t^\nu > \upsilon b$, all t, if and only if $W_0^\nu < W_0^*$;*
2. *$\Delta^\nu = 0$ and $\Lambda_t^\nu = \upsilon b$, all t, if and only if $W_0^\nu = W_0^*$;*
3. *$\Delta^\nu < 0$ and $\Lambda_t^\nu < \upsilon b$, all t, if and only if $W_0^\nu > W_0^*$.*

Proof 1. At all t, $W_t^\nu = W_t^*$ is equivalent to $\pi_t W_t^\nu = [p(1-A) - w_t L](1-A)^{-1} b$, or $p\omega_0^\nu = pA(1-A)^{-1}b$. Thus, if $W_t^\nu = W_t^*$, then $W_{t+1}^\nu = W_{t+1}^*$, all t. Similarly, $W_t^\nu > W_t^\mu$ implies $W_{t+1}^\nu > W_{t+1}^\mu$ for any ν, μ, and all t.

2. By Proposition 1 and the strict monotonicity of $p[b - \pi_t \omega_0^\nu]$ in W_t^ν at all t, $\Lambda_t^\nu > \upsilon b \Leftrightarrow W_t^\nu < W_t^*$, $\Lambda_t^\nu = \upsilon b \Leftrightarrow W_t^\nu = W_t^*$, and $\Lambda_t^\nu < \upsilon b \Leftrightarrow W_t^\nu > W_t^*$. Hence, by step 1, $\Lambda_0^\nu > \upsilon b$ implies $\Lambda_t^\nu > \upsilon b$ all $t > 0$, and thus $\Delta^\nu > 0$. Conversely, if $\Delta^\nu > 0$, it must be $\Lambda_t^\nu > \upsilon b$ for at least some $t \geqq 0$. However, as just shown, *WP* exploitation status cannot change over time, and thus $\Lambda_t^\nu > \upsilon b$, all t. The other two cases are proved similarly. ■

4.3 Exploitation, DOPA and Welfare Inequalities

Given Proposition 3, it is natural to focus on IRSs in order to analyse the links between exploitation and wealth. The next results derive the conditions under which Roemer's [23, 25] theory can be extended to the intertemporal context, and at the same time highlight the conceptual links and differences between his definition of exploitation and neoclassical welfare inequalities.

Theorem 2 *Let $\pi' = (1-\rho)/\rho$ and let (p, w') be the associated price vector. If $w_t = w'$ all t, and $pb \leq w'$, then for all ν, $s_t^\nu = 0$, all t, is optimal. Moreover, if T is finite, then $V(\omega_0^\nu) = \max\{0, (1-\rho^T)[\frac{p'b}{(1-\rho)} - \frac{p'\omega_0^\nu}{\rho}]\}$, while if $T \to \infty$, then $V(\omega_0^\nu) = \max\{0, \frac{p'b}{(1-\rho)} - \frac{p'\omega_0^\nu}{\rho}\}$, where $p' \equiv \frac{p}{w'}$.*

Proof 1. Suppose $p\omega_0^\nu \geqq pb\rho/(1-\rho)$. The vector ξ^ν such that $s_t^\nu = 0$ all t, and $y_t = y'$ all t, with $\pi' A y' = b$, is optimal and $\Lambda_t^\nu = 0$ all t.

Suppose $p\omega_0^\nu < pb\rho/(1-\rho)$, so that $\Lambda^\nu > 0$ for all $\xi^\nu \in \mathscr{O}^\nu(p, w')$. Write MP^ν using dynamic optimisation theory. Let $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ be the feasibility correspondence:

$$\Psi(\omega_t^\nu) = \{\omega_{t+1}^\nu \in \mathbb{R}_+ | \frac{p}{w_t}\omega_{t+1}^\nu \leqq 1 - \frac{p}{w_t}b + \frac{p}{w_t}\omega_t^\nu + \pi_t \frac{p}{w_t}\omega_t^\nu\}.$$

Given ω_0^ν, let

$$\Pi(\omega_0^\nu) = \left\{\omega^\nu \,|\omega_{t+1}^\nu \in \Psi(\omega_t^\nu) \text{ all } t, \,\&\, \omega_T^\nu \geqq \omega_0^\nu\right\}$$

be the set of feasible sequences ω^ν. Let $\Phi = \{(\omega_t^\nu, \omega_{t+1}^\nu) \in \mathbb{R}_+ \times \mathbb{R}_+ | \omega_{t+1}^\nu \in \Psi(\omega_t^\nu)\}$ be the graph of Ψ. The one-period return function $F : \Phi \to \mathbb{R}_+$ at t is $F(\omega_t^\nu, \omega_{t+1}^\nu) = \frac{p}{w_t} b + \frac{p}{w_t}(\omega_{t+1}^\nu - \omega_t^\nu) - \pi_t \frac{p}{w_t}\omega_t^\nu$. MP^ν becomes

$$V(\omega_0^\nu) = \min_{\omega^\nu \in \Pi(\omega_0^\nu)} \sum_{t=0}^{T-1} \rho^t \left[\frac{p}{w_t} b + \frac{p}{w_t}(\omega_{t+1}^\nu - \omega_t^\nu) - \pi_t \frac{p}{w_t}\omega_t^\nu\right].$$

If $\frac{pb - \pi_t p\omega_t^\nu}{w_t} \leqq 1$ for all t, then $\Psi(\omega_t^\nu) \neq \varnothing$ for all $\omega_t^\nu \in \mathbb{R}_+$. Then, since F is continuous and bounded, MP^ν is well defined for all T.

2. If $w_t = w'$ for all t, then $\frac{pb - \pi_t p\omega_t^\nu}{w_t} \leqq 1$ for all t, ν, and MP^ν becomes:

$$V(\omega_0^\nu) = \min_{\omega^\nu \in \Pi(\omega_0^\nu)} \sum_{t=0}^{T-1} \rho^t p'b + \rho^{T-1} p'\omega_T^\nu - (1+\pi')p'\omega_0^\nu, \text{ where } p' \equiv \frac{p}{w'}.$$

Therefore, for all T, any feasible ω^ν such that $\omega_T^\nu = \omega_0^\nu$ (or $\lim_{T\to\infty} \omega_T^\nu = \omega_0^\nu$, if $T \to \infty$) is optimal and $V(\omega_0^\nu)$ immediately follows.

3. The last part of the statement is straightforward. ■

At an IRS, if $\pi_t = \pi' = (1-\rho)/\rho$, all t, then $(p, w_t) = (p, w')$, all t, and so the IRS is a *stationary RS* (SRS).

Given Theorem 2, the next result characterises welfare inequalities and exploitation at a SRS, if agents discount future labour.

Theorem 3 *Let $1 > \rho$. Let (p, w') be a SRS for $E(\Omega_0)$ with $\pi' = (1-\rho)/\rho$, all t. Then:*

(i) for all $\nu, \mu \in \mathcal{N}$, if $p'\omega_0^\mu < \frac{p'b\rho}{(1-\rho)}$, then $V(\omega_0^\nu) < V(\omega_0^\mu)$ if and only if $p'\omega_0^\nu > p'\omega_0^\mu$, where $p' \equiv \frac{p}{w'}$;

(ii) There is a constant k^ν such that $\Lambda_t^\nu - \upsilon b = k^\nu$ all t, ν.

Proof Part (i). Directly from Theorem 2, since $V(\omega_0^\nu) = 0$ if and only if $p'\omega_0^\nu \geq p'b/\pi'$; while if $V(\omega_0^\nu) > 0$, then $V(\omega_0^\nu) - V(\omega_0^\mu) = (1-\rho^T)[p'\omega_0^\mu - p'\omega_0^\nu]/\rho$ when T is finite, and $V(\omega_0^\nu) - V(\omega_0^\mu) = [p'\omega_0^\mu - p'\omega_0^\nu]/\rho$ if $T \to \infty$.

Part (ii). Straightforward, given Proposition 1. ■

Theorems 2 and 3 complete the intertemporal generalisation of Roemer's theory: the dynamic economy with discounting displays the same pattern of *WP* and *WL* exploitation as the T-fold repetition of the static economy, and both *WP* and *WL* exploitation are persistent. Moreover, unlike in the static model, the introduction of time preference in the dynamic model clarifies that Roemer's interpretation of Marxian exploitation at the *WL* level as an objectivist measure of inequalities—"the

exploitation-welfare criterion" [23, p. 75]—and subjectivist neoclassical welfare inequalities are different in general: the former notion focuses on asset inequalities, which are independent of time preference, while the latter focuses on welfare inequalities, which depend on ρ. According to Theorems 2 and 3, the two views coincide at a SRS, but they are conceptually distinct.

4.4 The Property Relations Definition of Exploitation

The previous sections generalise Roemer's *UE approach* to exploitation to the dynamic context and show that his definition of Marxian exploitation is distinct from, but conceptually related to neoclassical welfare inequalities. This section generalises Roemer's [23] *game-theoretic* approach, which focuses on property relations, a more general concept than asset inequalities.

Let $(V^1, \ldots, V^N)$ be the agents' payoffs at the existing allocation: in this context, it is natural to consider $(V^1, \ldots, V^N)$ as *WL* values. For instance, at an RS for $E(\Omega_0)$, $V^1 = -V(\omega_0^1), \ldots, V^N = -V(\omega_0^N)$. Let $P(\mathcal{N})$ be the power set of $\mathcal{N}$ and let $K : P(\mathcal{N}) \to \mathbb{R}_+$ be a *characteristic function* which assigns to every coalition $\mathcal{J} \subseteq \mathcal{N}$ with J agents an aggregate payoff $K(\mathcal{J})$ if it withdraws from the economy.

Definition 10 ([23], pp. 194–195) Coalition $\mathcal{J} \subseteq \mathcal{N}$ is exploited at allocation $(V^1, \ldots, V^N)$ with respect to alternative K if and only if the complement to $\mathcal{J}$, $\mathcal{N} - \mathcal{J} = \mathcal{J}'$, is in a relation of dominance to $\mathcal{J}$ and

(i) $\sum_{\nu \in \mathcal{J}} V^\nu < K(\mathcal{J})$,
(ii) $\sum_{\nu \in \mathcal{J}'} V^\nu > K(\mathcal{J}')$.

Definition 10 captures various kinds of exploitation, including Marxian exploitation, by specifying different hypothetically feasible alternatives. The concept of exploitation is related to the *core* of an economy: the set of non-exploitative allocations coincides with the core of the game described by K [23, Theorem 7.1, p. 198]. The precise definition of exploitation depends on the function K. A coalition is *feudally exploited* at an allocation if it can improve by withdrawing from society with its own endowments and arranging production on its own. In $E(\Omega_0)$, the set of feudally non-exploitative allocations coincides with the *private ownership core* (POC). Formally, a coalition $\mathcal{J}$ is viable if it has enough assets to reproduce itself if it secedes from the parent economy [23, pp. 45–49].

Definition 11 A coalition $\mathcal{J} \subseteq \mathcal{N}$ is viable if $\sum_{\nu \in \mathcal{J}} \omega_0^\nu \geqq JA(1 - A)^{-1}b$.

A reproducible allocation is a profile of (not necessarily optimal) actions of all agents in $E(\Omega_0)$, that satisfy the feasibility and reproducibility constraints.

Definition 12 A *reproducible allocation* (RA) for $E(\Omega_0)$ is a profile of actions $\xi^\nu = (x^\nu, y^\nu, \gamma^\nu, s^\nu)$ for all ν, such that

1. $Lx_t^\nu + \gamma_t^\nu \leqq 1$, all ν, t;
2. $A(x_t + y_t) \leqq \omega_t$ all t;
3. $(x_t + y_t) \geqq A(x_t + y_t) + Nb + s_t$, all t;
4. $\omega_{t+1} = \omega_t + s_t$, all t;
5. $\omega_T \geqq \omega_0^\nu$.

A viable coalition $\mathscr{J}$ can block a RA $(\xi^\nu)_{\nu\in\mathscr{N}}$ if there is another RA for the smaller economy that yields higher welfare to its members.

Definition 13 A viable coalition $\mathscr{J}$ can *block* a RA $(\xi^\nu)_{\nu\in\mathscr{N}}$ if there is a profile $\left(\xi'^1, \ldots, \xi'^J\right)$ such that

1. $\sum_{t=0}^{T-1} \rho^t \Lambda_t'^\nu < \sum_{t=0}^{T-1} \rho^t \Lambda_t^\nu$, all $\nu \in \mathscr{J}$;
2. $A\sum_{\nu\in\mathscr{J}} x_t'^\nu \leqq \sum_{\nu\in\mathscr{J}} \omega_t^\nu$, all t;
3. $(1-A)\sum_{\nu\in\mathscr{J}} x_t'^\nu = Jb + \sum_{\nu\in\mathscr{J}} s_t'^\nu$, all t;
4. $\sum_{\nu\in\mathscr{J}} \omega_{t+1}^\nu = \sum_{\nu\in\mathscr{J}} \omega_t^\nu + \sum_{\nu\in\mathscr{J}} s_t'^\nu$, all t;
5. $\sum_{\nu\in\mathscr{J}} \omega_T^\nu \geqq \sum_{\nu\in\mathscr{J}} \omega_0^\nu$.

The POC of $E(\Omega_0)$ is the set of RAs which no coalition can block. Theorem 4 proves the absence of feudal exploitation in $E(\Omega_0)$.

Theorem 4 *Let $\rho \leqq 1$. Any IRS of $E(\Omega_0)$ lies in its private ownership core and thus displays no feudal exploitation.*

Proof 1. If $\pi_t = 0$, all t, the result is trivial. Hence, assume $\pi_0 > 0$.

2. Suppose that there is $\mathscr{J} \subseteq \mathscr{N}$ that can block the IRS. By Definition 13(1), no pure capitalist can belong to $\mathscr{J}$; thus, by Lemma 1 and Proposition 1, at an IRS $(p, \mathbf{w})$, $\pi_t \frac{p}{w_t}\omega_0^\nu = \frac{p}{w_t}b - \Lambda_t^\nu$ all t and all $\nu \in \mathscr{J}$. Summing over $\nu \in \mathscr{J}$ and t, $\sum_{t=0}^{T-1} \rho^t \pi_t \frac{p}{w_t} \sum_{\nu\in\mathscr{J}} \omega_0^\nu = \sum_{t=0}^{T-1} \rho^t J \frac{p}{w_t} b - \sum_{t=0}^{T-1} \rho^t \sum_{\nu\in\mathscr{J}} \Lambda_t^\nu$. By Proposition 2, $\sum_{t=0}^{T-1} \rho^t \pi_t \frac{p}{w_t} \sum_{\nu\in\mathscr{J}} \omega_0^\nu = [(1+\pi_0)\frac{p}{w_0} - \rho^{T-1}\frac{p}{w_{T-1}}] \sum_{\nu\in\mathscr{J}} \omega_0^\nu$.

3. If $\mathscr{J}$ can block the IRS, multiplying Definition 13(3) by $\rho^t \upsilon$ and summing over t, $\sum_{t=0}^{T-1} \rho^t \sum_{\nu\in\mathscr{J}} \Lambda_t'^\nu = \sum_{t=0}^{T-1} \rho^t J\upsilon b + \sum_{t=0}^{T-1} \rho^t \upsilon \sum_{\nu\in\mathscr{J}} s_t^\nu$. By Definition 13(1) and step 2: $\sum_{t=0}^{T-1} \rho^t J(\frac{p}{w_t} - \upsilon)b - \sum_{t=0}^{T-1} \rho^t \upsilon \sum_{\nu\in\mathscr{J}} s_t^\nu > [(1+\pi_0)\frac{p}{w_0} - \rho^{T-1}\frac{p}{w_{T-1}}] \sum_{\nu\in\mathscr{J}} \omega_0^\nu$.

4. If $\mathscr{J}$ can block the IRS, by Definition 13(2)–(3), $A(1-A)^{-1}(Jb + \sum_{\nu\in\mathscr{J}} s_t^\nu) \leqq \sum_{\nu\in\mathscr{J}} \omega_t^\nu$ all t; multiplying both sides by $\rho^t \pi_t \frac{p}{w_t}$, $\rho^t(\frac{p}{w_t} - \upsilon)Jb - \rho^t \upsilon \sum_{\nu\in\mathscr{J}} s_t^\nu \leqq \rho^t \pi_t \frac{p}{w_t} \sum_{\nu\in\mathscr{J}} \omega_t^\nu - \rho^t \frac{p}{w_t} \sum_{\nu\in\mathscr{J}} s_t^\nu$ all t. Summing over t, by Definition 13(4), the latter expression becomes $\sum_{t=0}^{T-1} \rho^t(\frac{p}{w_t} - \upsilon)Jb - \sum_{t=0}^{T-1} \rho^t \upsilon \sum_{\nu\in\mathscr{J}} s_t^\nu \leqq \sum_{t=0}^{T-1} \rho^t [(1+\pi_t)\frac{p}{w_t} \sum_{\nu\in\mathscr{J}} \omega_t^\nu - \frac{p}{w_t} \sum_{\nu\in\mathscr{J}} \omega_{t+1}^\nu]$. Then, using $\rho(1+\pi_{t+1})\frac{p}{w_t+1} = \frac{p}{w_t}$ all t, $\sum_{t=0}^{T-1} \rho^t(\frac{p}{w_t} - \upsilon)Jb - \sum_{t=0}^{T-1} \rho^t \upsilon \sum_{\nu\in\mathscr{J}} s_t^\nu \leqq (1+\pi_0)\frac{p}{w_0} \sum_{\nu\in\mathscr{J}} \omega_0^\nu - \rho^{T-1}\frac{p}{w_{T-1}} \sum_{\nu\in\mathscr{J}} \omega_T^\nu$.

5. The latter inequality and the inequality in step 3 can both hold only if $\sum_{\nu\in\mathscr{J}} \omega_T^\nu < \sum_{\nu\in\mathscr{J}} \omega_0^\nu$, which contradicts Definition 13(5). ■

In Roemer's interpretation of historical materialism as predicting the progressive disappearance of various forms of exploitation, Theorem 4 proves that capitalist relations of production eliminate feudal exploitation. It also clarifies the neoclassical claim concerning the absence of exploitation in a competitive economy: there is no *feudal* exploitation [23, pp. 205–208].

A different specification of K is necessary to define *capitalist* exploitation. Let $\omega_0^{\mathscr{J}} \equiv \frac{J}{N}\omega_0$ be coalition $\mathscr{J}$'s per-capita share of aggregate initial assets. Given the linear technology, all coalitions are viable if they withdraw with $\omega_0^{\mathscr{J}}$. Then, a coalition can communally block a RA if it can increase the welfare of its members by withdrawing with $\omega_0^{\mathscr{J}}$.

Definition 14 A coalition $\mathscr{J}$ can *communally block* a RA $(\xi^\nu)_{\nu\in\mathscr{N}}$ if there is a profile of vectors $\left(\xi'^1, \ldots, \xi'^J\right)$ such that

1. $\sum_{t=0}^{T-1} \rho^t \Lambda_t'^\nu < \sum_{t=0}^{T-1} \rho^t \Lambda_t^\nu$, all $\nu \in \mathscr{J}$;
2. $A\sum_{\nu\in\mathscr{J}} x_t'^\nu \leqq \omega_t^{\mathscr{J}}$, all t;
3. $(1-A)\sum_{\nu\in\mathscr{J}} x_t'^\nu = Jb + \sum_{\nu\in\mathscr{J}} s_t'^\nu$, all t;
4. $\omega_{t+1}^{\mathscr{J}} = \omega_t^{\mathscr{J}} + \sum_{\nu\in\mathscr{J}} s_t'^\nu$, all t;
5. $\omega_T^{\mathscr{J}} \geqq \omega_0^{\mathscr{J}}$.

The *communal core* of $E(\Omega_0)$ is the set of RAs which no coalition can communally block; a coalition is *capitalistically exploited* if it can communally block the RA; and a RA is *capitalist non-exploitative* if it lies in the communal core of the economy. Theorem 5 proves that Marxian exploitation and capitalist exploitation coincide in $E(\Omega_0)$ at an IRS.

Theorem 5 *Let $\rho \leqq 1$. At an IRS, a coalition is WL Marxian exploited if and only if it is capitalistically exploited.*

Proof If a coalition $\mathscr{J}$ is Marxian exploited, $\sum_{t=0}^{T-1}(\sum_{\nu\in\mathscr{J}} \Lambda_t^\nu - J\upsilon b) > 0$. But then by Proposition 3, at an IRS $\sum_{t=0}^{T-1} \rho^t(\sum_{\nu\in\mathscr{J}} \Lambda_t^\nu - J\upsilon b) > 0$, and $\mathscr{J}$ can communally block the allocation. The converse is proved similarly. ■

Theorem 5 suggests that Marxian exploitation can be seen as a special case of Roemer's Definition 10 in a linear economy with labour-minimising agents. The property-relation definition (which can be applied to a general set of economies; Roemer [23, chapter 7]) would then be a generalisation of Marx's theory that captures its essential normative content.

4.5 Power and the Persistence of Exploitation

The previous results provide a complete dynamic generalisation of Roemer's distributive approach, and thus may be seen as confirming Roemer's key theoretical insight that exploitation can be reduced to a concern for asset inequalities. This section raises some doubts on this conclusion. For DOPA is not necessary *and sufficient* to generate persistent exploitation.

Theorem 6 shows that if agents do not discount the future, profits and the UE of labour tend to decrease over time.

Theorem 6 *Let $\rho = 1$. Let $(p, \mathbf{w})$ be an IRS for $E(\Omega_0)$ with $\pi_0 > 0$. Then (i) $\pi_t > \pi_{t+1}$, all t. Moreover, (ii) for all $\nu \in \mathcal{N}$ such that $\Lambda^\nu > 0$ for all $\xi^\nu \in \mathcal{O}(p, \mathbf{w})$, at all t, if $W_t^\nu < W_t^*$ then $\Lambda_t^\nu > \Lambda_{t+1}^\nu$, if $W_t = W_t^*$ then $\Lambda_t^\nu = \Lambda_{t+1}^\nu$, and if $W_t > W_t^*$ then $\Lambda_t^\nu < \Lambda_{t+1}^\nu$.*

Proof *Part (i).* The result follows noting that $\frac{p}{w_t}$ is a continuous, increasing function of π_t, all t, while by Proposition 2, $\frac{p}{w_t} > \frac{p}{w_{t+1}}$ all t.

Part (ii). By Proposition 3 if $W_t^\nu = W_t^*$ then $\Lambda_t^\nu = \Lambda_{t+1}^\nu = \upsilon b$, all t. By Proposition 1, $\Lambda_{t+1}^\nu - \Lambda_t^\nu = (\frac{p}{w_{t+1}} - \frac{p}{w_t})b + \left(\pi_t \frac{p}{w_t} - \pi_{t+1} \frac{p}{w_{t+1}}\right) \omega_0^\nu$ or, equivalently, $\Lambda_{t+1}^\nu - \Lambda_t^\nu = (\frac{p}{w_{t+1}} - \frac{p}{w_t})b + \left(\pi_t - \frac{\pi_{t+1}}{1+\pi_{t+1}}\right) \frac{p}{w_t} \omega_0^\nu$. Therefore the result follows from part (i) and the monotonicity of the right hand side of the latter expression in W_t^ν. ■

Theorem 6 is rather counterintuitive. In the equilibrium that preserves the exploitation structure of the competitive economy, profits and *WP* exploitation decrease over time: *WP* exploiters work more while *WP* exploited agents work less, even if neither accumulates. The simple *possibility* of saving implies a decrease in the dispersion of agents' labour times around υb, due to the decrease in profits.

Theorem 7 strengthens these conclusions by looking at the long-run behaviour of the economy.

Theorem 7 *Let $\rho = 1$ and $T \to \infty$. Let $(p, \mathbf{w})$ be an IRS for $E(\Omega_0)$ with $\pi_0 > 0$. Then $\Lambda_t^\nu \to \upsilon b$ and $\frac{p}{w_t} \omega_t^\nu \to \upsilon \omega_0^\nu$, all ν, as $t \to \infty$.*

Proof By $\pi_t = [\frac{p}{w_t}(1 - A) - L]/\frac{p}{w_t} A$ and Proposition 2, if $\rho = 1$ then $\frac{p}{w_{t+1}} = \frac{p}{w_t} A + L$ at IRS, thus $\frac{p}{w_t} = [\frac{p}{w_0} - L(1 - A)^{-1}]A^t + L(1 - A)^{-1}$, which implies by $A \in (0, 1)$ that $\frac{p}{w_t} \to \upsilon$ and $\pi_t \to 0$ as $t \to \infty$. ■

Theorem 7 completes the analysis. The previous sections extend Roemer's theory to the intertemporal context, but the key results crucially depend on the assumption that $\rho < 1$. If $\rho = 1$, at any SRS, Proposition 2 implies zero profits leading to a non-exploitative allocation. Theorems 6 and 7 generalise this conclusion. In the equilibrium which preserves DOPA and the exploitation structure of the economy, profits and *WP* exploitation decrease over time and tend to disappear in the long run,

even if capital scarcity persists (unlike in accumulation models, such as Devine and Dymsky [3]).[9]

This conclusion is robust. The above results extend and generalise analogous conclusions by Veneziani [33, 34] in dynamic subsistence economies. Moreover, Veneziani and Yoshihara [35] have shown that, whenever $\rho = 1$, *WP* exploitation also tends to disappear in more general dynamic general equilibrium models with convex technologies and standard utility functions defined over consumption and leisure. The question, then, concerns the implications of these results for exploitation theory.

Proposition 3, and Theorems 5 and 7 can be interpreted as identifying asset inequalities *and* a strictly positive rate of time preference as the necessary and sufficient conditions for the persistence of exploitation in a neoclassical dynamic framework. This questions Roemer's claim that *all* the relevant normative intuitions about exploitation are captured by DOPA.

In fact, as noted above, although Roemer's argument about the focal, indeed exclusive relevance of asset inequalities is *normative* in nature, it crucially rests on a *positive* claim that "differential distribution of property and competitive markets are sufficient institutions to generate an exploitation phenomenon, under the simplest possible assumptions" [23, p. 43]. This suggests that the Marxian concept of exploitation can be reduced to an asset-based approach, and provides the foundations for Definition 10. As Skillman [30, p. 311] aptly noted, "the legitimacy of Roemer's reformulation depends in large part on the validity of his claims concerning the role of DOPA in capitalist exploitation".

By significantly qualifying Roemer's *positive* claim, Proposition 3 and Theorems 5 and 7 raise some doubts on Definition 10 both per se and as a generalisation of Marx's theory. For they prove that at a RS where no agent accumulates and capital scarcity persists, DOPA is necessary to generate UE exploitation, but it is not sufficient for it to persist. Thus, the persistence of DOPA per se is not a sufficient statistic of the unfairness of labour/capital relations (and more generally, of a society). Something else is indispensable to guarantee the persistence of exploitation, which would be normatively at least as important as DOPA itself. Definition 10 may be seen as incorporating a different moral concern, rather than as a generalisation of Marx's definition. More generally, the question arises whether DOPA should be a basic moral concern, both in itself and in a theory of exploitation, or rather a different role of DOPA should be stressed as a causally primary, but normatively secondary wrong.

To be sure, it may be argued that the above analysis shows that exploitation *is* persistent, provided agents discount the future. This objection is not entirely

[9] Okishio [20] also shows that in a dynamic capitalist economy with neither population growth nor technical change, competition among capitalists may drive profits and exploitation to zero. According to Okishio [20], this profit squeeze derives from the increase in the real wage rate due to capital accumulation. Okishio's [20] results, however, are based on simulation methods and only hold for a specific choice of parameters.

compelling. As Veneziani [34] has argued, Roemer's key argument is a *logical* claim about the sufficiency of DOPA—and of the key institutional features of capitalist economies (i.e. competitive markets)—to generate exploitation. The specific value of a parameter of the agents' utility function should not be relevant at this level of abstraction. Moreover, whether agents do, or do not, display impatient time preferences is a purely *empirical* issue, and one which has a priori little to do with DOPA or with the fundamental features of a capitalist economy.

Finally, a theoretical argument that crucially relies on time preference seems at odds with the core intuitions of UE exploitation theory which emphasise the structural features of capitalist economies. As Roemer [25, 60ff] himself notes, the normative relevance of a theory of exploitation critically relying on such exogenous factors would be rather unclear. This is particularly relevant in our model because, by Theorems 2 and 3, both the persistence *and* the magnitude of exploitation and welfare inequalities depend on time preference. Given the positive relation between the profit rate, and welfare inequalities and exploitation, the higher ρ, the lower the equilibrium profit rate, and the lower UE exploitation, ceteris paribus.

In summary, the above results provide a robust criticism of Roemer's core claim that DOPA is the fundamental cause of exploitation. This claim crucially depends on very restrictive assumptions, such as the impossibility of savings. If savings are allowed, DOPA is necessary but not sufficient to generate persistent exploitation, and an emphasis on asset inequalities while exploitation disappears seems misplaced. Therefore, the intertemporal model raises serious doubts on the claim that exploitation theory can be reduced to a form of resource egalitarianism.[10]

It is certainly possible, and interesting, to investigate some mechanisms that guarantee the persistent abundance of *labour* in a capitalist economy. Skillman [30] and Okishio [20], for example, suggest that a dynamic model including growth in the labour force and/or labour-saving technical change might provide micro-foundations to persistent exploitation in a Marxian framework. This would be consistent with Marx's own approach, which focused on (long-run) equilibria with unemployment due to labour-saving technical progress, whereas the analysis above focuses on the neoclassical full employment equilibrium, which is not the standard feature of capitalist economy for Marx even if $\rho = 1$. However, even if exploitation could be proved to be persistent under those assumptions, it is unclear whether our main conclusions would change. For DOPA and competitive markets would still be insufficient to yield persistent *WP* exploitation.

[10]It might be objected that *WL* exploitation does not disappear, even if $\rho = 1$, and the relationship between initial wealth and *WL* exploitation status is preserved. Thus, from a mathematical viewpoint, the model may be interpreted as a generalisation of Roemer's theory under the *WL* definition. Yet, this does not affect our main conclusions. First, given the theoretical relevance of the *WP* definition, Marxian exploitation should arguably be micro-founded as a persistent *WP* phenomenon. Second, not only is the tendential disappearance of *WP* exploitation disturbing per se; it also implies that ceteris paribus, *WL* exploitation, too, is lower in the dynamic model with agents living for T periods than in the T-fold iteration of the static model.

5 Conclusions

This paper has presented and extended some recent work in exploitation theory. Three main conclusions can be drawn from our analysis. First, the notion of exploitation as the unequal exchange of labour is logically coherent and can be meaningfully defined in economies that are significantly more general than those usually analysed in mathematical Marxian economics. The model set up in Sect. 2 allows for choice of technique, joint production, heterogeneous intertemporally optimising agents with general preferences over consumption and leisure, and different endowments of physical and human capital. Although the model incorporates some features that are not standard in neoclassical theory, such as the time structure of production and the reproducibility of all capital goods, the dynamic equilibrium notion is conceptually cognate to standard notions used in optimal growth models.

Second, the normative foundations of UE exploitation can be analysed by adopting the axiomatic method, and an extension of the "New Interpretation" [4, 5, 10, 11] is the only definition (among the main approaches in the literature) satisfying a number of weak and desirable properties in the general economies analysed in this paper.

Third, exploitation cannot be reduced to a focus on asset inequalities: even if capital scarcity and DOPA persist, in dynamic subsistence economies UE exploitation tends to disappear. A concern for power, dominance, or coercion is an integral part of the notion of exploitation and can contribute to mitigate the "distributive bias" of normative economics.

It is important to stress that this paper does not provide the final word on exploitation theory. It suggests that UE exploitation can be analysed rigorously with the standard tools of normative economics and social choice theory, and that a logically coherent and normatively interesting notion of exploitation can be formulated in general economies. Yet many important issues remain unanswered which represent promising lines for further research. In the rest of this section, we discuss some of them.

First, the axiomatic analysis in Sect. 3 is based on the "*contribution view*" of exploitation theory: exploitative relations are characterised by systematic differences between the amount of *effective* labour that agents contribute to the economy and the labour received by them. As noted by an anonymous referee, however, one may emphasise the "*welfare view*" of exploitation theory, whereby the normative relevance of UE exploitation derives from the fact that income and labour *time* are fundamental determinants of individual *well-being freedom* (e.g. Rawls [21]). If one adopted the *welfare view*, then perhaps both the definitions of exploitation in Sect. 3.1 and the main domain axiom **LE** should be expressed in terms of labour time.

Second, Sect. 4 raises the issue of the appropriate definition of exploitation and in particular the role of distributional and power-related concerns in exploitation theory. Our results suggest that, contrary to Roemer's claim, Marxian exploitation

cannot be reduced to asset inequalities and the resulting welfare inequalities. This raises two sets of issues.

At a normative level, as Veneziani [34] has argued, the dominance condition in Definition 10, is not just necessary "to rule out some bizarre examples" [23, p. 195]: asymmetric relations of power, or dominance should play a definitional role in a theory of exploitation as a social relation in competitive economies. Exploitation should thus be conceived of as involving *both* the outcome *and* the structure of the interaction between agents, as it diagnoses the process through which "certain inequalities in incomes are generated by inequalities in rights and powers over productive resources: the inequalities occur, in part at least, through the ways in which the exploiters, by virtue of their exclusionary rights and powers over resources, are able to appropriate labour effort of the exploited" [38, p. 1563].

At a positive level, the question arises as to the key determinants of the persistence of exploitation in capitalist economies. Arguably, here too, a focus on power, or dominance, may contribute to a more satisfactory explanation of persistent exploitative relations based on the structural features of capitalist economies. As Devine and Dymski [3] noted, two implicit assumptions are necessary in Roemer's theory in order to generate persistent exploitation: capital scarcity and exogenous labour intensity. The former disappears when capital accumulation is introduced, the latter is violated when labour contracts are incomplete. Without complete contracts, exploitative relations may not arise even in a static setting because of the profit-squeeze caused by the lack of labour-discipline in production.[11] Building on this point, Yoshihara [39] integrates incomplete labour contracts into the standard general equilibrium framework of Marxian exploitation theory, and shows that the degree of exploitation is related to the strength of the power relationship which is in turn affected by the degree of asset inequalities: poor agents are forced to provide a higher level of labour intensity per wage rate than wealthier agents.

Given its concern with power and the emphasis on the role of physical assets in explaining hierarchical relations and the existence of firms, the *property rights theory of the firm* [12] may also provide an interesting theoretical framework to analyse exploitative relations which goes beyond purely distributive views and is consistent with the idea that asset inequalities are causally primary, but normatively secondary.

Acknowledgements A preliminary version of the paper was presented at CEPET 2010 Workshop in honour of Nick Baigent, where Nick Baigent, Constanze Binder, Giulio Codognato, Rajat Deb, and Yongsheng Xu provided insightful and useful comments. We are also grateful to Meghnad Desai, Woojin Lee, John Roemer, Ian Steedman, and participants in the "Capitalism, Socialism and Democracy" conference (Amherst), the MANCEPT Workshops and seminars at the LSE, the University of Siena, and the Institute of Education (London) for comments and suggestions on an earlier version of this paper. Special thanks go to Gil Skillman, the editors and two anonymous referees for long and detailed comments. The usual disclaimer applies.

[11] This point is also raised by Bowles and Gintis [2].

References

1. Baigent N (1981) Decompositions of minimal liberalism. Econ Lett 7:29–32
2. Bowles S, Gintis H (1990) Contested exchange: new microfoundations of the political economy of capitalism. Polit Soc 18:165–222
3. Devine J, Dymski G (1991) Roemer's 'general' theory of exploitation is a special case. Econ Philos 7:235–275
4. Duménil G (1980) De la valeur aux prix de production. Economica, Paris
5. Duménil G (1984) The so-called 'transformation problem' revisited: a brief comment. J Econ Theory 33:340–348
6. Duménil G, Foley DK, Lévy D (2009) A note on the formal treatment of exploitation in a model with heterogeneous labor. Metroeconomica 60:560–567
7. Flaschel P (1983) Actual labor values in a general model of production. Econometrica 51:435–454
8. Fleurbaey M (1996) Théories Économiques de la justice. Economica, Paris
9. Fleurbaey M (2013) The facets of exploitation. J Theor Polit 26:653–676
10. Foley DK (1982) The value of money, the value of labour power, and the Marxian transformation problem. Rev Radic Polit Econ 14:37–47
11. Foley DK (1982) Realization and accumulation in a Marxian model of the circuit of capital. J Econ Theory 28:300–319
12. Hart OD (1995) Firms, contracts, and financial structure. Oxford University Press, Oxford
13. Krause U (1982) Heterogeneous labour and the fundamental Marxian theorem. Rev Econ Stud 48:173–178
14. Maniquet F (2002) A study of proportionality and robustness in economies with a commonly owned technology. Rev Econ Des 7:1–15
15. Matsuo T (2008) Profit, Surplus product, exploitation and less than maximized utility. Metroeconomica 59:249–265
16. Morishima M (1969) Theory of economic growth. Oxford University Press, Oxford
17. Morishima M (1973) Marx's economics. Cambridge University Press, Cambridge
18. Morishima M (1974). Marx in the light of modern economic theory. Econometrica 42:611–632
19. Okishio N (1963) A mathematical note on Marxian Theorems. Weltwirtschaftliches Archiv 91:287–299
20. Okishio N (2000) Competition and production prices. Camb J Econ 25:493–501
21. Rawls J (1971) A theory of justice. Harvard University Press, Cambridge
22. Roemer JE (1981) Analytical foundations of Marxian economic theory. Harvard University Press, Cambridge
23. Roemer JE (1982) A general theory of exploitation and class. Harvard University Press, Cambridge
24. Roemer JE (1982) New directions in the Marxian theory of exploitation and classes. Polit Soc 11:253–287
25. Roemer JE (1988) Free to lose. Harvard University Press, Cambridge
26. Roemer JE (1989) Marxism and contemporary social science. Rev Soc Econ 47:377–391
27. Roemer JE (1994) Egalitarian perspectives: essays in philosophical economics. Cambridge University Press, Cambridge
28. Roemer JE (2010) Kantian equilibrium. Scand J Econ 112:1–24
29. Roemer JE, Silvestre JE (1993) The proportional solution for economies with both private and public ownership. J Econ Theory 59:426–444
30. Skillman GL (1995) *Ne Hic Saltaveris*: the Marxian theory of exploitation after Roemer. Econ Philos 11:309–331
31. Steiner H (2013) Liberalism, neutrality and exploitation. Polit Philos Econ 12:335–344
32. van Donselaar G (2009) The right to exploit. Oxford University Press, Oxford
33. Veneziani R (2007) Exploitation and time. J Econ Theory 132:189–207
34. Veneziani R (2013) Exploitation, inequality, and power. J Theor Polit 25:526–545

35. Veneziani R, Yoshihara N (2012) Globalisation and inequalities, Mimeo, Queen Mary University of London, and IER, Hitotsubashi University
36. Veneziani R, Yoshihara N (2015) Exploitation in economies with heterogeneous preferences, skills and assets: an axiomatic approach. J Theor Polit 27:8–33
37. Vrousalis N (2013) Exploitation, vulnerability, and social domination. Philos Public Aff 41:131–157
38. Wright EO (2000) Class, exploitation, and economic rents: reflections on Sorensen's 'Sounder Basis'. Am J Sociol 105:1559–1571
39. Yoshihara N (1998) Wealth, exploitation, and labor discipline in the contemporary capitalist economy. Metroeconomica 49:23–61
40. Yoshihara N (2010) Class and exploitation in general convex cone economies. J Econ Behav Organ 75:281–296
41. Yoshihara N, Veneziani R (2009) Exploitation as the unequal exchange of labour: an axiomatic approach. Working paper N.655, Queen Mary University of London
42. Yoshihara N, Veneziani R (2011) Strong subjectivism in the theory of exploitation: a critique. Metroeconomica 62:53–68
43. Yoshihara N, Veneziani R (2013) Exploitation of labour and exploitation of commodities: a 'New Interpretation.' Rev Radic Polit Econ 45:517–524
44. Ypi L (2010) On the Confusion between ideal and non-ideal in recent debates on global justice. Polit Stud 58:536–555

The Merits of Merit Wants

Richard Sturn

Abstract Merit wants are a multi-faceted concept cutting through a complex array of problems associated with different levels of analysis. They are considered in this paper as a shorthand notion for concerns that are respectable and important, assuming a broadly individualist conception of welfare. So why are merit wants not a firmly established part of modern normative economics, given that simplifying, but still meaningful notions are suitable as conceptual starting point for a research program? In this paper I try to link the answer to this question with making explicit three levels of problems (limits of reason, higher order preferences, collective choice) which may be useful to locate and scrutinize various interpretations of and approaches to merit wants.

Keywords Bounded rationality • Communal preferences • Higher order preferences • Merit wants/goods • Social choice

1 Introduction

Merit wants are a multi-faceted concept cutting through a complex array of problems. Those problems are associated with different levels of analysis, which have in common that they are somehow related to the limits of standard concepts of consumer sovereignty. They either refer to choice situations where there are good reasons to conceptualize human agents not as consumers (but as individuals expressing their values, or as citizens), or they relate to situations where they are conceptualized as consumers, but apparently lack sovereignty in the appropriate sense. Merit wants can be seen as a shorthand notion for various respectable and important concerns; those concerns are particularly important in the context of theoretical frameworks taking human agency seriously as a foundational ingredient of evaluation and explanation of social states. Merit wants are not a normatively

R. Sturn (✉)
Institute of Public Economics, University of Graz, Universitaetsstrasse 15, 8010 Graz, Austria
e-mail: richard.sturn@uni-graz.at

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_16

empty box, as they were occasionally called [20]. Or so I will argue in this paper.

But why are merit wants not a firmly established part of modern normative economics, given that simplifying, but still meaningful notions are in principle suitable as a conceptual core of research programs? Is the ambiguous status of this concept perhaps caused by the excessive complexity lurking in the background of the variety of interpretations? Does under-conceptualized complexity prevent crucial issues from being made explicit at the proper level of analysis? If that is the case (and I will argue that it is the case to a certain extent), the kind of tension aptly summarized by Kaushik Basu [6, p. 220] should come as no surprise: "that some goods should be treated as merit goods most of us agree, but fail to express why we do so, with any rigour". Referring to the dubious status of merit goods, Nick Baigent [4, p. 301] observes that they "have always been postulated in an ad hoc way without any real justification."

In this paper I try to link the answer to the kind of questions sketched by way of introduction with making explicit three levels of analysis: call them for short (1) "behavioral limits of reason", (2) "higher order preferences", and (3) "collective choice". By means of (1)–(3), various interpretations of and approaches to merit wants can be brought into some kind of preliminary structure. The social choice aspects of merit wants has been addressed by Nick Baigent [4], suggesting a possibility of integration of merit wants in the framework of Arrowian Social Choice theory. Issues related to the second level of analysis (referring to the epistemological status of preferences in general and higher order preferences in particular) are discussed in another paper by Nick Baigent [5], entitled "Behind the veil of preferences". Behavioral limits of reason are the core of Munro's [23] discussion of merit wants, taking on board pertinent findings of experimental and psychological economics.

Discussion of the crucial issues underlying merit wants shows that it is misleading to think of the disputes regarding merit wants as exhibiting a divide between the clear water of individual sovereignty and the muddy mixtures of collectivism, paternalism, and authoritarian elitism (words which are often used when referring to the problematic status of merit wants). I am arguing that merit wants as discussed by Musgrave and others reveal a divide within the individualist camp, amongst economists and philosophers sticking to liberal principles. While conceptions of merit wants are certainly not without difficulties, steering clear of the problems which they are meant to address is neither equivalent to their solution, nor can it be justified in terms of "staying at the safe side."

Moreover, in a perspective emphasizing social choice as a dimension of merit wants, their recent revival in the wake of behavioral economics and libertarian paternalism (cf. e.g. Thaler and Sunstein [38, 39] and Munro [23]) must be regarded as conceptually unfinished business as yet. This revival is focused on findings of behavioral economics that provide evidence for persistent "mistakes" in individual choice behavior. To be sure, the methods and findings of behavioral economics allow for important progress regarding the empirics of context-dependent behavior

and pertinent limits of reason: this is a step forward in Mill's [21, V.xi] program of a situation-specific diagnosis of classes of circumstances rendering consumer sovereignty problematic. I will come back to that in Sects. 3 and 6, where some advantages of a conceptually complete theory of merit wants are suggested.

The remainder of this paper is organized as follows. In the following section, it is argued that merit wants are not an empty box. In Sect. 3, three levels of merit wants discussions are sketched, suggesting that each of the levels connotes a crucial question. Those questions (and the answers to them) may be seen as a first step to make visible the way in which alternative approaches to answering those questions are interrelated across various levels of analysis. Using that three-level scheme as interpretive background, Sects. 4 and 5 provide a selective overview of the pertinent literature. Section 4 deals with the pre-history of merit wants. The motivation for that rough and cursory overview is to show how old and pervasive the awareness of problems underlying merit wants conceptions is, and how various thinkers struggled to come to grips with its problems and to reconcile them with individualist frameworks. Section 5 offers some brief remarks on the history of merit wants. This supports the view that the merit wants concept may pose problems not so much on the grounds of lack of normative substance, but rather because it is complexly intertwined with "too many" potentially relevant ideas. The final section expands on some issues primarily related to collective choice and to identity/higher order preferences, complemented by some suggestions of how the different levels of merit wants may be interrelated. By way of conclusion, I suggest a tentative answer to the question of why merit wants may be a useful concept despite all difficulties.

2 Why Merit Wants Are Not an Empty Box

I am going to argue that merit wants (or at least the underlying problems) should be taken seriously. Nonetheless, three difficulties with the "traditional" merit want literature ought to be taken into explicit account. Those difficulties may have contributed to the perception that merit wants are an "empty box" and to the fact that the pertinent literature was limited to a "steady trickle", as John Head [17] put it.

First, in the literature we can find interesting and sometimes inspiring suggestions of how to deal with *various aspects of merit wants* (including various sorts of higher order preferences and community preferences, or the distinction between consumer/voter/reflective sovereignty suggested by Brennan and Lomasky [10]), but there is no unified theoretical foundation of merit wants which could serve as a focal point integrating pertinent discussions. Second, in the first decades after Musgrave [24], pertinent conceptualizations seemed difficult to integrate within the research programs of empirical economics. In certain respects, this obstacle is now overcome due to the progress made in behavioral economics. Third, merit wants were difficult to integrate in the framework of individualist welfare economics, a

problem that now may be alleviated by the progress of non-welfarist conceptions in normative economics.

Musgrave [24, p. 341] was well aware of those difficulties. In his first paper introducing merit wants, he suggested that an important aspect of merit wants could be accommodated in his Distribution Branch. Beyond that, he argues, this concept "introduces a new feature so far there is no place in our theoretical framework". Despite various attempts of integration by Musgrave and others, the ambiguity concerning the question whether and how this can be done constrained the usefulness of the concept throughout the following decades.

Authors such as Charles McLure [20] concluded that merit wants are "a normatively empty box", as the pertinent concepts either can be re-formulated in terms of individualistic market failure theory or else invoke inacceptable authoritarian or collectivistic premises. In the present paper, it is argued that merit wants neither are a normatively empty box, nor should they be considered as a concept of "essentially residual nature" [15], a view which is of course invited by the often-used "definition", according to which merit wants apply to all cases where consumer sovereignty is a problematic assumption.

The real question is not so much whether that box has any specific content at all, but whether it is packed with too many different approaches, too many important and difficult issues, and perhaps too many levels of analysis. Yet in order to reject the empty box-interpretation, two kinds of arguments have to be dealt with. The first concerns the possible reconstruction of certain apparent cases of merit wants in terms of standard theory. The second is related to the claim that merit wants (insofar they cannot be reconstructed along the lines of standard theory) are at odds with standard conceptions of individualism, as they rely on heavily collectivistic views of society.

Let us start with the second argument. Do merit wants indeed rely on collectivistic views incompatible with the kind of individualism which is a foundational part of normative and positive economics? The answer is no: at least some of the approaches making use of merit arguments need to be taken seriously from a "broadly" individualist perspective. Put another way, unless we rule "broad" conceptions of normative individualism out of court from the very beginning and stick to narrow ones, we have good reasons to be interested in merit wants. What I call a "broad individualism" simply amounts to rejecting the equation of individualism with claims according to which (in cases not covered by traditional market failure[1]) market-mediated individual choices are the only possible sources of normative authority. While "broad" individualism requires moving beyond narrow versions of consumer sovereignty, it does not entail philosophically more demanding conceptions of autonomy or positive freedom (cf. Berlin [8] for a crisp account of problems implied by the latter) or a rejection of welfarism. Notice in particular that "broad" individualism does not imply the kind of "broad" rationality (a notion suggested by

[1]For an even narrower version of subjectivist individualism endorsed by Libertarian/Austrian Economics, the qualification added in parenthesis must be dropped (cf. e.g. [18, pp. 3–26]).

Munro [23, pp. 5–6]) which includes the theoretical option in favor of a reflective stance with regard to goals (addressing the question whether goals are reasonable in terms of "true interest"), along with instrumental rationality (viz. consistency either in choices or in preferences). While that kind of "broad" rationality no doubt prepares the ground for interesting interpretations (notably those hinging on higher order preferences and multiple selves), analyzing discrepancies between "desires" and "satisfactions" as emphasized by Pigou [27] does not presuppose a fully-fledged reflective stance with regard to goals, but may be discussed at the level "mistakes" of instrumental rationality. The distinction between desires (wantings) and satisfactions (likings) is rather common in discussions on merit wants (see e.g. Head [17, p. 232]) and libertarian paternalism. Similar issues have been discussed in-depth in the context of utilitarian ethics. Utilitarianism incorporates various axioms capturing its consequentialist, welfarist, and individualist nature. But even utilitarians who reject discriminating between different kinds of preference satisfaction ("pushpin is as good as poetry") are not committed to the view that all individuals will (in any situation) understand their own interests best, and act accordingly.

If empirical evidence provided by behavioral economics shows some systematic violation of axioms capturing instrumental rationality, there is a prima facie reason to take merit policies into consideration. The case for such policies is made stronger if a definite cause triggering the violation can be identified and if the concomitant distortion can be addressed by certain policies. It is by no means clear that "broad rationality" needs to be invoked for this kind of argument. A less demanding notion of "broad individualism" may be sufficient, hinging only on the assumptions required for the diagnosis of "irrationality" (perhaps including implications for personal identity regarding the continuity of selves, but not necessarily a reflective self; see Sects. 3 and 6). Neither broad individualism nor broad rationality implies the imposition of collectivist values.[2]

I now move on to the other argument supporting the empty box view, claiming that the "interesting cases" of merit wants are generally reducible to cases of standard market failure. To be sure, there are some enlightening attempts showing that various apparent cases of merit wants are well captured by meaningful extensions of public goods or externalities. Head [17, pp. 240–7] includes an overview of such attempts: in one type of cases, shortcomings in individual decision making can be explained in terms of lack of information; lack of information (which can be assumed to be a public good) may be caused by its undersupply in the private-provision equilibrium. Hence public information policies (already stressed by Mill [21, V.xi]) are a remedy justified by standard public good-arguments.

[2]Broad rationality including reflective preferences can be taken into consideration in a manner which *takes individualist concerns seriously* and is combined with a clearly critical (but not a priori dismissive) stance regarding the normative status of reflective preferences. An example in case is the discussion of reflective preferences, p-preferences and m-preferences by Brennan and Lomasky [10], located in a merit wants context.

(Such a kind of argument could be probably extended to cases where framing effects are shown to distort individual choices: establishing a less distortive frame could be construed as a public good.) Other reconstructions include "psychic" externalities[3] or preference externalities, which inter alia were used to explain merit aspects of redistributive policies. Furthermore, one may argue that merit wants associated with what Musgrave (e.g. [25]) calls community preferences are often related to certain historically emerging solutions of public good problems (cf. [17, p. 246]). For instance, national defense is a public good in abstracto, but particular military arrangements and expenses can perhaps only be explained in terms of community preferences, reflecting the values, traditions and culture of some particular community.

To sum up: some kinds of prima facie merit-based public policies (providing useful information or better "framing") can be explained or justified as giving effect to the first-order preferences of individuals, who in absence of those policies are caught in a social dilemma. Institutional remedies suitable to overcome social dilemma situations (such as public good problems) often play an important role in policies designed to address merit wants-problems.

Successful exercises in clarification of that kind certainly are most welcome. Merit wants were often invoked in an ad hoc fashion, or in ways offering no resources to take adequate care of elitist-authoritarian implications. While the range of implications of the above-sketched reconstructions for the status of merit wants is far from obvious, they are in any case preferable to ad hoc-reasoning. But even if meaningful reconstructions are possible, some of the examples seem to suggest that merit wants-reasoning still has a role to play. Take the case of distribution as an example. As Musgrave (e.g. [25]) aptly observes, voluntary giving as well as redistribution through a voting procedure sometimes takes a specific "paternalistic" form, the donors favoring in-kind transfers instead of cash transfers. Now it is certainly possible to address this issue in terms of preference externalities. (I derive utility from seeing the recipients obtaining in-kind transfers such as educational vouchers, while I derive zero or even negative utility from seeing them getting money, which *they* would prefer to the in-kind transfers, and which I believe they are likely to spend on alcohol.) While this kind of reasoning may *explain* certain patterns and modes of redistribution,[4] important *normative* questions remain open. Should we accept whatever pattern emerges as a case of "Pareto optimal

[3] Andel [1, p. 635] credits Mugrave's graduate student Charles Tiebout for having been involved in suggesting an analysis of merit goods based on "psychic" externalities [40, p. 414].

[4] The contingent empirical fact that many people *factually endorse* an in-kind transfer policy similar to one which is, say, inspired by Rawls's [28] conception of basic goods or Sen's (e.g. [33, pp. 86–89]) capabilities and functionings could be taken as evidence that pertinent evaluative standards are supported by actual moral sentiments. That kind of support may be considered as important in the context of a theory of "justice as the first virtue of social institutions", as Rawls [28, p. 3] puts it.

redistribution"[5]? If yes, does this kind of consumer sovereignty also apply to other preference externalities, which came to be known as *nosy preferences* with regard to other people's way of life? John Stuart Mill [21, V.xi] categorically rejects policies which would give effect to such nosy preferences, even though (in the same section) he strongly advocates merit policies in the sphere of education. His pertinent arguments include the following: "Those who most need to be made wiser and better, usually desire it least, and if they desired it, would be incapable of finding the way to it by their own lights" [21, p. 868]. It seems obvious that making a distinction between mere nosiness (which carries with it, as Mill recognized, destructive potentials for a liberal society) and defensible "merit externalities" requires a step beyond the framework of consumer sovereignty, as it cannot be reconstructed in the standard externality framework. The involved issues cannot be meaningfully discussed without reference to what I call the second and third level of merit wants, including issues of higher order preferences and social choice. It may turn out that in-kind transfer policies are *justified* under certain assumptions which need to be discussed at those levels of argument.

3 Three Levels of Merit Wants Analysis

In the previous section, two kinds of arguments were scrutinized which support the position that merit wants are an empty box. It was argued that merit wants are not an empty box under the premise of broadly individualist frameworks of analysis and evaluation. This remains true if we take into account the explanatory potential of standard market failure theory in order to shed light on various cases of purported merit wants. Put another way, merit wants are not a normatively empty box, given that our starting point is the comparably cautious broadening of individualism described above, even if we take on board possibilities of reconstructing parts of their presumed domain in terms of public goods and externalities.

The concept of merit wants may not be an empty box, but is it useful? It appears as a catch-all conception summarizing aspects of human agency which cannot be addressed in frameworks confined to exogenous first-order preferences or choices revealed in a market context. How should we deal with the embarrass de richesses which seems to be on offer when considering the pertinent literature aiming at some sort of individualistic foundation for merit goods? The fragmented discussions include various aspects of

- robust and systematic "mistakes" in the choice behavior of many individuals,
- custodial choices,
- frame/context dependent and endogenous preferences,

[5] See Andel [1, p. 636] for a useful summary of the turn of the merit wants discussion towards the issue of so-called Pareto optimal redistribution in the late 1960s and early 1970s.

- "psychic" externalities, other-regarding and expressive preferences,
- concepts of multiple selves and higher order preferences,
- communal preferences and social choice,
- ethical values and conceptions of well-being and distributive justice invoking specific goods.

I suggest dividing the various discussions associated to merit wants into three broad categories:

1. "Limits of reason: This level primarily concerns the empirical identification of the limits of consumer sovereignty occasioned either by bounded rationality or by decision heuristics which are inadequate with regard to a given choice context;
2. "Higher order preferences": This includes normative issues, dealing with the quest for an adequate framework locating normative authority when "*broad*" rationality is taken into consideration, including a reflective stance regarding goals;
3. "Collective choice": The social choice level, making explicit the ways in which choices giving effect to merit wants are related to the dimension of social choice, including distinctions such as the one between tastes and values in Arrowian Social Choice.

As indicated above, each of the three levels connotes at least one issue of major importance:

1. "Classes of people vs. types of choice situations".
2. "Unavoidable paternalism vs. higher order preferences".
3. "Political choice as if it were market choice vs. social choice".

1. *Classes of people vs. types of choice situations.* In a social order based on enlightenment values of individualism and rationalism, the question needs to be addressed: When can we rely on people properly using their reason, and when not? A traditional answer puts the focus on classes of people: minors and people with clinical symptoms of mental disorder are subject to custodial choices and excluded from political decision making and freedom of contract. But the class-specific view of mental maturity was extended to adult women (married women in particular) and indigenous people. Both were widely (either implicitly or explicitly) regarded as unfit to use their reason in important domains of modern social life. Hence in much of the liberal era, the problems of the imperfections of human agency were dealt with in terms of diagnostic categories with clearly discriminatory connotations. By contrast, a strand of reasoning from Mill to modern behavioural economics approaches this problem in terms of "difficult" choice situations where humans tend to go wrong, without invoking distinctions of race, class and gender.
2. *Unavoidable paternalism vs. higher order preferences.* Sunstein and Thaler argue that paternalism is inevitable (see D'Amico ([15, Sect. 1], for a crisp summary): real-live choice does not take place in an undistorted state of nature, but is often subject to *circumstances* that influence choices in a way that fails to be

welfare-promoting: framing effects and status-quo biases are responsible for the effects of circumstances such as default rules, anchors, and the way in which the menu of choice is presented (neither of which should play a role according to standard choice theory). Put another way, some kind of "choice architecture" is in place anyway, whether produced by purposive private action (e.g. manipulative marketing) or as a spontaneous by-product of the evolution of markets and institutions. So why not improve this choice architecture in a welfare-enhancing way using the tools of libertarian (or soft) paternalism?

I do not argue that this kind of argument (as far as it goes) is off the mark. But Sugden's [36, p. 229] objection to the conception of normative economics implicit in the above-sketched perspective should be taken seriously: according to Sugden, the libertarian paternalist "presupposes a planner with the responsibility to collate information about individuals' preferences and then, guided by that information, to promote the overall social good." The core of Sugden's objection is the "single, neutral point of view" which must be assumed to guide the design of a benevolent planner engaged in the implementation of libertarian paternalist schemes. Sugden criticizes the underlying neutrality assumption, arguing that the planner always must rely on (somehow imposed) non-neutral normative judgments when determining whether individual choices are distorted or not. The diagnosis of distortions triggers his intervention and the provision of a "superior" choice architecture. The answer to Sugden's objection is not a watertight theory of the benevolent libertarian-paternalist planner. A more promising way of dealing with such objections is to be expected from challenging the linkage between merit wants and paternalism (whether libertarian, coercive, or otherwise) in a qualified way. Challenging the concept of paternalism in the context of the present problems will be accompanied by a richer concept of agency, including "broad" rationality as mentioned above. This can be achieved if a *credible account of multiple selves and higher order preferences* is provided—as a basis for models of self-correction and self-commitment, of dealing with one's own imperfections, and perhaps even of character-planning and sociability.[6]

3. *Market choice vs. social choice.* In his later writings on the subject, Musgrave (e.g. [25]) considers community preferences as the core of merit wants. There are good reasons for this emphasis. A first approach bringing to the fore what is at stake here is to be found in Kenneth Arrow's [2] distinction between "tastes" and "values" (referring to aspects of social states such as the overall distribution of income and wealth) in the context of Social Choice. In a perspective making explicit the institutional dimension, social choice conceptualized as value-based (or expressive) political choice (distinct from the taste-based choices typical for market allocations) can be considered in its relation to institutions which enable individuals to give effect to their attempts towards self-correction and self-

[6]In a programmatic article, Sen [31] discusses a whole range of issues pertinent to this level of merit want analysis, including the possible role of higher order preferences.

commitment, to their higher order preferences, or to their communal preferences. (See Brennan and Lomasky [10], for an analysis of markets and collective choice procedures as different environments with respect to the way in which the preferences of "acting" selves or "reflective" selves may find their articulation.) The controversial issue at this level is this: Can we really make a strong case for genuinely *social* choice in the above-sketched sense? The alternatives involved here can be grasped by considering positions such as the one defended by Robert Sugden. Sugden [36, p. 243] argues that the market implies "the privileging of the *acting self*—the self as buyer, seller and consumer, rather than the self as the maker of plans or as the source of reflective judgments about the well-being of the continuing person. Or more accurately, the market privileges the preferences of acting *selves*. Sugden defends a view which tends to identify the agent with her "acting self". This goes along with a vindication of markets and *reduced forms* of social choice which mimic markets in that they "privilege" the tastes of the "acting selves".

4 On the Pre-history of Merit Wants: Aristotle, Locke, John Stuart Mill

Merit wants more often than not were invoked in an ad-hoc fashion, but not all discussions of merit wants lack theoretical foundations. Moreover, some of those theoretical foundations make it clear that "paternalism" is a rather simplistic and perhaps misleading way to express the problems of the merit wants-concept. Summarizing those problems by means of notions like paternalism or collectivism appears even less convincing when we consider the pre-history of merit wants and related developments which are sometimes (but not always) explicitly linked to merit wants.

Merit wants are based on claims that individual tastes are insufficient as a basis of evaluation and/or explanation in certain classes of cases. The pre-history of merit wants indicates the enduring relevance of some of the pertinent concerns. It moreover provides some illustrations of how those concerns are expressed in a variety of theoretical frameworks.

Aristotle is remarkable for the comprehensive treatment of almost the whole range of issues relevant in the context of merit wants. First, according to Aristotle, pursuing the appropriate course of action is executively difficult: Aristotle devotes considerable attention to weakness of will (akrasia). Second, it is epistemically difficult: to determine the right course of action requires knowledge and judgmental powers presupposing complex learning processes. Those learning processes include practice and experience, culminating in the development of virtues which are conducive to true happiness. Third, the communal dimension looms large, articulated by the conception of humans as political animals; political live has an important place in Aristotle's universe of values. All in all, a view of human agency with its

certain imperfections and possible achievements emerges which seems congenial to the concerns coming to the fore in the discussions on merit wants. Aristotle's view of agency and sociability indeed was a source of inspiration for modern currents of thought for which ideas coming close to merit wants are of key importance: think of some versions of communitarian thought, of Martha Nussbaum's capabilities approach, or of Carl Menger, whose integration of communal wants in a staunchly individualist approach may be related to his Aristotelian background. A further aspect is perhaps even more important for modern discussions: the philosophical critique of Aristotelian virtue ethics (taking issue with (a) the conception of virtue connoting a kind of epistemological privilege and (b) the related universe of values which implies a thick concept of the good) may be relevant for establishing a defensible concept of merit wants in a society characterized by cultural pluralism, where thick concepts of the good are problematic. This critique hence exhibits some of the possible pitfalls of merit wants policies. It provides some background to the refinements of some common lines of critique focusing the "paternalistic" implications of merit wants: defensible merit want-policies must not be vulnerable to the objection that they end up with the attempt of one sub-cultural group educating the other(s).

The development of liberal thought from John Locke to John Stuart Mill should not be understood as an attenuation of individualism, but as a progressive shift of focus. Let us begin with Locke [19]. After emphasizing that only Adam was created with full powers of agency, and that it is difficult to determine whether actual persons are sufficiently endowed with agency-related powers, Locke's [19, §52–61] considerations are oriented towards the question, Which classes of people can and which cannot be relied upon as capable of properly using their reason? Locke's focus is on minors and people with clinical symptoms of mental disorder. As pointed out above, in much of the liberal era, the problems of the im-perfections of human agency were dealt with in a discriminatory fashion, excluding other classes of individuals as well.

Adam Smith's [34] theory of human agency presents the whole range of issues underlying merit wants in a specific way: from systematic biases such as overconfidence causing mistakes of instrumental rationality to the complex learning processes, including the virtues of the statesman as discussed in a part added to the 6th edition of the Theory of Moral Sentiments (1790, VI), where a non-technocratic idea of leadership in a polycentric modern society is suggested.

For sake of brevity, I skip Hegel in my historical sketch. Hegel's importance in the context of merit goods comes to the fore in W. Ver Eecke's [42, pp. 63–65] discussion of merit characteristics of various dimensions of social policy.

If we take Locke's briefly sketched discussion of agency-related powers as point of reference, Mill's [21] progressive shift introduces a new focus. While in some passages the older tendency to identify classes of agents with insufficient agency-related powers is shining through, Mill's innovative focus clearly is on types of situations (in which eventually any individual tends to act in ways which are not conducive to her welfare). Hence it is not without reason that Mill is one of the favorite references of authors dealing with merit wants (cf. e.g. [6] or [1]). The tools

of modern Behavioral Economics are suitable to pursue this agenda and to transform it into systematic research strategies.

Mill's [21, V.xi] Principles provide a subtle discussion of the limits of laissez faire. After mentioning protection against force and fraud as core functions of the state, Mill considers the regulation of external effects. Interestingly, he includes effects based on "the moral influence of example" in this discussion, remarking that such psychic externalities are not a proper foundation for public intervention: the prude who suffers during sleepless nights as he cannot get the images of the corruptive lifestyle of the lewd out of his mind, cannot hope for public regulation, unless that lifestyle causes tangible externalities as well (see the discussion on the tension between libertarianism[7] and the Pareto principle discussed by Sen [32, pp. 285–326]).

Protection against force, fraud and regulation of tangible externalities are seen as part of the basic rules of a System of Natural Liberty (to use Adam Smith's phrase). Apart from that, Mill discusses five further types of situations where some kind of public intervention may be justified. The onus of argument always falls on those who are in favor of the intervention, as Mill emphasizes in a famous passage. Interestingly, Mill first discusses some types of situations which are associated with "overruling" the judgment of individuals who, for various reasons, cannot be assumed to be the best judges of their interests (i.e. merit wants cases). Undeveloped judgmental powers, lack of opportunity to learn from experience and choices entailing far-reaching or/and irreversible commitments are cases in point discussed by Mill. Along those lines, Mill [22] also argues in favor of limits to contract freedom in the case of slave contracts, while undeveloped judgmental powers are invoked in the case of educational choices by Mill [21, p. 868]: "Those who most need to be made wiser and better, usually desire it least, and if they desired it, would be incapable of finding the way to it by their own lights."

The fourth type of situation, as Mill explicitly states, is associated with giving effect to the actual preferences of individuals: Mill says that the task of public governance in that class of cases is "not to overrule the judgment of individuals respecting their own interest, but to give effect to that judgment; they being unable to give effect to it except by concert, which concert again cannot be effectual until it receives validity and sanction from the law." This clearly is a specific formulation of the social dilemma structure that may emerge in cases of public good provision. As a fifth class, Mill discusses other-regarding choices in a distributive context together with aspects of empowerment, including policies with first-order effects on future generations. The latter are related to merit wants in a particularly complex way, insofar we are dealing with the preferences and opportunities of individuals not yet born.

The programmatic association of behavioral economics with libertarian paternalism (including the design of a tool-box of instruments allowing for better targeting of policies) is not merely a terminological provocation whose oxymoronic flavor provides genuine food for thought. It offers the perspective of policies that are better

[7]The status of minimal libertarianism is analyzed by Baigent [3].

targeted to the diagnosed problems as compared to traditional patterns of public intervention. "Better targeted" should not be understood in a merely pragmatic or technocratic sense. It refers to improved diagnosis of classes of choice situations in which merit policies are commendable. Declaring certain classes of individuals as unfit to choose on their own lights (motivating class-specific merit policies) is rendered obsolete to some extent. In this sense, the association of behavioral economics and libertarian paternalism is a program very much in the spirit of Mill.

Of all currents of thought in which issues related to merit wants played a role, the German language public finance literature is the most direct source of inspiration for Musgrave's conceptual development. Musgrave became conversant with enlightened versions of this tradition during his studies in Heidelberg, and was reminded of it by the contributions of another German-Jewish emigrant to the U.S., Gerhard Colm [13, 14], economic advisor to President Truman. In the German language public finance literature (which included Austrian and Swedish authors) a conceptual differentiation emerged in the course of the nineteenth and early twentieth century: disambiguating what was originally summarized by the notion of "collective needs" in two different classes of phenomena: (1) individual needs or wants which are accommodated by goods which are in different degrees non-rival and non-excludable, preparing the ground for Wicksell's and Lindahl's pertinent contributions and the literature on public goods; and (2) needs or wants which presuppose the existence of some form of community, even though they are felt by individuals (communal wants). A doctoral dissertation by Margit Cassel [12] comes close to a complete overview of the whole range of problems and pertinent concepts. Before Cassel [12], a discussion including Austrian, Italian and Scandinavian authors (notably Sax, Mazzola, Wicksell, and Margit Cassel's father Gustav) had successfully transformed the concepts inherited from German public finance within a marginal utility framework, including the terminological transformation eventually leading to the notion of "public goods" instead of social wants.

By way of conclusion of Sect. 4, a remark on Charles Taylor's [37, pp. 127–45] concept of "irreducibly social goods" is in order. Taylor's discussion is helpful in getting clear about the differences between public goods and the kind of merit wants which are directly related to a genuinely communal level. Taylor's arguments can be summarized as follows: Merit wants of this kind are accommodated by irreducibly public goods, i.e. by goods or services for which there are no private substitutes. The goods considered here are categorically lacking private substitutes (i.e. private substitutes are not merely unattractive for empirically contingent reasons, such as currently available technologies and relative prices). A fair distribution may be seen as an irreducibly public good in that sense, but further kinds of goods may correspond to that definition: goods which are related to goals/wants that make no sense in absence of a pre-existing community with its culture and its values. Think for instance of the goals guiding Britain's defense effort during the Battle of Britain, which can be addressed as national goals (this terminology is used by Colm [13, 14]). Hence there appears to exist a close connection between communal preferences (as discussed in the German Public Finance tradition) and Taylor's irreducibly social goods.

5 A Brief Retrospective on Merit Wants

Since Richard Musgrave coined the notion of merit wants[8] in 1956, a modest, but steady trickle of articles explicitly dealt with their conceptual clarification. Musgrave's pertinent contributions throughout the following decades show some shifts of emphasis (recapitulated by [1]). A succinct summary of conceptual issues is to be found in Musgrave [25].

All in all, in the more than 50 years after the introduction of merit wants by Richard Musgrave, its development exhibited considerable vicissitudes. It was not an unambiguous success story. Musgrave's original problem was posed at the explanatory level: a considerable share of public sector activity is difficult to explain solely on the basis of standard micro-based market failure theory. Initially, Musgrave's focus was public provision of rival and excludable goods and services: merit wants were seen as a possibility to explain the provision of such goods. Later he made clear that the merit dimension is independent of the degree of rivalry and excludability. More generally, the way in which merit wants were located in his overall theoretical framework of market failure seems to have changed over time (cf. [1]). In particular, his emphasis shifted to the issue of "leadership in a democratic society" and communal wants as the core of merit wants (see [25]), a perspective which is also pursued in a number of articles in the 1990s where his roots in the German language Public Finance tradition are better visible than in earlier work.

While Musgrave's discussions of merit wants are full of perceptive remarks throughout the various shifts of emphasis, he developed no unified theory of merit wants, let alone a coherent integration within the modern micro-based theory of the public sector and modern normative economics. Despite various attempts to define them in a more specific way, the shifts of emphasis may have contributed to the view (explicit or implicit) that they are best understood as a residual category, covering all the cases where consumer sovereignty fails to be a convincing concept either as basis of normative authority or of explanation. Such a residual view provides few resources against invoking merit want-arguments in an ad-hoc fashion. Hence as a potential guide to matters of public policy, it is notoriously associated with the dangers of elitist/authoritarian imposition of values, insofar there is no justifiable answer to questions such as: Whose values ought to matter? Which values ought to matter, and how? Having in mind suchlike objections, Musgrave himself was rather cautious regarding the scope and usefulness of the merit wants-concept.

[8] A referee suggested to comment on the terminological ambiguity of merit wants vs. merit goods. In keeping with the German-language Public Finance tradition, in his early contributions Musgrave used the notion of merit wants as well as social wants (for public goods). As he was eager to get rid of terminological heritage which might hinder the development of a unified micro-based theory of market failure, he soon adopted the now common terminology. This terminological shift included merit goods, even though the latter are mostly explained in terms of properties of individual values or preferences, whereas rivalry and excludability are properties of goods rather than of wants.

Yet there exists respectable work which contributed to the conceptual clarification of merit wants. In parts of it, the roots of the merit wants concepts (owing much to German Public Finance and Gerhard Colm in particular) is made explicit. Geoff Brennan and Loren Lomasky [10] were among the first who contributed to the conceptual development of the communal wants dimension of merit wants beyond Colm [13, 14]. They explicitly referred to Colm's distinction between homo oeconomicus and homo politicus (and Colm's corresponding distinction between market and politico-fiscal processes), developing those ideas in the context of a democratic institutional perspective with an explanatory agenda. Nick Baigent's [4] attempt to integrate merit wants within Arrowian Social Choice theory can be seen as the counterpart in the sphere of normative economics. Colm emphasized the conceptualization of the judgments of homo politicus as referring to overarching national goals (e.g. defense or education policies) in a specific way. Arrow's individual "values" (complementing the tastes of consumer theory) refer to aspects of social states other than individually consumed quantities of goods. While the analytical difference between the two kinds of distinction would require some further discussion, their main thrust is somewhat similar.

Brennan and Lomasky [10] introduced distinctions such as the one between m-preferences (typically related to outcomes in a market environment), reflective preferences (which are not immediately outcome-oriented in the way m-preferences are) and p-preferences, which are "intimately related" to reflective preferences as political choice processes typically allow for the expression of preferences with small expected costs. Kaushik Basu's [6] distinction between actual choice and retrospective choice is based on a different dimension of multiple selves, namely between present and future selves. This distinction is important in particular in combination with circumstances such as choices related to learning processes and important irreversible consequences, discussed already by Mill [21, 22]. Mill is also a source of inspiration in Basu's [7] more recent writings on the limits of the "principle of free contract".

John Head [17] provides the most comprehensive survey on the literature dealing with conceptual issues.[9] This survey does not yet reflect the perspectives of re-invigoration of that concept made available by advances in behavioral economics. The latter are explicitly combined with conceptual work on merit wants in Alistair Munro's monograph [23] entitled "Bounded Rationality and Public Policy."

A different level of merit goods is emphasized by W. Ver Eecke [41, 42], whose discussion is framed by a constitutional perspective based on a voluntary exchange conception of public goods provision. According to Ver Eecke, the specific difference of merit goods is that they cannot be based on voluntary exchange, as their provision implies losers. Ver Eecke starts with a rather widely shared justification of markets as social arrangements giving effect to individual preferences. Merit

[9]Other authors using the concept of merit wants were not so much interested in conceptual issues but in implications not least for the economics of taxation. See e.g. Pazner [26], Besley [9], Schroyen [29], Capéau and Ooghe [11].

wants in turn are justified as a specific class of possibility conditions either (along "Kantian" lines) for the functioning of market economies or (along "Hegelian" lines) for preserving a sufficient degree of freedom for all individuals in a market society, notably by means of social policies in the face of economic hardship. Both kinds of conditions are characterized by the impossibility of a policy design which would render them beneficial for everybody. The implementation of pertinent policies is accompanied by losses of some individuals.

The issue of distribution has ever been present in Musgrave's discussions on merit wants. In Ver Eecke's writings, it is put to the center of stage. His idea of distribution-sensitive "possibility conditions" is interesting. Yet two questions need further discussion in order to prevent overstretching the concept: (1) How does all this relate to the well-known limitations of a Wicksell-Lindahl voluntary exchange approach which can be discussed in a conventional public good-framework? (2) How does it relate to analyses of strategic structures representing social dilemma situations in which certain socially desirable "solutions" are available, but are not symmetric in terms of individual advantages?

6 Conclusion: Towards a Conceptually Complete Theory of Merit Wants?

Taking stock of some contributions to the "steady trickle", approaches seeking to provide theoretical clarification fall into three distinctive classes: (1) the diagnostic level of choice frames, contexts, heuristics and bounded rationality, (2) various versions of higher order preferences, and (3) various ways of associating merit wants with the logic of collective choice.

An integrated (conceptually complete) theory of merit wants across the three levels is complex and difficult. But is it attractive as a research agenda? To be sure, those three levels were present from the beginning in the writings of Richard Musgrave and especially Gerhard Colm [13, 14], who was an important source of inspiration for Musgrave. Taken together, Colm's and Musgrave's pertinent writings suggest that the analysis of institutionalized mechanisms (presumably) giving effect to merit wants conceptually and practically will involve all three levels. But this is not more than a hint that a conceptually complete theory of merit wants could be a plausible research agenda. Hence by way of conclusion, I try to provide some additional hints regarding the attractiveness of such an agenda.

The present revival of merit wants is primarily triggered by developments in behavioral economics. They by and large entail a focus on level (1), accompanied by some discussions at the level of (2). Munro [23, p. 6] draws a map of merit wants-aspects composed of four regions related to various diagnoses of limits of reason; i.e., the map *prima facie* refers almost exclusively to (1):

(i) Defective telescopic faculties;
(ii) Defective information processing;

(iii) Frame dependent preferences; and
(iv) Preference misalignment.

Given the progress of behavioral economics, (i), (ii) and (iii) may seem to be fairly well-explored territory. Preference misalignment (iv) is located in the core area of Munro's map and can be seen as still partly uncharted territory. In Munro's exposition, (iv) seems to be the residual region of merit wants; it is in keeping with the residual character which is sometimes attributed to merit wants as a whole.

Behavioral economics brought about unambiguous progress on the diagnostic level; there can be little doubt that its tools and findings will keep a firm place in the economist's toolbox. Yet upon closer inspection some loose ends become apparent which invite further thought. Here are three examples illustrating that claim. First, the problems related to (i) ("defective telescopic faculties") may raise the question of whether or not issues of a temporal sequence of selves are involved, where later selves may regret some irreversible decisions of the former selves (e.g. to forego educational opportunities, preferring the pleasures of consumption). Basu [6] reconstructs merit wants along those lines. Corresponding merit want policies may be understood in terms of giving effect to the preferences of later selves. This sounds plausible, but there is a problem: according to such policies, the preferences of later selves eventually are made to trump those of the earlier selves. (Why) Can this be legitimate? While there are arguments supporting public policies acting as advocates of later selves (Basu aptly draws on Mill [21, V.xi]), the issue is not so simply resolved. Mill's arguments include irreversibility and systematic lack of experience-based imagination of some aspects of future states. Under certain circumstances (indicated by Mill), the present self cannot be expected to be an effective advocate of the later self.

Notice though that such arguments favor some conceptions of personal identity and exclude others. Consider, for instance, the foundational discussion on "libertarian paternalism" as a program motivated by certain cases of limits of reason. Robert Sugden's (e.g. [35, 36]) suggestion of an interpretation of consumer sovereignty which does not hinge upon coherent preferences, along with the related idea of "privileging the acting self" mentioned in the above, not only amounts to a rejection of libertarian paternalism, but also of other kinds of merit wants policies motivated by the arguments just sketched.

It can be argued that this provides a motivation for a conceptually complete account of merit wants as a research agenda. Sugden's explicit critique of libertarian paternalism (and the implicit critique of a larger class of merit policies) entails a specific conception of personal identity as a sequence of time-slice selves (for a summary, see D'Amico [15, Sect. 2]). At a collective choice-level it entails a specific conception of contractarianism treating political choice *as if it were market choice*. Hence considering discussions on merit wants which first seem to be confined to issues of bounded rationality and the corresponding toolbox of libertarian paternalism at level (1), a significant divide pertinent to levels (2) and (3) becomes visible. Considering those discussions (for an overview see [15]), it becomes obvious that any kind of systematic suggestion regarding the evaluative

implications of behavioral findings (such as Sugden's interpretation of consumer sovereignty which does not rely on coherent preferences) explicitly or implicitly hinges on certain answers to the crucial questions to be posed at levels 2 and 3. Ultimately, the divide concerns the status of collectives and collective choice, and it is related to different conceptions of personal identity and agency. The nature and the implications of the divide cannot be properly addressed when the third level, the level of social choice is not taken into explicit consideration. Hence we have to move to level 2 and to level 3 for a proper discussion of the underlying problems, unless we subscribe to a kind of naïve technocratic conception of policy which is difficult to defend, given a pluralistic and polycentric society.

Integrating levels 2 and 3 also seem to make sense in cases such as nosy preferences. Behavioral economists now do experiments where nosy preferences play a role: people deviate from what they are expected to do in the standard model, for reasons that could be circumscribed with "nosy preferences" (see [30] for a brief summary of findings). Yet the normative implications of such findings are unclear, motivating a more encompassing theorization including levels 2 and 3. Even if "nosy preferences" can be shown to be efficiency-enhancing in some cases (e.g. when people punish free-riders) and to inefficient situations in others, the normative status of the "nosy preferences" as such needs to be clarified, as well as the procedural legitimacy of mechanisms either giving effect to those nosy preferences or laundering them.

Last but not least, the scope of frame dependence may be sufficiently wide as to make integration of the three levels sometimes indispensable: the market and the political forum may be regarded as two different decision contexts—and as two different behavioral frames. In an essay entitled "The market and the forum", Jon Elster [16] explains some possible backgrounds of such contextual differences (but cf. also the paper by Brennan and Lomasky [10], quoted above). Communal wants with their cultural connotations invoke a specifically far-reaching form of frame-dependence, insofar they are presupposing the existence of the specific frame of a given community (cf. Cassel [12, §§16–22]).

To sum up: Important progress at the diagnostic level notwithstanding, and acknowledging the multifaceted aspects of pertinent diagnosis of the limits of reason (see Munro [23, pp. 5–6]), discussions concerned with the integration of recent behavioral findings within normative economics indicate that steering clear of levels 2 and 3 leaves open crucial questions regarding the status of merit wants and related concepts. A conceptually complete theory of merit wants encompassing the three levels is no doubt complex. But it is an agenda which is foreshadowed in pertinent work (particularly by Musgrave and Colm), and which could add considerable leverage to the advances made in behavioral economics.

In the years after Musgrave [24], merit wants served as a reminder that public goods and externalities may not accommodate all explanatory challenges with regard to non-market goods and public interference. Colm and Musgrave framed the problem more or less in this way. Moreover, it served as a reminder that the concepts of New Welfare Economics may not solve all evaluative problems.

Research programs such as behavioural economics and non-welfarist Social Choice have changed the situation. Merit wants could now serve as a heuristic device, or as a focal concept encompassing and combining the three above-sketched levels of analysis. If it is at all useful today, it is not useful as a minor chapel off the main naves of public economics, normative economics and Social Choice. It is not useful as a concept suitable to discuss a residuum of some abstruse cases. If it has a function, its scope is more encompassing and foundational.

Acknowledgements The paper has benefitted from comments by two anonymous referees and Maxime Desmarais-Tremblay.

References

1. Andel N (1984) Zum Konzept der meritorischen Güter. Finanzarchiv 42:630–648
2. Arrow K (1963) Social choice and individual values, 2nd edn. Yale University Press, New Haven
3. Baigent N (1981) Decompositions of minimal libertarianism. Econ Lett 7:29–32
4. Baigent N (1981) Social choice and merit goods. Econ Lett 7:301–305
5. Baigent N (1995) Behind the veil of preference. Jpn Econ Rev 46:88–101
6. Basu K (1975/1976) Retrospective choice and merit goods. Finanzarchiv 34:220–225
7. Basu K (2011) Beyond the invisible hand: groundwork for a new economics. Princeton University Press, Princeton
8. Berlin I (1969) Four essays on liberty. Oxford University Press, Oxford
9. Besley T (1988) A simple model for merit good arguments. J Public Econ 35:371–384
10. Brennan G, Lomasky L (1983) Institutional aspects of 'merit goods' analysis. Finanzarchiv 41:183–206
11. Capéau B, Ooghe E (2003) Merit goods and phantom agents. Econ Bull 8:1–5
12. Cassel M (1924) Die Gemeinwirtschaft oder Die Gründe einer öffentlichen Haushaltung. Inaugural-Dissertation. A. Deichertsche Verlagsbuchhandlung Dr. Werner Scholl, Leipzig
13. Colm G (1955) Essays in public finance and fiscal policy. Oxford University Press, New York
14. Colm G (1965) National goals analysis and marginal utility economics. Some non-technical comments on a highly technical topic. Finanzarchiv 24:209–224
15. D'Amico D (2009) Merit goods, paternalism, and responsibility. Mimeo, University of Pavia, Pavia
16. Elster J (1989) The market and the forum. In: Elster J, Hylland A (eds) Foundations of social choice theory. Cambridge University Press, Cambridge, pp 103–132
17. Head J (1991) Merit wants: analysis and taxonomy. In: Eden L (ed) Retrospectives on public finance. Duke University Press, Durham, pp 229–252
18. Hoppe H-H (1993) The economics and ethics of private property. Kluwer, Boston
19. Locke J (1690) Two treatises of government. Awnsham Churchill, London
20. McLure C (1968) Merit wants: a normatively empty box. Finanzarchiv 27:474–483
21. Mill JS (1848) Principles of political economy. Longman, London
22. Mill JS (1859) On liberty. Longman, London
23. Munro A (2009) Bounded rationality and public policy. A perspective from behavioural economics. The economics of non-market goods and resources, vol 12. Springer, Heidelberg
24. Musgrave R (1956/1957) A multiple theory of budget determination. Finanzarchiv 17:333–343
25. Musgrave R (1987) Merit goods. In: Eatwell J, Milgate M, Newman P (eds) The new palgrave: a dictionary of economics. Macmillan, London, pp 452–453
26. Pazner EA (1972) Merit wants and the theory of taxation. Public Finance 27:460–472

27. Pigou AC (1932) Economics of welfare, 4th edn. Macmillan, London
28. Rawls J (1971) A theory of justice. Harvard University Press, Cambridge
29. Schroyen F (2003) An alternative way to model merit good arguments. Paper, Department of Economics, Norwegian School of Economics & Business Adminstration, Bergen
30. Schulz JF, Thoeni C (2013) Paternalismus, Rationalität, systematische Fehler, nudges: Befunde der experimentellen Ökonomik. Jahrbuch für normative institutionelle Grundfragen der Ökonomik 12:63–82
31. Sen A (1977) Rational fools: a critique of the behavioural foundations of economic theory. Philos Public Aff 6:317–344
32. Sen A (1982) Choice, welfare and measurement. Blackwell, Oxford
33. Sen A (2002) Rationality and freedom. The Belknap Press of Harvard University Press, Cambridge
34. Smith A (1759, 1790, [1976]) The theory of moral sentiments. In: Raphael DD, Macfie AL (eds) Glasgow edition of the works and correspondence of Adam Smith, vol 1. Clarendon Press, Oxford
35. Sugden R (2004) The opportunity criterion. Consumer sovereignty without the assumption of coherent preferences. Am Econ Rev 94:1014–1033
36. Sugden R (2008) Why incoherent preferences do not justify paternalism. Const Polit Econ 19:226–248
37. Taylor C (1995) Philosophical arguments. Harvard University Press, Cambridge
38. Thaler R, Sunstein C (2003) Libertarian paternalism is not an oxymoron. Univ Chicago Law Rev 70:1159–1202
39. Thaler R, Sunstein C (2008) Nudge. Yale University Press, New Haven
40. Tiebout CM, Houston DB (1962) Metropolitan finance reconsidered. Rev Econ Stat 44:412–417
41. Ver Eecke W (2007) An anthology regarding merit goods. Purdue University Press, West Lafayette
42. Ver Eecke W (2013) Ethical reflections on the financial crisis 2007/2008. Springer briefs in economics. Springer, Berlin

An Extraordinary Maximizing Utilitarianism

Jonathan Riley

Abstract This chapter interprets John Stuart Mill's liberal version of utilitarianism, which is extraordinary in at least three respects. First, Mill distinguishes among different kinds of utilities conceived as pleasant feelings (including relief from pain) of different intrinsic qualities irrespective of quantity. A feeling of security associated with the moral sentiment of justice is said to be higher in quality as pleasure than any competing kind of pleasure, where justice is conceived in terms of a social code that distributes and sanctions equal rights and duties for all who have a voice in constructing the rules. Second, the utilitarian aggregation procedure is restricted to this higher moral kind of utility and may be depicted as a social welfare functional which operates within a limited sphere of morality and law. The sole purpose of the aggregation procedure is to generate an optimal social code of justice so that individuals are then free from coercive interference to act and pursue non-moral kinds of pleasures in accordance with their optimal rights and duties recognized under the code. Finally, Mill never discusses a standard utilitarian aggregation mechanism and seems instead to have in mind a democratic voting procedure, which can be seen as a purely ordinalist utilitarian procedure, for aggregating over the higher moral kind of utilities expressed by moral individuals who are competently acquainted with the different kinds of utilities.

Keywords Art of living • Borda procedure • Democratic voting • Freedom • Happiness • John Stuart Mill • Justice • Kinds of utility • Optimal code • Pleasure • Social welfare functional • Utilitarianism

1 An Extraordinary Utilitarianism

Utilitarianism, as commonly understood by economists and philosophers ever since the late Victorian period when Henry Sidgwick and early neoclassical economists such as Stanley Jevons, Francis Y. Edgeworth and Alfred Marshall established themselves as the leading utilitarian thinkers and rejected John Stuart Mill's

J. Riley (✉)
Department of Philosophy, Tulane University, New Orleans, LA 70118, USA
e-mail: jonriley@tulane.edu

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_17

extraordinary version of the doctrine as incoherent, is a deeply flawed doctrine which is unable to capture our considered beliefs about justice and equal rights.[1] This flawed doctrine, which may be referred to as standard utilitarianism, comes in different forms, including, among others, act utilitarianism, rule utilitarianism, and an indirect approach that relies on a non-utilitarian decision procedure (which might assign intrinsic importance to moral rights, for instance) to achieve the standard utilitarian goal of maximizing the sum total of utility. Despite their differences, these different forms of standard utilitarianism all endorse two assumptions: first, utility (however defined, as happiness, for example, or as preference-satisfaction, or as an index of some objective list of valuable things) is homogenous in quality and so varies only in terms of quantity; and, second, individual utilities are cardinal and comparable so that an incontestable sum total of utility can be calculated for any feasible option, at least in principle, as is required to find an option that maximizes the sum total. Without such rich utility information, standard utilitarianism is unworkable.[2]

Most scholars, even those who are sympathetic to his project of providing a utilitarian foundation for liberal justice, attempt to read Mill as a standard utilitarian.[3] But Mill's version of maximizing utilitarianism as presented in Utilitarianism [1861] is so different in structure from standard utilitarianism that it obliterates the traditional understanding of what maximizing utilitarianism is. The most important difference is that Mill rejects the assumption that utility, which he conceives of as happiness in the sense of pleasure including relief from pain, is homogenous in quality across its various sources. Instead, he puts forward his controversial doctrine of higher pleasures, which holds that some kinds of pleasant feeling have a much higher intrinsic value or quality than others do as pleasure, irrespective of quantity. In his view, "the pleasures of the intellect, of the feelings and imagination, and of the moral sentiments" have "a much higher value as pleasures than . . . those of mere sensation."[4] Indeed, I shall argue that, for him, the pleasures of the moral sentiments,

[1]For a compelling critique of utilitarianism as commonly understood from the standpoint of our considered beliefs in justice, see Will Kymlicka, Contemporary Political Philosophy: An Introduction, 2nd ed. (Oxford: Oxford University Press, 2002), pp. 10–52, and the many references cited therein. See, also, Amartya Sen, On Ethics and Economics (Oxford: Basil Blackwell, 1987); and Sen, The Idea of Justice (Cambridge, MA: The Belknap Press of Harvard University Press, 2009).

[2]For further critical discussion of standard utilitarianism and its central assumptions, see Jonathan Riley, "The Interpretation of Maximizing Utilitarianism," Social Philosophy and Policy 26 (2009): 286–325.

[3]Perhaps the most frequent strategy is to read Mill as some type of rule utilitarian. See, for example, John Rawls, Lectures on the History of Political Philosophy, ed. Samuel Freeman (Cambridge, MA: Harvard University Press, 2007), pp. 249–316; and Dale E. Miller, J.S. Mill (Cambridge: Polity Press, 2010).

[4]John Stuart Mill, Utilitarianism [1861], in John M. Robson, gen. ed., Collected Works of J.S. Mill [CW], 33 vols. (London and Toronto: Routledge and University of Toronto Press, 1963–91), vol. X, chapter II, paragraph 4, p. 211. All references are to this edition. Henceforth, I shall provide references in abbreviated form, as follows: Mill, Util II. 4, p. 211.

and in particular of the sentiment of justice, have a much higher value as pleasures than any competing kinds of pleasures.

Mill emphasizes that he thinks with mankind "that justice is a more sacred thing than policy, and that the latter ought only to be listened to after the former has been satisfied."[5] He also argues that there is not a single principle or rule of justice but many, and that the plural maxims can come into conflict: "justice is not some one rule, principle or maxim; but many, which [may conflict]."[6] Unlike pluralists who assert that reason may be unable to resolve these conflicts, however, he insists that reasonable choices are those that may reasonably be expected to maximize the sum total of happiness. Even so, it will emerge that in light of his understanding of how a utilitarian must go about calculating a sum total of utility, he in effect claims that an open democratic process with suitable checks and balances is needed to determine what justice requires in any situation. What Mill is mainly concerned to rebut is the intuitionist account which holds that dictates of justice can be recognized by simple introspection, without any need to go through a democratic process in which the members of the community each have a voice in the matter: "if justice be totally independent of utility, and be a standard per se, which the mind can recognize by simple introspection of itself; it is hard to understand why that internal oracle is so ambiguous."[7]

Unfortunately, utilitarians have generally followed Sidgwick and his student G.E. Moore in maintaining that the higher pleasures doctrine is incompatible with utilitarianism. Anti-utilitarians have little incentive to question the consensus. Millian qualitative superiorities rely on non-utility values to draw the distinctions of quality, it is commonly charged, contrary to the utilitarian principle that utility is the sole basic value. It follows that Mill is at best some form of value pluralist if he is not merely confused. But the common charge is misplaced. Mill's claim that pleasure exhibits differences of quality irrespective of quantity is not inconsistent with utilitarianism, and the claim deserves careful study. Moreover, unless the doctrine of higher pleasures is properly taken into account, it is impossible to understand the way in which his extraordinary maximizing utilitarianism seeks to bring about "an existence exempt as far as possible from pain, and as rich as possible in enjoyments, both in point of quantity and quality."[8]

A second difference between Mill's utilitarianism and standard utilitarianism is that, even for individual utilities of the same kind or quality, Mill does not make the standard assumption that the utilities are, at least in principle, cardinally measurable

[5]Mill, Util V. 32, p. 255. The quoted sentence is one of many indications that Mill assigns absolute priority to considerations of justice over considerations of mere general expediency or policy. As he also says: "moral rules [of justice] which forbid mankind to hurt one another [are] . . . more vital to human wellbeing than any [policy] maxims, *however important*" (Util V.33, p. 255, emphasis added).

[6]Mill, Util V.27, p. 252.

[7]Util.V.26, p. 251.

[8]Util II.10, p. 214.

and interpersonally comparable, as is required to calculate an objective sum total of utility. Instead of assuming that individuals' feelings of pleasure can be measured and compared so that they can be added together in an incontestable mechanical fashion, he apparently assumes only that purely ordinal utility information is available. In other words, the only information about people's pleasant feelings of any kind which he relies on is the quite impoverished information contained in their distinct and typically conflicting individual preference orderings defined over the feasible sources of that kind of enjoyment, taking for granted that a rational individual prefers more pleasure to less of the same kind. As a hedonist, Mill believes that competent individuals are ultimately motivated by their expectations of pleasure, including freedom from pain, so that their preference rankings reveal their (possibly mistaken) expectations of pleasure. Each individual is assumed to rank the feasible sources in accord with his rough estimates of the amount of pleasant feeling of the relevant kind which he expects from the sources. (Since Mill's hedonistic idea of utility differs from the modern technical idea of utility as the value of a mathematical function that represents a preference ordering, it might make sense to rename the former as an idea of welfare so as to avoid confusion. For further discussion, see Jonathan Riley, "Welfare (philosophical aspects)," in James D. Wright, chief ed., International Encyclopedia of the Behavioral and Social Sciences, 2nd edition (Amsterdam: Elsevier), forthcoming. I shall not pursue that strategy here, however. In this regard, it must not be thought that hedonistic utility is necessarily incompatible with technical utility. Under hedonism, the individual's preferences are motivated by his expected pleasures from the feasible options, and a rational agent prefers more pleasure to less of the same kind. His preference ranking of a given kind reveals his expected utilities from the sources, where utility means pleasant feeling of that kind. If the individual's ranking is an ordering, that is, a complete, reflexive and transitive ranking of the feasible options, then the ordering may be represented by a purely technical utility function, where utility is not pleasure but merely a numerical representation of the preference ordering without reference to what motivates it. So there is no incompatibility if hedonism is a true psychology and individuals are motivated to form preference orderings rather than, say, inconsistent preferences.) Even if attention is restricted to pleasant feelings of the same quality, therefore, it is misleading to read in the traditional manner Mill's constant endorsement of conduct that increases, or tends to increase, the sum total of happiness. It is misleading because he never assumes that enjoyable feelings can be cardinally measured and interpersonally compared so as to yield a precise sum total that all reasonable people must accept. Thus, an unorthodox way of reading the sum total criterion must be found, even apart from the need to make "proper allowance" for different kinds of utilities.

How, then, should we understand Mill's qualitative distinction between higher and lower kinds of pleasant feelings? What does he mean that "proper allowance" must be made for the different kinds in order to maximize the general happiness? And how should we interpret his suggestions that, for any given kind of pleasant feeling, the sum total of that kind of pleasure ought to be maximized consistently with making proper allowance for the different kinds? How is it even possible to

maximize the total quantity of any kind of pleasure subject to the constraint that proper allowance must be made for the different kinds of pleasure?

2 Higher Pleasures

When he introduces his doctrine of higher pleasures in Utilitarianism, II, Mill says that a higher kind of pleasant feeling is superior in quality to a lower kind, regardless of the quantities of the different kinds of pleasures.[9] Qualitative superiority means intrinsic superiority, that is, the higher of two kinds of pleasant feelings is intrinsically more valuable as pleasure in virtue of its felt quality, regardless of quantity. This can in turn be interpreted as infinite superiority in the sense that even a bit of higher pleasure is more valuable as pleasure than any quantity of lower pleasure, no matter how much lower pleasure is amassed. No amount of lower pleasure can ever be equal in value to a higher pleasure. It is important not to confuse the claim that qualitative superiority means infinite superiority with the entirely distinct claim that humans are capable of experiencing limitless pleasure. When he says that competent people who have experienced both kinds prefer the higher pleasure to any amount of the lower pleasure which "their nature is capable of," Mill is taking for granted that humans are only capable of experiencing finite pleasures rather than unlimited pleasures of any kind.[10]

To make proper allowance for the different kinds of pleasant feelings, the different kinds must be arranged into a hierarchy such that a higher quality of pleasant feeling takes absolute priority over a lower quality in cases of conflict. The different kinds of pleasure have different sources, and these different sources can be treated as different aspects or features of possible outcomes. An outcome that is reasonably expected to bring any amount of higher pleasure must be ranked above an outcome that brings only lower pleasures, no matter how much lower pleasure is expected. Individuals who have developed the capacities required to competently experience the different kinds of pleasures do reveal such preferences, Mill insists, or, if there is disagreement, the majority of them do.

Although I cannot discuss in detail the much-maligned doctrine of higher pleasures, a couple of points deserve emphasis. First, despite the scorn of numerous critics that arises in large part from their own misunderstandings, Mill's doctrine can be interpreted in such a way that it becomes highly plausible. Given a suitable interpretation, most if not all people who are competently acquainted with the

[9]Mill, Util II.5, p. 211. Mill goes on to draw an important distinction between happiness and contentment in the next paragraph.

[10]For different interpretations of the higher pleasures doctrine which cannot be squared either with hedonism or with Mill's texts, see, e.g., Rawls, Lectures on the History of Political Philosophy, pp. 258–65; L.Wayne Sumner, Welfare, Happiness, and Ethics (Oxford: Clarendon Press, 1996), pp. 87–112; and Miller, J.S. Mill, pp. 35–36, 54–70.

different kinds of pleasant feelings arguably do confirm the qualitative ranking suggested by Mill. In this regard, the ranking of the different kinds of pleasures in terms of quality is not a purely subjective ranking. Rather, the different kinds of pleasant feelings are latent in human nature, with the higher kinds awaiting discovery by those who develop their intellectual, moral and practical capacities. The preferences or judgments of those who have developed the capacities needed to experience the higher kinds merely confirm the natural qualitative ranking as opposed to create it.

Second, the suggestion of a natural qualitative ranking is only plausible if we can identify a ranking which is endorsed, at least implicitly, by most if not all competent adults with suitable experience. But what specific qualitative ranking does Mill propose? As already indicated, he maintains as a first step that the pleasant feelings inseparably associated with the workings of our higher mental faculties are qualitatively superior to the inchoate physical sensations of pleasure registered by our body, that is, by our "animal nature" when assumed to be "disjoined" from our mental capacities. Humans, like other animals, experience simple physical sensations of pleasure and pain automatically through the nervous system, independently of the will and other higher mental faculties. Mental pleasures are more complex than simple physical sensations of pleasure, he argues, because any mental pleasure is a quasi-chemical combination of various ingredients, including an idea of some object or activity together with physical sensations of pleasure, or their traces in memory and imagination, expected from that object or activity. These various ingredients can melt together so that the mental pleasure feels like a whole new feeling with its own emergent properties, including the property of qualitative superiority over the pleasant physical sensations among its ingredients.(Cognitive scientists suggest that consciousness itself is a higher-level phenomenon that emerges from complex interactions among the various components of a neural network. Just as we may never be able to explain how or why mental states emerge from interactions among components of the body, so we may never be able to explain how or why higher feelings of pleasure emerge from interactions among their ingredients including lower pleasures. We can only observe through introspection that consciousness and higher pleasures do in fact exist.) The ingredients typically vanish from our consciousness as separate elements: the pleasant physical sensations become inseparably associated with the idea of the object or activity so that we forget that the sensations of pleasure and the idea are separate elements, unless the mental feeling is subjected to a psychological analysis.

Mill's view seems to be that any person capable of experiencing the mental kind of pleasures will not voluntarily sacrifice even a bit of mental pleasure for any amount of the mere bodily kind of pleasure. Since mental pleasures are complex feelings that already include pleasant physical sensations or their traces among other ingredients, however, the only way mental and bodily kinds can come into conflict is when mental pleasure must be sacrificed altogether in return for a greater amount of the mere physical sensation of pleasure that is already a component of the mental

pleasure.[11] In short, the issue is: would anyone who can exercise his mental faculties voluntarily give them up, along with the mental pleasure they make possible, in return for any amount of the purely sensual pleasure above and beyond that already contained in his mental pleasure? Mill's negative answer is plausible, keeping in mind his admonition that people may involuntarily sacrifice mental pleasures for mere physical sensations because "they have become incapable of exercising their mental faculties or making choices" as a result of abuse, disease, or neglect.[12]

But all this is only a first step. Within the broad category of mental pleasures, some kinds of mental pleasures are qualitatively superior to others. Our higher mental capacities include abilities to form and remember ideas of objects and activities, to construct propositions and reason to conclusions, to creatively imagine novel and fictitious things, to imaginatively sympathize with other people and even with members of other species, to construct moral rules designed to protect any human or any sentient creature from suffering harms reasonably judged to be wrongful, to direct resentment and punishment against those who intentionally, knowingly, recklessly or negligently break the moral rules, and so forth. So far we have only spoken of what might be termed everyday mental pleasures, that is, a kind of pleasant mental feeling associated with objects or activities that we find useful or merely expedient for our daily life, keeping in mind that these things may relieve us of suffering as well as give us delight. But Mill is clear that the kind of pleasant feeling associated with the moral sentiments, of which the sentiment of justice is the most important and sets the tone for the others such as the sentiment of charity or kindness, is qualitatively superior to any competing kinds of mental pleasures. This kind of pleasure grows up around the idea of justice understood as an essential code of rules, the "very groundwork" of our existence, which distributes and sanctions equal rights and duties to a social group whose members each have a voice in the construction of the code. I shall say more in due course about this kind of pleasure, which Mill calls the pleasure of "security."[13]

Even so, the kind of pleasure associated with aesthetic sentiments of beauty and sublimity may be qualitatively supreme if genuine aesthetic pleasures never conflict with moral pleasures. For Mill, aesthetic pleasure is apparently associated with lofty ideas or ideals of harmony, symmetry, infinity and so forth that direct our attention to an imaginary more perfect world or utopia or heaven that transcends the imperfect world of our experience.(For further discussion relating to the qualitative supremacy of the kind of pleasant feeling associated with genuine aesthetic emotion, see

[11]Of course, different mental pleasures, in which the physical sensation of pleasures is fused with ideas of different objects or activities, may conflict so that an agent must choose one rather than another. But that is not the issue here.

[12]Util II.5–9, pp. 211–214.

[13]Mill, Util V.25, p. 251. For Mill's analysis of the ingredients of the moral sentiment of justice and the pleasant feeling of security which in his view is inseparably associated with it, see Util V.14–25, pp. 246–251. See also Jonathan Riley, "Happiness and the Moral Sentiment of Justice," in Leonard Kahn, ed., Mill on Justice (London: Palgrave Macmillan, 2012), pp. 158–183.

Jonathan Riley, "Optimal Moral Rules and Supererogatory Acts," in B. Eggleston, D. Miller, and D. Weinstein, eds., John Stuart Mill and the Art of Life: A Critical Reader (Oxford: Oxford University Press, 2010), pp. 185–229; and Riley, "Mill's Greek Ideal of Individuality," in K.N. Demetriou and A. Loizides, eds., John Stuart Mill: A British Socrates (London: Palgrave Macmillan, 2013), pp. 97–125.)

The implicit ranking of the different kinds of pleasures in terms of increasing quality, from purely physical sensations through intellectual and moral feelings up to aesthetic emotions, deserves further consideration. There seems to be widespread agreement among suitably competent people that demands of justice and right are overriding, for instance, even if many refuse to account for this in terms of a higher moral kind of pleasure including relief from suffering. And perhaps it is widely agreed that genuine spiritual emotions of beauty and sublimity cannot arise in association with perceived immorality and are of supreme value for a flourishing life, even if many refuse to account for this value in terms of an aesthetic kind of pleasure. (Some may agree with an anonymous referee who complains that there is no reason to see aesthetic pleasure as superior in quality to moral pleasure, even if genuine aesthetic emotion cannot come into conflict with the moral sentiment of justice. Consistently with this, for example, an ideal society of virtuous people who all habitually respect each other's equal rights may be seen as an aesthetic and moral target that has not yet been, and perhaps never will be, achieved by any observed society. Beauty and justice may be in harmony in this case but there is no reason to believe that the aesthetic pleasure of imagining such an ideal society, or of living in one if ever achieved, is qualitatively superior to the pleasant feeling of justice itself. Even so, Mill allows for the possibility of sublime supererogatory actions in which an individual sacrifices her own moral rights (perhaps even her life) to help others: the individual has an aesthetic personal ideal that demands more sacrifice from her than is demanded by her recognized moral duties to others even in an ideal society. Moreover, he allows that the individual may freely pursue artistic or religious commitments in her self-regarding sphere which is properly beyond morality: a supreme kind of aesthetic pleasure might be associated with those non-moral personal projects. In any case, Mill does not insist that experienced judges must find aesthetic pleasures to be superior in quality to moral pleasures, only that they may do so. As indicated later in the main body of this chapter, well-developed human beings are free to disagree over this so long as they fulfil their recognized moral duties: in Mill's unusual doctrine, the utilitarian aggregation procedure is used solely to construct a social code that distributes and sanctions moral rights and duties. Thus, aggregation is restricted to the moral kind of pleasures whereas the individual is given liberty to pursue or not pursue other kinds of pleasures (including aesthetic pleasures) in accord with her own judgment and wishes.)

In any case, a ranking of the different kinds of pleasant feelings in terms of their inherent qualities is a very special sort of preference ordering, namely, a lexicographical (or lexical) ordering. The lexical ordering is a very special ranking because it captures the discontinuities of intrinsic value produced by the infinite superiority of higher pleasures over lower ones: no finite lower pleasure, however large, can ever be equal in value as pleasure to even a bit of higher pleasure.

Apart from his argument that some kinds of pleasant feelings have a superior inherent quality compared to other kinds irrespective of quantity, Mill also seems to dispense with the assumption, commonly thought indispensable to utilitarian ethics, that rich cardinal and interpersonally comparable utility information is available. He seems to rely entirely on individual preference rankings of the feasible sources to obtain competent yet fallible estimates of the amounts and kinds of pleasant feelings to be expected from the sources. Given his reliance on such purely ordinal information about pleasures, he cannot suppose that an objective sum total of any kind of pleasure can be determined in the standard utilitarian manner.

3 Democratic Voting

Instead of a standard utilitarian procedure that adds up pleasant feelings of the same kind, Mill apparently has in mind a democratic voting process in which individual preferences defined over the sources of a given kind of gratification are given equal positive scores or votes and the votes are added up to select an outcome that has the greatest sum total of votes. Indeed, a traditional utilitarian calculus logically reduces to such a democratic process in the context of purely ordinal utility information.[14] Thus, Mill seems to favor what may be called a "purely ordinalist utilitarian" procedure, in other words, a democratic procedure that does not rely on either cardinal utility or interpersonal utility comparisons to generate collective preferences or judgments. (It is worth remarking that Mill does not rely on sympathy as a device for making interpersonal comparisons of utility. Rather, like his father James Mill in A Fragment on MacKintosh and Adam Smith in The Theory of Moral Sentiments, he takes for granted that an agent who imagines himself in another's shoes can only infer what his own feelings of pleasure and pain would be, and cannot be certain of the other's actual feelings either as to intensity or kind. The agent sympathizes with the other only if the other acts, speaks and gestures in ways that persuade the agent to accept that the other's actual feelings may mirror his own hypothetical feelings while imagining himself in the other's position. Even if he does sympathize with the other, the agent does not compare different persons' actual feelings but instead merely adds his own hypothetical feelings at the other's position to his own actual feelings at his own position, when ranking the feasible outcomes. More importantly, even if different individuals do sympathize with others in this way, there is no reason to expect everyone to agree on the ranking of the outcomes. Preferences will remain diverse unless everyone sympathizes with each other to just the same extent while in each others' positions, in which case everyone identifies with one another so completely that society may be treated as if it were composed of a single individual. In that case, the members of society discover by

[14] See Jonathan Riley, "Utilitarian Ethics and Democratic Government," Ethics 100 (1990): 335–348.

putting themselves in each other's shoes that they are already in effect the same person. They are not having to make any comparisons of different persons' actual feelings or to balance conflicting personal preferences so as to calculate an outcome that maximizes a factual sum total of utility. See also Robert Sugden, "Beyond Sympathy and Empathy: Adam Smith's Concept of Fellow Feeling," Economics & Philosophy 18 (2002): 63–87.)

Following Condorcet, it might be argued that a democratic voting process provides a maximum likelihood estimate of a traditional utilitarian outcome under certain conditions. If individuals form preference orderings, are more likely to be correct than mistaken in their estimates of pleasure to be expected from the sources, and cast their votes independently, then an outcome that receives the greatest sum total of votes is very likely to be an optimal outcome at which the sum total of pleasant feelings of the relevant kind is in fact maximized. Even if these quite stringent conditions are satisfied, however, a maximum likelihood estimate of a best option does not tell us the actual amount of the greatest sum total of happiness. Nor does it imply that anybody has any idea of how to add up in a meaningful way the actual feelings of pleasure experienced by different individuals.

Mill makes clear in Considerations on Representative Government that by "true democracy" he means a political system in which the popular majority has ultimate control over political decisions. Consistently with this, the popular majority may employ constitutional measures to help them to arrive at prudent and fair legislative judgments. Such measures include elected representatives as well as various checks and balances designed to encourage the representatives to engage in discussion and deliberation and discourage them from abusing constituted individual rights.[15] The key point is that Mill's idea of a utilitarian aggregation procedure is not a traditional utilitarian calculus but rather a democratic process that relies on purely ordinalist utility information, to wit, a majority decision process that in principle incorporates suitable discussion and debate as well as respect for basic rights. This is consistent with his remarks in *Utilitarianism*. Just as majorities with competent experience of the different kinds of pleasant feelings ought to determine the ranking of the kinds in terms of their intrinsic quality regardless of quantity, he says, so majorities competently acquainted with a given kind of pleasant feeling must determine the relative quantities of that particular kind of utility which can reasonably be expected from its feasible sources.

When he speaks of a sum total of utility, Mill is apparently referring to the estimated sum total which can be inferred from the competent yet fallible majority's ranking of the feasible sources of the particular kind of pleasure. Each competent person's estimate of the quantities of pleasure which he expects from the possible options, expressed merely in terms of more or less pleasure as revealed by his own

[15]For further discussion of the form of representative democracy which Mill recommends as best for any civil society, see Jonathan Riley, "Mill's Neo-Athenian Model of Liberal Democracy", in N. Urbinati and A. Zakaras, eds., J.S. Mill's Political Thought: A Bicentennial Reassessment (Cambridge and New York: Cambridge University Press, 2007), pp. 221–249.

ranking of the options, is counted for exactly as much as another's in a majority voting procedure, to produce an overall estimate of the total quantities of pleasure, as revealed by the majority's ranking of the outcomes. Total pleasure is maximized at an outcome with the most votes. If most people rank x above y, for instance, then x is reasonably expected to yield a greater sum total of the relevant kind of satisfaction than y is, because any person's ranking of the two outcomes is the only (purely ordinal) measure of the amounts of pleasure he expects from those outcomes, and each person's ranking must be counted equally. (If every competent person prefers x to y, then it is obvious that x is reasonably expected to yield a greater total amount of the relevant kind of pleasure than y is, because there is no conflict about which outcome yields more pleasure. Even if individual rankings conflict, however, the majority's ranking of x above y indicates that greater total pleasure may reasonably be expected from x than from y, given the assumption of purely ordinal information about pleasures: more individuals expect greater pleasure from x as opposed to y, fewer individuals expect greater pleasure from y as opposed to x, the individual rankings are the sole measures of the amounts of pleasure expected from the outcomes, and each person's ranking must be given equal positive weight. It does not matter in this context precisely how much actual pleasure will arise from x as compared to y according to any person's estimates.) True, majority rankings may be incoherent. But majority preference cycles, in which a majority prefers x to y, y to z, and z to x, cannot arise if we take an adequate view of what is meant by counting individual preferences equally, to wit, equal positive weights must be applied to each person's preferences. For example, each person might be given one ballot to cast for his top-ranked option and the votes are then added up to select an outcome with the most votes. Strictly speaking, majority rule per se, without the ballot, does not give equal positive weight to different persons' rankings. Rather, it impartially counts the individual rankings over each pair of possible outcomes without attempting to weigh the rankings relative to one another, and simply reflects the shared binary rankings of the greater number. It makes no attempt to compare the different persons' rankings of the amounts of pleasure anticipated from the feasible sources, and is constrained to recognize that each person's ranking does not specify the gain or loss of pleasure between any two sources but instead refers only to "more or less" pleasure.

To remove any possibility of preference cycles, a scoring function, or positional rule, can be superimposed to mimic the way in which votes are distributed to individuals, so that different individuals' preferences are not only treated impartially but also counted equally in the stronger sense that each gets equal positive weight for determining the collective ranking of the outcomes. Although the ballot typically covers only the voter's top-ranked option instead of his entire ranking, the scoring function can be extended to cover his entire ordering. A scoring function such as Borda count might be employed, for instance. Under Borda count, each person ranks the m possible outcomes from best to worst, and $m-1, m-2, \ldots, 1, 0$ points are assigned to the best, next-best, ... second-worst, worst outcomes, respectively. The procedure then selects an outcome with the greatest aggregate point total. (As is well known, Borda rule escapes from Arrow's famous "impossibility theorem" by

replacing Arrow's binary "independence of irrelevant alternatives" condition with a less restrictive independence condition that permits the social choice over any pair of options x and y to be influenced by individual preferences defined over "irrelevant" options such as w and z. See Kenneth J. Arrow, Social Choice and Individual Values, 2nd ed. (New Haven: Yale University Press, 1963); and Amartya K. Sen, Collective Choice and Social Welfare (San Francisco: Holden-Day, 1970).)

More generally, to escape from the Borda method's insistence on equal spacing of points awarded for first choice, second choice, and so forth, a generalized scoring function could be employed whose points assignments are unique only up to a positive monotonic transformation so long as such transformations are applied to every person's ranking. Although it may look as if it implies that interpersonal comparisons of utility are being carried out, this generalized scoring function is just a way to consistently count one person's ranking (and thus his estimates of pleasure to be expected from the outcomes) for exactly as much as another's. There is no attempt to justify the function in terms of interpersonal comparisons of actual pleasures or satisfactions. Rather, the scoring rule is merely an artificial yet impersonal device for implementing the classical utilitarian norm that prescribes giving equal positive weight to equal amounts of pleasure of the same kind, where only purely ordinal information about pleasure is available, and it is supposed that any two person's pleasures are equal in degree if the pleasures are expected to come from outcomes that occupy the same relative positions in the two persons' rankings of the outcomes. On this view, the utilitarian norm itself is the expression of a moral attitude that individuals ought to be treated in this strongly impartial fashion to arrive at collective decisions, independently of any claim to be able to cardinally measure or compare the individuals' actual pleasant feelings. (This view of the principle of utility seems to introduce an element of the type of metaethical theory that Simon Blackburn calls "projectivism" and Allan Gibbard calls "expressivism." See, for example, Blackburn, Essays in Quasi-Realism (New York: Oxford University Press, 1993); Blackburn, Ruling Passions (Oxford: Clarendon Press, 1998); Gibbard, Wise Choices, Apt Feelings (Cambridge, MA: Harvard University Press, 1990); and Gibbard, How Should We Live? (Cambridge, MA: Harvard University Press, 2003). Despite its possible attractions as a way of achieving the utilitarian aim of impartially counting everyone for one and only one, however, the scoring approach does occasionally fail to choose Condorcet winners when they exist; that is, it may fail to select an outcome that is majority-preferred to every other outcome in a series of binary contests. In any case, I shall henceforth assume that some such scoring procedure is used to remove any inconsistencies otherwise associated with majority voting. For recent critical surveys of the vast literature relating to scoring rules including Borda count, see Steven J. Brams and Peter C. Fishburn, "Voting Procedures," in K.J. Arrow, A.K. Sen and K. Suzumura, eds., Handbook of Social Choice and Welfare, vol. 1 (Amsterdam: Elsevier, 2002), 173–206; and Prasanta K. Pattanaik, "Positional Rules of Collective Decision-Making," in ibid., 361–394. Needless to say, Bentham and Mill never discuss such a scoring procedure but it is arguably in the spirit of utilitarianism as they conceive it.)

As I read him, then, Mill supposes that the only available utility information is contained in suitably competent yet fallible individuals' estimates of the various kinds and amounts of pleasant feelings to be expected with respect to any given domain of possible outcomes. The estimates are embodied in individual preference orderings defined over the relevant domain. There are plural kinds of preferences, keeping in mind that the different kinds of pleasant feelings have different sources which may be treated as different aspects of the feasible outcomes. A suitably competent person has a set of $k + 1$ separate preference rankings, as follows: k rankings defined respectively over the k different aspects of the outcomes, one ranking for each of the k kinds of pleasant feelings, each ranking reflecting his estimates of the quantities of the relevant kind of pleasant feeling which he expects from the aspect that is its source, $k > 1$; and a lexical ranking of the k preference rankings which reflects his judgment of the different qualities of the k kinds of pleasant feelings irrespective of quantity. An individual will not form a complete set of these $k + 1$ preferences, however, unless he has developed the intellectual, moral and aesthetic capacities required to competently experience the higher pleasures. I shall now claim that, for Mill, the democratic process, or purely ordinalist utilitarian aggregation procedure, is properly restricted to the construction of an optimal social code of justice that is expected to maximize the sum total of the higher moral kind pleasant feeling (which Mill calls a feeling of security) enjoyed by the members of society.

4 The Restriction of Democracy to Justice and Right

In principle, $k + 1$ distinct democratic processes could be employed to generate $k + 1$ distinct collective preference rankings from the corresponding sets of potentially diverse individual preference rankings, $k + 1$ separate preferences for each individual. As it turns out, however, a democratic voting process is only needed to generate a collective preference ranking with respect to feasible sources of the higher kind of pleasant feeling that is inseparably associated with the moral sentiment of justice, given that majorities competently acquainted with the different kinds of pleasure confirm that this moral kind of pleasant feeling is qualitatively superior to any conflicting kinds of enjoyments irrespective of quantity. Social codes that distribute equal rights and duties as well as sanctions to discourage rule-breakers are the sole source of this moral kind of pleasure. Thus, the democratic process properly considers only individual preference rankings defined over alternative feasible codes, each of which distributes equal rights and duties of a distinctive content to the members of society. (To avoid unnecessary complications, I shall focus exclusively on the moral sentiment of justice, which Mill says sets the tone of all of morality. I am ignoring other moral sentiments such as the sentiments of kindness and beneficence. Whereas moral rules of justice distribute and sanction individual rights and correlative "perfect" obligations (which must invariably be satisfied if demanded by the right-holder or his agent), moral rules of charity and

beneficence distribute and sanction "imperfect" obligations that are not correlative with individual rights. The duty-holder has discretion with respect to fulfillment of his imperfect obligations: he may not wish to give aid to some people because he disapproves of their character, for instance, or because he is no position to give help to anyone—he has no surplus wealth or time to give to others for projects which are not considered matters of justice. Consistently with this, a person may have perfect obligations which are correlative with others' rights to basic subsistence or to aid in situations where they face grievous harms that do not necessarily originate with the duty-holder but instead may arise from third parties or natural causes. Moral rules of charity can be said to provide security for everyone's vital interest in receiving a customary level of help above and beyond the minimum required by justice in the given community, with the caveats that no person has a right to charity and that potential donors are morally permitted to assess whether a person is likely to reciprocate (perhaps to third parties) when/if in a position to do so and/or otherwise satisfy his obligations under the recognized moral rules of the community. In short, moral rules of charity and of justice are both associated with the same higher kind of pleasure, namely, the feeling of security, even though the rules of charity do not assign rights. But, as Mill confirms, the security provided by the rules of charity is "far less in degree" than that provided by rules of justice (Util.V.33, p. 255). See, also, Util.V.37, p. 259, for an illustration that duties of positive beneficence can rise to the level of perfect duties of justice.)

Any competent yet fallible individual ranks the possible codes in accord with his rough estimates of the amounts of security to be expected from their particular rights and duties and sanctions. The democratic process then generates a collective ranking of the codes such that a top-ranked code is recognized and enforced by society as an authoritative code. According to the popular majority or its representatives after discussion and deliberation, a maximum amount of security for the shared vital interests of individuals is reasonably to be expected from the particular rights and duties distributed and sanctioned by this optimal code. Under suitable conditions, this collective estimate may be treated as a maximum likelihood estimate of the true amount of security to be experienced in total by the collection of individuals in society.

The construction of an authoritative social code may be a gradual and piecemeal dynamic process, with particular rules or subsets of rules rather than the whole code coming up for democratic discussion and debate from time to time. Moreover, as individuals acquire more information about the amounts of security to be expected from feasible codes or as circumstances change such that novel rights, duties and sanctions may be required, the given set of individual preference rankings may evolve. As a result, the collective ranking of codes, and thus the authoritative code for society, may change over time, with the important caveat that efforts will be made to maintain consistency among the rules making up the authoritative code at any given time. Indeed, this dynamic process may continue indefinitely as people continue to disagree over which of the many feasible codes is reasonably expected to bring the greatest total security for all. A fixed and final optimal social code of justice that is accepted as best by everyone, and not merely by a majority, is properly

seen as an ideal target, to be approached over time, perhaps, but never to be actually achieved.

It deserves emphasis that, for Mill, an individual ought to be free from coercive interference to pursue in his own way all other kinds of pleasant feelings besides security, as long as he fulfils his duties to others distributed and sanctioned by the authoritative social code. Majorities competently acquainted with the different kinds of pleasures recognize that there is no call for a voting procedure to arrive at enforceable collective judgments with respect to the other kinds of pleasant feelings. Instead, the individual is given capacious freedom to live his own life as he sees fit, in accord with his recognized moral rights and duties. This is an important feature of Mill's utilitarian "art of life," to which I shall return later in the discussion.

Admittedly, it is fair to object that a democratic process is also needed to select social rules for merely expedient purposes independent of morality and justice. Many social rules are established as elements of effective social policy and ought to be followed for that reason, for example, parking regulations that force people to pay a fee if they wish to park their vehicles downtown, customs regulations that prohibit the importation of goods at artificially low prices, health regulations that force people into quarantine if they catch a dangerous infectious disease, and, more generally, what H.L.A. Hart calls "secondary" rules that confer powers on public officials and private persons.[16] By themselves, such expedient social rules do not impose duties—individuals do not have duties to park downtown, or to import goods, or to catch an infectious disease, or to exercise powers such as the power to negotiate contracts—and thus are not part of morality which, as Mill understands it, is marked by the deservingness of punishment for failure to fulfil one's duties to others.[17] Such rules by themselves are not sources of the pleasant feeling of security but rather of some lower-quality merely expedient kind of mental pleasure.

Nevertheless, these merely expedient social rules do specify conditions under which people shall incur duties to others, and once duties are incurred, failure to fulfil them promises to meet with penal sanctions distributed by distinct penal rules. If the merely expedient rules are seen in combination with the penal rules, the combinations properly constitute moral rules. An individual does have a moral duty to pay the parking fee if he parks his car downtown, and a duty not to import goods at dumped prices, and a duty to enter quarantine once he catches a dangerous infection, and a duty to keep to the terms of his contracts as well as a duty to use his discretion to exercise the powers of his office once he has been elected or otherwise appointed to the office. The individual deserves some form of punishment if he fails to satisfy these duties. As these examples attest, a merely expedient rule combined with a sanctioning rule may be said to extend morality in generally expedient directions. Thus, such combinations may be included among the rules comprising a moral code of equal rights and duties, even though it is also true that a utilitarian social code

[16]H.L.A. Hart, The Concept of Law, 2nd ed. (Oxford: Oxford University Press, 1994), pp. 26–49, 80–81.

[17]Mill, Util V 14–15, pp. 246–248.

(including a legal code) does contain rules that, considered in isolation, are not part of morality and justice.

It is also worth emphasizing that, when competent majorities do consider enacting merely expedient social rules, the merely expedient kind of mental pleasure which is reasonably expected from such rules is lower in quality than the higher kind of pleasant feeling associated with the moral sentiment of justice. This is important because it implies that majorities cannot legitimately extend morality by establishing merely expedient social rules that trample over or otherwise impinge upon the individual moral rights and correlative duties which are distributed and sanctioned by rules of justice. Those basic rights and duties of justice protect the vital personal interests shared by individuals and are the only source of the higher pleasure of security. In contrast, merely expedient social laws and policies do not distribute such basic moral claims and duties. Rather, the merely expedient rules by themselves distribute non-moral legal permissions and prohibitions, for example, a permission or privilege to import goods upon payment of a tariff but a prohibition to import the same goods at dumped prices, or a permission to enter into contracts, or a prohibition to move freely about the country when infected with a serious illness. In effect, rules of this sort either provide opportunities which the majority decides are generally expedient but which the individual is not morally obligated to pursue, such as an opportunity to negotiate contracts, or they deny opportunities which the majority decides are inexpedient and which the individual has no moral right to pursue, such as an opportunity to import goods at dumped prices or to go around infecting others with a serious disease.

Once the merely expedient rules are made binding by being combined with suitable sanctioning rules, however, the combinations do give rise to moral duties to obey the law that are enforced by legal penalties. Such moral duties are apparently correlative with the rights of certain public officials, designated as society's representatives, to enforce the merely expedient law on behalf of the public. In other words, the duties are not correlative with moral rights that protect the vital personal interests of all individuals but rather with rights that protect the official interests of certain persons designated to promote the public welfare as determined by the competent majority. Thus, a person who does choose to negotiate a contract has a duty to fulfill the terms of the contract, for instance, and a person who does catch a serious infectious disease has a duty to enter quarantine. Those duties are enforced by officials of the state, even though it is true that no person has a right that others make contracts with him just as no individual has a right that infected people enter quarantine. Anyone who breaks his contracts without an excuse or who escapes from quarantine when ill with a serious infection deserves punishment. But I cannot further discuss these matters.

Such complications do not frustrate my claim that, in Mill's approach, a democratic voting procedure is employed only to generate collective rankings of the distinct codes of equal rights and duties which are the feasible sources of the pleasant moral feeling of security. If individuals can form only partial orderings of the distinct codes for one reason or another, then democratic social choice can yield a maximal code rather than an optimal security-maximizing one, keeping in mind

that a maximal code cannot be said to yield more security for everyone's shared vital personal concerns than that yielded by other possible maximal codes. Nevertheless, it may be assumed for convenience that competent individuals do form complete (though possibly mistaken) preference orderings, and that a suitable democratic voting procedure such as Borda rule can be employed to generate complete social orderings. It may well be the case that the orderings exhibit extensive ranges of indifference in some social contexts, where even competent people do not have very highly developed capacities and lack the information needed to make discriminating comparative judgments of the amounts of security to be expected from distinct codes. Such indifference does not prevent the democratic selection of an optimal code under the circumstances, although multiple codes may appear to be tied as optimal, any one of which may be picked. Moreover, as they develop their capacities further and acquire additional information about the security effects of distinct codes of equal rights and duties, competent people may alter their judgments, cease to be indifferent among so many distinct codes, and perhaps even converge on a particular code as best.

5 Justice and Right as Maximization of Security

In Mill's utilitarianism, justice is properly understood as the art of maximizing the higher moral kind of pleasant feeling which he labels as security. This higher kind of pleasure (or utility or interest) can only be experienced under an authoritative social code that distributes and sanctions equal rights and duties for a group of peers. The rights and duties which are distributed and enforced by such a code are the source of the individual's moral feeling of security:

> When we call anything a person's right, we mean that he has a valid claim on society to *protect* him in the possession of it, either by the force of law, or by that of education and opinion. If he has what we consider a sufficient claim, on whatever account, to have something guaranteed to him by society, we say that he has a right to it.[18]

Security is a variable, however, which can only be maximized under an optimal social code that distributes and sanctions particular equal rights and duties for all members of society each of whom has a voice in the construction of the code. In short, a democratic process is required. No code can be optimal if it is imposed against the considered opinions of the popular majority or its elected representatives. Security cannot be maximized in an aristocracy or oligarchy, for example, where rights are distributed and sanctioned exclusively for some minority group of peers. Nor can it be maximized in an absolute monarchy or dictatorship, where rights and privileges are exclusively the leader's and everyone else has only duties to respect the leader's claims.

[18] Util V.24, p. 250, emphasis added.

After making clear that everyone feels the immense importance of enjoying security from grievous injuries because "on it we depend for all our immunity from evil, and for the whole value of all and every good, beyond the passing moment," Mill goes on to say that this "extraordinarily important and impressive kind of utility . . . cannot be had, unless the machinery for providing it is kept unintermittedly in active play."[19] For convenience, let us assume that an optimal social code of justice is wholly a legal code enacted and enforced by the official "machinery" of the state. This is not an essential assumption—some moral rules, including rules of justice, cannot be expediently enforced as legal rules. But it allows us to simplify things in order to concentrate on core aspects of security-maximization that remain central even when the assumption that moral rules are always established as laws is relaxed. For instance, it allows us to conceive of the democratic procedure as a formal political process which aims to construct a legal code of justice. Social sanctions can also be identified with legal sanctions.

Given the simplifying assumption, we can say that the "machinery for providing [security]" is essentially the official political machinery which is "kept unintermittedly in active play" to enact, interpret and enforce the law. A legal code of justice must be constructed, modified and enforced over time through the cooperative efforts of various public officials, the most senior of whom are elected by the citizenry. Moreover, officials and private individuals alike must be taught the importance of the rule of law and constantly reminded of their legal rights and duties. This continuous active play of the machinery for providing the pleasant feeling of security works to heighten the intensity of the feeling for anyone who accepts the legal and moral code, so much so that the feeling comes to be felt as a distinctive kind of utility, qualitatively superior to competing kinds:

> Our notion, therefore, of the claim we have on our fellow creatures to join in making safe for us the very groundwork of our existence, gathers feelings round it so much more intense than those concerned in any of the more common cases of utility, that the difference of degree (as is often the case in psychology) becomes a real difference in kind. The claim assumes that *character of absoluteness, that apparent infinity, and incommensurability with all other considerations,* which constitute the distinction between the feeling of right and wrong and that of ordinary expediency and inexpediency.[20]

Given that the security afforded by a legal and moral code of recognized equal rights and duties is infinitely more valuable than any competing enjoyments, this social code of justice takes absolute priority over any competing considerations of expediency. An individual's equal rights can never be legitimately overridden without his consent to promote other people's happiness because even a bit of the higher pleasure of security—no matter who feels it—is intrinsically more valuable as pleasure than any amount of lower pleasure, however large, which human nature

[19]Util V.25, p. 251.

[20]Util V.25, p. 251, emphasis added.

is capable of experiencing—no matter how many different persons are assumed to experience it.[21]

The qualitative superiority of the pleasure of justice over competing kinds of pleasures implies that any person who is competently acquainted with the different kinds gives absolute priority to considerations of justice over competing considerations in order to maximize his own happiness both in point of quality and quantity. Any such individual will voluntarily give such priority to justice in his interactions with like individuals, because they can be trusted to reciprocate. Moreover, his maximization of his own happiness is logically compatible with his fellows' maximization of their personal happiness, and thus with maximization of the general happiness regarded as nothing but the simultaneous maximization of everyone's personal happiness. His own moral feeling of security is maximized if and only if everyone else's is also maximized because the moral feeling implies that equal rights must be distributed to all. Codes that distribute and sanction equal rights and duties are the sole source of the higher pleasure of security associated with the moral sentiment of justice. No competing kinds of enjoyments can ever be equal in value as pleasure to this enjoyable feeling of security according to most people competently acquainted with the different kinds. Thus, a competent individual's own happiness necessarily coheres with like individuals' personal happiness to the extent of their equal rights. Neither personal utility nor general utility can be promoted by violating rights.

To even form a moral sentiment of justice, an individual must be able to identify the particular social rules of justice with which to comply. Until he knows the particular code which ought to be accepted, he cannot know the particular equal rights and duties which ought to be recognized by everyone within his community as belonging to him and anyone else in like circumstances. But to establish rights and duties that are publicly endorsed by his society in its laws and conventions, the individual must participate with his fellows in a political process, given that an ideal observer is not available to determine the best moral and legal code. An open democratic process of free discussion and debate is essential for fallible beings to assess proposals and converge on an optimal social code, that is, a code that impartially distributes those particular equal rights and duties which, at least so far as competent yet fallible people can tell, maximize the feeling of security enjoyed by anyone and everyone who possesses them. In short, the individual's sentiment of justice presupposes a voting procedure to select an optimal code upon which any fair-minded individual must rely to guide his interactions with his fellows.

Perhaps we can begin to appreciate how Mill conceives of a utilitarian doctrine that ultimately aims to maximizes happiness both in point of quantity and quality.

[21]Given the simplifying assumption, any suitably competent individual who accepts legal rules as reasons for action also by definition accepts them as moral reasons. This necessary connection between law and morality ceases to exist, however, when the simplifying assumption is relaxed, as it must be. So I do not mean to reject (or endorse) Hart's influential version of inclusive legal positivism. See H.L.A. Hart, The Concept of Law; and Hart, Essays on Bentham. (Oxford: Clarendon Press, 1982).

Because the qualitative superiority of the moral kind of pleasure over competing kinds of gratifications means that a social code selected by competent yet fallible majorities takes absolute priority over conflicting considerations, there is no fundamental conflict between personal happiness and the happiness of all. Instead, the ultimate aim is a comprehensive social outcome in which each and every individual who is competently acquainted with the different kinds of pleasures simultaneously maximizes his personal happiness in point of quantity and quality. As Mill puts it, "the ultimate end, with reference to and for the sake of which all other things are desirable (whether we are considering *our own good or that of other people*), is an existence exempt as far as possible from pain, and as rich as possible in enjoyments, both in point of quantity and quality."[22]

6 The Millian Utilitarian Social Welfare Functional[23]

Given our assumptions, Mill's extraordinary utilitarianism can be represented as a doctrine in which a purely ordinalist utilitarian—that is, democratic—social welfare functional (SWFL) such as Borda rule is restricted in its operation to a higher moral kind of preferences ultimately motivated by the higher pleasure of security that is inseparably associated with the moral sentiment of justice, with justice conceived in terms of a moral and legal code that distributes and sanctions equal rights and duties (as well as privileges, powers, immunities and their correlative positions) for all. This Millian SWFL takes as input any given set of individual preference orderings defined over alternative feasible codes (or portions of codes) of equal rights and duties, and generates as output a social preference ordering of the feasible codes such that a top-ranked code is an optimal code, in other words, a code that may reasonably be expected to maximize security for everyone's vital personal interests, and one that most participants in the democratic aggregation process agree ought to be chosen and recognized by their society as authoritative. It must be re-emphasized that the democratic aggregation process can only be run without auxiliary internal checks and balances on the assumption that participants are competently acquainted with the different kinds of pleasant feelings, in which case they are moral agents who recognize that considerations of justice are far more valuable than competing considerations are for their own personal happiness as well as collective happiness. In non-ideal social contexts, where at least some individuals are narrowly selfish rather than moral agents, a universal democratic franchise must be combined with

[22] Util II.10, p. 214, emphasis added.

[23] Various forms of social welfare functionals and the axioms or conditions that characterize them are brilliantly summarized by Amartya Sen, "Social choice theory," in K. J. Arrow and M. D. Intriligator, eds., Handbook of Mathematical Economics, Vol III (Amsterdam: North-Holland, 1986), pp. 1073–1128.

an expedient scheme of checks and balances to promote deliberation and discourage abuse of power.[24]

Confining attention to the ideal case, an optimal social code of justice distributes and sanctions the particular rights and duties which, in the competent majority's judgment, provide the best protection for what the majority considers the vital personal concerns shared by the members of the community. The majority's judgment, though fallible, is very likely to result in a code that maximizes the sum total of the moral gratification of security actually enjoyed by the equal right-holders, although it is not claimed that we can precisely measure, compare and add up the amounts of this moral feeling experienced in fact by the different individuals. It is possible that different individuals do experience different amounts of the feeling of security from the same rights, but we do not claim to possess the rich utility information needed to decide one way or the other.

As indicated earlier, the Millian SWFL may be run repeatedly over time as individual preferences change in response to new information and unforeseen social situations, so that the process of constructing an optimal code is not only piecemeal and gradual but may also continue indefinitely as disaffected minorities continue to push for reconsideration of which rules and rights are best for promoting security. In effect, a code recognized as optimal at a given point in time may come to be seen as sub-optimal and so be replaced by another code, distinct in some respects from the earlier code, which a new competent majority now chooses as authoritative.

The Millian SWFL may be said to constitute the core of Mill's utilitarianism in so far as the SWFL generates an optimal code of justice and right that secures "the very groundwork of our existence." But there is much more to Mill's utilitarianism than the SWFL. For the rights and liberties distributed and sanctioned by an optimal social code of justice give rise to spheres of life and conduct which are distinct from the sphere of morality and law within which the SWFL's operation is confined. Individuals are free to pursue their non-moral kinds of preferences, which are ultimately motivated by kinds of pleasant feelings other than the pleasant feeling of security, in accord with their recognized rights and duties because society has no need to map these other kinds of preferences into a social or moral preference ranking. Indeed, society ought not to even try to map any non-moral kind of individual preferences into a social preference because that sort of exercise is ruled out by the individual rights and liberties distributed and sanctioned by the rules of justice. Since the rights and liberties of justice are the sole source of the pleasures of security and individuality, which are far more valuable as pleasure than any competing kinds of pleasure, Mill's utilitarianism insists on the importance of spheres of individual freedom beyond the realm of morality and law and the reach of the SWFL.

[24]For discussion of the scheme of checks and balances that Mill recommends in his proposed plan of representative democracy, see Riley, "Mill's Neo-Athenian Model of Liberal Democracy."

For Mill's utilitarianism to work, the conditions imposed on the SWFL must be suitably restricted in scope such that they refer only to the moral kind of preferences and do not apply to the non-moral kinds. In effect, the Millian utilitarian SWFL, which generates a moral kind of social preference from any given set of a moral kind of personal preferences, all of which can be represented by technical utility functions, makes possible an "enlarged" utilitarian "art of life" that is compatible with, though it extends beyond, the utilitarian morality. The enlarged philosophy prescribes freedom for individuals and groups in domains of life and conduct that are beyond the moral realm of the SWFL, including a "purely self-regarding" domain as well as a distinct other-regarding domain of competitive freedom. On this occasion, I cannot further discuss in detail these spheres of freedom beyond morality and law. But there is no presumption that individuals will form preference orderings that can be represented by technical utility functions in these non-moral domains. (For my interpretation of the purely self-regarding sphere in which the individual has a moral right to complete liberty in the sense of doing whatever she pleases, see Jonathan Riley, Mill on Liberty (London: Routledge, 1998; 2nd ed., 2015). The distinct Millian sphere of competitive freedom can accommodate Robert Sugden's convincing account of competitive markets and the opportunities they present for mutual advantage. See, e.g., Robert Sugden, "Opportunity as a Space for Individuality: Its Value, and the Impossibility of Measuring It," Ethics 113 (2003): 783–809; Sugden, "The Opportunity Criterion: Consumer Sovereignty without the Assumption of Coherent Preferences," American Economic Review 94 (2004): 1014–1033; Sugden, "Opportunity as Mutual Advantage," Economics and Philosophy 26 (2010): 47–68; and Ben McQuillin and Robert Sugden, "How the Market Responds to Dynamically Inconsistent Preferences," Social Choice and Welfare 38 (2012): 617–634.)

7 The Millian Utilitarian Art of Life

It emerges that Mill's utilitarian moral theory, whose central core is justice, is only a part, although a crucial part, of a larger utilitarian art of life that, besides aiming to maximize the higher moral pleasure of security within a limited moral domain of life, also aims to maximize non-moral kinds of pleasures in domains of life beyond moral and legal regulation. Mill leaves no doubt that the principle of utility, "considered as the directive rule of human conduct," is also the "first principle of morals" that ultimately determines right and wrong conduct.[25] Given that there are different kinds of utility of different intrinsic qualities, the principle of utility is a multidimensional principle. The highest abstract standard of justice and morality is the component of the principle of utility which ultimately directs

[25]Mill, Util II.9, p. 214, and Util V.36, p. 257.

humans to maximize security, the higher kind of pleasure associated with the moral sentiment of justice, a kind of utility that is intrinsically more valuable as utility than any competing kinds of utility. According to this abstract standard of justice, the suitably competent yet fallible majority ought to construct (no doubt in a gradual and piecemeal fashion) a social code that impartially distributes and sanctions the particular equal rights and duties which, upon reflection, are believed to maximize the pleasant feeling of security for all in the given civil society. Fair-minded individuals can then look to these recognized rights and duties for direction as to how to perform right acts and omissions and avoid wrong ones in their community. In short, morality is the art of maximizing security. As Mill says, "morality ... may accordingly be defined [as] the rules and precepts for human conduct, by the observance of which an existence such as has been described [that is, a happy existence in terms of both quantity and quality of pleasure, including freedom from pain] might be, to the greatest extent possible, *secured* to all mankind; and not to them only, but, so far as the nature of things admits, to the whole sentient creation."[26]

But the multidimensional principle of utility is more than the first principle of morals. It also directs individuals to maximize non-moral kinds of utility in non-moral domains of life. For instance, the rules of justice can distribute and sanction equal rights to complete liberty of purely self-regarding conduct because such conduct does not directly harm others or, if it does, only with their genuine consent and participation, where harm is defined as any form of perceptible damage to external objects of concern to others (including their bodies, material wealth, reputations, contracts, and so forth) but excludes their mere dislike. Since no person suffers any non-consensual harm, every person can do as she likes and avoid what she dislikes without impeding or obstructing others. Given that liberty in the sense of choosing as one pleases is the only permanent and unfailing source of individuality or self-development, and that individuality is "a principal ingredient of happiness," Mill claims that individuals "capable of rational persuasion" should choose as they like in their self-regarding affairs so as to develop their higher faculties of intellect, imagination, and moral sentiment. The individual should freely pursue his personal projects and aesthetic commitments, for example, as long as he does not inflict any form of perceptible damage on others without their consent. This "one very simple principle of liberty" is, for him, a component of the principle of utility.

As well, the rules of justice can distribute and sanction private property rights in accord with the principle that producers deserve the fruits of their own labor and saving, and also distribute liberties or permissions to compete without force or fraud over the marketing of scarce material goods. (Mill also leaves open the possibility that capitalist property rights might eventually be replaced by rights of democratic participation in worker cooperatives in a decentralized system of market socialism. But he emphasizes that egalitarian reforms of the capitalist system must be the concern of progressive thinkers for the foreseeable future. See, e.g., Jonathan Riley,

[26]Util II.10, p. 214, emphasis added.

"J.S. Mill's Liberal Utilitarian Assessment of Capitalism Versus Socialism," Utilitas 8 (1996): 39–71; Riley, "Mill's Political Economy: Ricardian Science and Liberal Utilitarian Art," in John Skorupski, ed., The Cambridge Companion to John Stuart Mill (Cambridge: Cambridge University Press, 1998), pp. 293–337; and Riley, J. S. Mill: Principles of Political Economy and Chapters on Socialism, abridged ed. (Oxford: Oxford University Press, 2008).) Even though successful competitors impose non-consensual harms on the losers in such a sphere of competitive freedom, society properly decides that the harms of wasted exertions, lost income and perhaps even bankruptcy are not wrongful because they are outweighed in the opinion of competent majorities by the social benefits of efficient allocation of resources and economic growth made possible by the free competition, always assuming that the competition is conducted without force or fraud. Since competent majorities do not consider these non-consensual harms suffered by losing competitors to be immoral, the rules of justice do not distribute duties to prevent them or to punish the successful competitors. In other words, competent people agree that a claim to be protected from this kind of suffering is not needed to promote security for our vital concerns: "society admits no right, either legal or moral, in the disappointed competitors, to immunity from this kind of suffering; and feels called on to interfere, only when means of success have been employed which it is contrary to the general interest to permit—namely, fraud or treachery, and force."[27] Instead, such competitive behavior is morally permissible, and ought to be encouraged to the extent that it is reasonably expected to bring collective benefits that outweigh harms of the kind suffered by the losing competitors. This principle of an efficient social policy of laissez-faire within the bounds of justice and right is also, for Mill, an emanation from the multidimensional principle of utility. (It deserves emphasis that the permissions to compete distributed by the rules of justice are not naked liberties such that an individual has no duty not to compete on any terms. Instead, the liberties are backed up and qualified by such just claims as the right not to be physically prevented from entering the competition, the right not to have to compete against fraudulent sellers who market fake cheap goods, and so forth. Competitors do have duties correlative with these just claims.)

In contrast to these spheres of individual freedom, individuals have duties to others in the sphere of morality and law, and can in principle be legitimately compelled to obey the rules of justice established by competent majorities. Anybody who violates another's recognized equal rights, as distributed by an optimal code, causes the kind of harm to others which is considered by the competent majority to be wrongful and deserving of some form of punishment, including legal penalties, public stigma and humiliation, and a guilty conscience. The wrongdoer has failed to fulfil his duty correlative to the other's right, and thus has damaged some vital concern of the right-holder's which the majority believes ought to be secured by right. But it must not be thought that Mill is committed to a negative theory of justice that seeks merely to prevent individuals from inflicting wrongful physical

[27] Mill, OL V.3, p. 293.

and economic injuries on others without their consent. Rather, justice requires respect for equal rights, and these rights may include positive rights of assistance from others, who have correlative duties to provide the assistance. Mill defends an equal right to poor relief or subsistence funded by the general taxpayer, for instance, although he also believes that those receiving support should work for it if able to do so, and that anyone receiving support should not be entrusted with the franchise. In any case, security from starvation or abject poverty is certainly something that a majority of suitably competent people can be expected to recognize as a vital personal interest that ought to be guaranteed by society, at least in societies where a threshold level of collective wealth has been achieved. Moreover, justice demands that individuals have rights to be secure from brute bad luck at least to some feasible extent. Thus, in situations where an individual would suffer grievous damage to her important interests through no fault or choice of her own, as when a comatose diabetic person falls face down in a puddle of water, others have perfect duties to help her if they can easily do so without risk of grievous harm to themselves. They deserve punishment if they fail to do their duty even though the harms to the victim do not arise from their own actions.

Much more needs to be said to clarify Mill's extraordinary version of utilitarianism. But enough indication has been provided, I hope, to encourage further study of this remarkable doctrine. Indeed, my view is that Millian utilitarianism has the resources to deflect the familiar objections which are fatal to standard utilitarianism, including the charges that utilitarianism cannot adequately take account of the importance of distributive justice or of beautiful personal projects and commitments. (For an excellent critical discussion of what's wrong with standard utilitarianism, see Will Kymlicka, Contemporary Political Philosophy: An Introduction, 2nd ed. (Oxford: Oxford University Press, 2002), pp. 10–52. Kymlicka's considered objections are all the more noteworthy because he attempts to provide a generous justification for standard utilitarianism, seeing it as a way of interpreting the abstract ideal of equal concern and respect for persons rather than as a teleological doctrine that aims to achieve a state of affairs in which the sum total of utility is maximized. Other leading critics of standard utilitarianism include John Rawls, A Theory of Justice (Cambridge: Harvard University Press, 1971; rev. ed. 1999); Bernard Williams, "A Critique of Utilitarianism," in J.J.C. Smart and Bernard Williams, Utilitarianism: For and Against (Cambridge: Cambridge University Press, 1973), pp.75–150; Williams, Ethics and the Limits of Philosophy (Cambridge, Ma.: Harvard University Press, 1985), esp. pp. 71–129, 174–202; Samuel Scheffler, The Rejection of Consequentialism, rev.ed. (Oxford: Clarendon Press, 1994); Scheffler, Boundaries and Allegiances: Problems of Justice and Responsibility in Liberal Thought (Oxford: Oxford University Press, 2001), pp. 149–215; and Amartya Sen, The Idea of Justice (Cambridge: Harvard University Press, 2011), pp. 269–290. Rawls largely exempts Mill's utilitarianism as he understands it from his critique of standard utilitarianism. See John Rawls, Lectures on the History of Political Philosophy, ed. Samuel Freeman (Cambridge, MA: Harvard University

Press, 2007), pp. 249–316. For a critical assessment of Rawls's interpretation of Mill's utilitarianism, see Jonathan Riley, "Rawls, Mill and Utilitarianism," in A Companion to Rawls, eds. Jon Mandle and David Reidy (Oxford: Wiley-Blackwell, 2014), pp. 397–412. But further discussion of these matters must be left for another occasion.

Acknowledgements I wish to thank two anonymous referees and the editors of this volume for their comments on an earlier draft. I am also grateful to the editors for the invitation to contribute to this festschrift for my friend and former colleague Nick Baigent. Responsibility for the views expressed remains mine alone.

Lindahl and Equilibrium

Anne van den Nouweland

Abstract This paper demonstrates that there is a discrepancy between the ideas expressed by Lindahl in 1919 and the current-day definition of Lindahl equilibrium. It describes how the ideas expressed by Lindahl developed into the equilibrium concept for public good economies that now carries Lindahl's name. The paper also touches on a seemingly forgotten equilibrium concept for public good economies known as ratio equilibrium, and explains that from an axiomatic perspective this equilibrium concept is a better fit with the ideas expressed by Lindahl.

Keywords Lindahl equilibrium • Ratio equilibrium • Public good economies

1 Introduction

Lindahl equilibrium and ratio equilibrium are different equilibrium concepts for public good economies. Lindahl equilibrium is based on consumers paying personalized prices for public goods, whereas ratio equilibrium is based on consumers paying personalized shares of the costs of public good production. The two concepts coincide when production of public goods exhibits constant returns to scale, but they differ for more general production processes.

Lindahl equilibrium carries Lindahl's name and it is commonly accepted that it was introduced by Lindahl [11], and later formalized by Samuelson [19], Johansen [7], and Foley [5]. However, in van den Nouweland et al. [23] we showed that not Lindahl equilibrium, but ratio equilibrium, which was formalized by Kaneko [9], accurately represents the cost-share ideas expressed in Lindahl [11]. This discrepancy motivates the current paper. I describe how the literature developed from Lindahl [11] to Lindahl equilibrium and I discuss the relation between ratio equilibrium and the ideas in Lindahl [11].

Lindahl [11] does not contain a mathematical definition of an equilibrium concept, but ideas expressed in text and a picture. These ideas are of individual agents determining their demand for public good based on shares of the cost of

A. van den Nouweland (✉)
Department of Economics, University of Oregon, Eugene, OR 97403-1285, USA
e-mail: annev@uoregon.edu

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*,
Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_18

public good production. I show that in the literature in which the ideas in Lindahl [11] were developed into the current-day concept known as Lindahl equilibrium, several things have happened. First, the meaning of the word *price* has evolved from sometimes meaning a total amount to be paid for a certain quantity of a good, into a fixed amount to be paid per unit of a good. Second, at a time when the literature was transitioning from explaining ideas verbally and graphically into using mathematical notation to formalize them, and many economists were unfamiliar with the use of mathematical methods, *price* came to have a very specific meaning and notation that was introduced for the special case of constant returns to scale came to be used in the definition of Lindahl equilibrium. Third, the assumption of constant returns to scale that was initially carefully mentioned, has been dropped over time. This is acutely relevant given that personalized prices for public good are not equivalent to personalized shares of the cost of its production when the production technology for the public good does not exhibit constant returns to scale.

The paper is organized as follows. In Sect. 2, I provide the current-day definition of Lindahl equilibrium and some discussion of this definition and its implications. In Sect. 3, I outline the ideas for an equilibrium concept that were explained in Lindahl [11]. In Sect. 4, I explain how the literature evolved from the ideas expressed in Lindahl [11] to the definition of Lindahl equilibrium as a concept based on personalized prices. In Sect. 5, I describe the axiomatic link between Lindahl [11] and ratio equilibrium. I conclude in Sect. 6.

This paper contains many direct quotes. Rather than explicitly attributing each and every quote to its source, whenever there can be no confusion about the source of a particular quote, I will simply surround it by the two signs “ and ”.[1]

2 Lindahl Equilibrium and Personalized Prices

In this section I demonstrate that Lindahl equilibrium is a concept based on personalized prices and discuss some implications of this aspect of the definition. I first provide a definition of a pure public good economy and a general definition of Lindahl equilibrium. I follow that up with descriptions of Lindahl equilibrium taken from three different sources that corroborate that in the literature Lindahl equilibrium is indeed defined using personalized prices. I end this section with a discussion of the nature of Lindahl equilibrium and examples that illustrate some potential problems with this concept.

[1] In some cases the signs “ and ” also appear in the quotes themselves.

A *pure public good economy* (with one public good and one private good) is a list $E = \langle N; (w_i)_{i \in N}; (u_i)_{i \in N}; c \rangle$ consisting of the following elements. N is a non-empty finite set of consumers. Each consumer $i \in N$ has an initial endowment w_i of the private good, and a utility function $u_i(x, y_i)$ for consumption of amounts x of the public good and y_i of the private good. There is a single producer of the public good and $c(x)$ is the cost in terms of private good for producing an amount x of the public good.

2.1 Definition of Lindahl Equilibrium

Lindahl equilibrium is a concept for pure public good economies that mirrors the definition of competitive equilibrium in private-good economies. In a Lindahl equilibrium, each consumer takes prices of all goods as given and demands levels of goods that maximize her utility among the bundles of goods that she can afford given her endowment and those prices. However, unlike in private-goods economies, each consumer is assumed to face a *personalized price* for units of the public good and this price can be different for each consumer. The personalized prices of all consumers are added to find the price at which a producer of a public good can sell its output and the producer determines a profit-maximizing production level given this jointly determined price. In case the producer has a positive profit, this is distributed among consumers according to some exogenously given distribution rule and a consumer's share of the profits is added to her initial endowment when determining her budget constraint. Finally, the market-clearing condition for public goods requires that all consumers demand the same level of public good (as opposed to the condition for private goods that the sum of demands by all consumers equals available amounts), and that this demand coincides with the producer's profit-maximizing supply.

Different rules for the distribution of profits may lead to different decisions by consumers and thus Lindahl equilibrium is defined with respect to a distribution rule. A distribution rule is a vector $d = (d_i)_{i \in N}$, with $\sum_{i \in N} d_i = 1$, where d_i is the proportion of the profits of public-good production that fall to consumer i and which this consumer can either consume as private good or use to pay for public good. A *d-Lindahl equilibrium* consists of a vector of personalized prices $\mathbf{p}^* = (p_i^*)_{i \in N}$, an amount of public good x^*, and amounts of private good $(y_i^*)_{i \in N}$ such that

1. x^* is a solution to

$$\max_x \left(\sum_{j \in N} p_j^* \right) x - c(x)$$

2. For each $i \in N$, (x^*, y_i^*) is a solution to

$$\max_{(x,y_i)} \; u_i(x, y_i)$$

$$\text{subject to} \quad p_i^* x + y_i \leq w_i + d_i \left(\left(\sum_{j \in N} p_j^* \right) x^* - c(x^*) \right)$$

This definition of Lindahl equilibrium captures the basic elements that the current-day definitions in the literature have in common. Of course, there are variations in descriptions, notations, and context in different papers, as the following three subsections illustrate.

2.2 *Lindahl Equilibrium in Mas-Colell et al. [13]*

Mas-Colell et al. [13], the standard text used in many graduate programs in economics, describes Lindahl equilibrium on pages 363–364. This description involves, for each consumer i, a market for the public good "as experienced by consumer i" and a price p_i of this personalized good. The prices p_i may differ across consumers. Given the equilibrium price p_i^{**}, each consumer i decides on the equilibrium amount x_i^{**} of public good (s)he wants to consume so as to solve

$$\max_{x_i \geq 0} \phi_i(x_i) - p_i^{**} x_i,$$

where $\phi_i(x_i)$ is consumer i's utility from consuming an amount x_i of the public good. A firm produces a bundle of I goods – I being the number of consumers – using a fixed-proportions technology such that the level of production of each personalized good is necessarily the same. The firm's equilibrium level of output q^{**} solves

$$\max_{q \geq 0} \left(\sum_{i=1}^{I} p_i^{**} q \right) - c(q),$$

where $c(q)$ is the cost of supplying an amount q of the public good. In equilibrium, the market-clearing condition $x_i^{**} = q^{**}$ has to hold for all i.

Mas-Colell et al. [13] assume that the cost function $c(\cdot)$ has a strictly positive second derivative at all $q \geq 0$, and thus they exclude cases where public-good production exhibits constant returns to scale.[2] They state that the type of equilibrium

[2] I am highlighting this feature because we will see in Example 2 that personalized prices can be problematic when marginal costs are not constant.

just described "is known as a *Lindahl equilibrium* after Lindahl (1919)" and refer the reader to Milleron [14] for a further discussion.[3]

2.3 Lindahl Equilibrium in Kreps [10]

Kreps [10] is a new text whose targeted audience is "first-year graduate students who are taking the standard" theory sequence "and would like to go more deeply into a selection of foundation issues". This text describes Lindahl equilibrium on pages 381–382, as part of a discussion of externalities in the setting of Coase [2], and states that "Lindahl proposed this equilibrium before Coase (in 1919)". Kreps [10] takes a full page to describe Lindahl equilibrium, which I have condensed to the following.

Each consumer h maximizes the utility $u^h(\mathbf{x}, \mathbf{z})$ that she accrues, subject to the budget constraint

$$p\mathbf{x}^h + \sum_{h' \neq h} r_{hh'}\mathbf{x}^h \leq pe^h + \sum_f s^{fh}\left[pz^f - \sum_{h'} q_{fh'}\mathbf{z}^f\right] + \sum_f q_{fh}\mathbf{z}^f + \sum_{h \neq h'} r_{hh'}\mathbf{x}^{h'},$$

where p denotes prices for the goods, the $r_{hh'}$ denote transfer prices that record transfers from h to h' made for the choice of $\mathbf{x}^h$ by h, e^h denotes endowment, the s^{fh} denote shares in the profits of the firms, and the q_{fh} denote transfers made from firm f to consumer h made for f's choice of $\mathbf{z}^f$. Also, given prices, a firm f chooses a production plan z^f to maximize the transfer-induced profits $pz^f - \sum_h q_{fh}z^f$. Kreps [10] states "A Lindahl equilibrium is a vector $(p, q, r, \mathbf{x}, \mathbf{z})$, where: firm f, taking prices as given, maximizes its net-of-transfer profits at $\mathbf{z}^f$; consumer h, taking prices as given, maximizes her preferences at $(\mathbf{x}, \mathbf{z})$, given the budget constraint above; and markets clear, in the usual fashion. N.B., every consumer chooses the full vector $(\mathbf{x}, \mathbf{z})$, and it is a condition of equilibrium that these choices agree."

2.4 Lindahl Equilibrium in Oakland [18]

Oakland [18] is the chapter on the theory of public goods in *The Handbook of Public Economics*. He refers the reader to Johansen [7][4] for the source of the "so-called Lindahl approach", which he describes (verbally) on page 525 as requiring "that individuals be charged (taxed) an amount equal to their marginal valuation times the level of public good supplied". Oakland explains that this approach involves the

[3] I will discuss Milleron [14] in Sect. 4.

[4] I will discuss Johansen [7] in Sect. 4.

Marshallian demand curve of each individual for the public good, which "shows the amount of public good desired at each constant tax price". These demand curves are summed vertically to find the aggregate marginal valuation for each particular level of the public good and then the resultant curve is intersected with the marginal cost schedule for the public good to find the "efficient level of public good". It is pointed out that "each individual's tax price will vary - the highest price charged to the individual with the greatest marginal valuation".

In a footnote, it is pointed out that, strictly speaking, the approach described "is correct only if the public good is produced under conditions of constant costs", because otherwise "other taxes and subsidies will be required which will affect the underlying demand curves".

2.5 Discussion

The preceding three subsections clearly demonstrate that Lindahl equilibrium is a concept based on prices. Mas-Colell et al. [13], Kreps [10], and Oakland [18] all, each in their own way, describe consumers who maximize their utilities taking into account a budget constraint in which quantities of public good are multiplied by some per-unit price. They also all include, explicitly or implicitly, a profit-maximization condition, and Kreps [10] explicitly includes terms related to the distribution of profits among the consumers, whereas Oakland [18] hints at the need for a way to distribute profits if production does not exhibit constant returns to scale. In the remainder of this discussion, I will work with the definition in Sect. 2.1, which makes all these aspects of Lindahl equilibrium explicit with a minimum of notation.

I discuss some of the issues surrounding Lindahl equilibrium by means of two simple examples, one of a public good economy with constant returns to scale in public good production, and one with decreasing returns to scale.

Example 1 Lindahl equilibrium in an economy with constant returns to scale.

Consider the 2-consumer public good economy with $c(x) = 2x$, $N = \{1, 2\}$, $w_1 = w_2 = 8$, $u_1(x, y_1) = x^{1/4} y_1^{3/4}$, and $u_2(x, y_2) = x^{3/4} y_2^{1/4}$. Obviously, the profit-maximization problem of the producer only has a non-zero solution if $p_1^* + p_2^* = 2$ and if this is the case, then all levels of public good result in maximal profits of 0. Thus, the Lindahl equilibrium is independent of the way in which profits are distributed among the consumers. Because the utility functions of both consumers are strictly increasing, the budget constraints will hold with equality in equilibrium. Using substitution, we thus can write the utility-maximization problem for consumer 1 as

$$\max_x \; x^{1/4}(8 - p_1^* x)^{3/4}.$$

From this, it is easily derived that consumer 1 demands an amount $2/p_1^*$ of the public good. Similarly, we find that consumer 2 demands an amount $6/p_2^*$ of the

public good. In equilibrium, these amounts have to be equal, which together with $p_1^* + p_2^* = 2$ leads to $p_1^* = 1/2$ and $p_2^* = 3/2$. Thus, we find that $x^* = 4$, $y_1^* = 6$, and $y_2^* = 2$.

This example illustrates that if public good production exhibits constant returns to scale, profit maximization implies that, in equilibrium, the public good is only produced if the sum of the personalized prices is equal to the constant marginal cost. As a result, the maximal profit of the producer is 0 and the equilibrium is independent of the profit distribution among consumers. In such cases there is a clear relationship between a consumer's budget constraint ($p_i^* x + y_i \leq w_i$) and the costs of public-good production, because $p_i^* x = \frac{p_i^*}{m} c(x)$ for all levels x of public good, where m is the constant marginal cost.

One important instance of constant returns to scale in public good production occurs when the public good is measured in terms of expenditures. However, in cases where the cost function c does not exhibit constant returns to scale, clearly the public good must be measured in some other way than expenditures. Such a case is exhibited in the following example.

Example 2 Lindahl equilibrium in an economy with decreasing returns to scale.

Consider the 2-consumer public good economy with $c(x) = x^2$, $N = \{1, 2\}$, $w_1 = 4, w_2 = 6, u_1(x, y_1) = x + y_1$, and $u_2(x, y_2) = 3x + y_2$. Profit maximization for the producer results in $x^* = \frac{p_1^* + p_2^*}{2}$ and the producer has a profit of $\left(\frac{p_1^* + p_2^*}{2}\right)^2$, which needs to be distributed among the consumers. For consumer 1's utility-maximization problem to have an interior solution, the consumer's budget line needs to have the same slope as any of her indifference curves; $p_1^* = 1$. Similarly, $p_2^* = 3$ in a Lindahl equilibrium. This leads to $x^* = 2$ and the firm making a profit equal to 4. Thus, with a distribution rule $d = (d_1, d_2)$ for profits, we find that $y_1^* = 4 - 2 + 4d_1 = 2 + 4d_1$ and $y_2^* = 4d_2$.

In this example, the profit-maximization condition guarantees a unique level of public-good production in equilibrium. However, the Lindahl equilibrium conditions put no restrictions on the distribution rule and there are many Lindahl equilibria with different outcomes for the consumers. This multiplicity (and the need to re-distribute profits) stems from the fact that the consumers do not take the actual cost $c(x)$ of production into account in their utility-maximization problem, but consider a linear budget constraint $p_i^* x + y_i \leq w_i + d_i \left(\left(\sum_{j \in N} p_j^*\right) x^* - c(x^*)\right)$ that is obtained from a fictional personalized per-unit price p_i^* for the public good.

There is a large literature related to Lindahl equilibrium, mainly focussed on the efficiency and re-distribution properties that our last example illustrates.[5] It is not my goal to cover this literature. Instead, I am interested in the personalized prices-based nature of Lindahl equilibrium and the origins of this particular feature of the concept.

[5] See, for example, Silvestre [22].

3 Lindahl [11]

In this section I discuss the ideas for an equilibrium expressed in Lindahl (1919)[6] and show that these ideas involve consumers anticipating having to pay shares of the cost of public good production. Of course, in general, this results in budget sets that are not linear and therefore are very different from the ones obtained when consumers are assumed to face fixed per-unit prices for public good. The ideas in Lindahl [11] are more closely related to the costs of public good production than the current-day definition of Lindahl equilibrium expresses.

As was customary at the time, Lindahl [11] contains very little mathematical elaboration. Particularly, it does not contain an explicit mathematical definition of an equilibrium concept, but expresses ideas verbally and graphically. The discussion below reflects these aspects of the paper.

Lindahl states "We may begin by assuming that there are only two categories of taxpayers [...]. Within each category all individuals must pay the same price for their participation in public good consumption. The problem is the relative amount of the two prices, *i.e.* the distribution of the total cost of the collective goods between the two groups." and "One party's demand for certain collective goods at a certain price appears from the other party's point of view as a supply of these goods at a price corresponding to the remaining part of total cost: for collective activity can only be undertaken if the sum of the prices paid is just sufficient to cover the cost." We see in these two quotes that price is used in the meaning of 'part of total cost', which implies that Lindahl uses the term *price* not as we use it today in economics – meaning as a variable to be multiplied by a quantity in order to figure out how much to pay – but as a total amount to be paid.

This interpretation is confirmed when Lindahl continues discussing in terms of shares of total cost. He states that "the question of distribution really means how big a share of certain total costs each party has to bear" and his idea is that "since the extent of collective activity is not given a priori, but is one of the variables of the problem, the absolute amount of taxation has to be determined at the same time as its distribution" and thus "the extent of collective activity desired by the tax payers becomes largely decisive for their cost share". Thus, in Lindahl's view the issue is the determination of the shares of the cost of public good provision to be paid by each of the parties. Moreover, because the total costs depend on the amount of public good provided, the determination of the cost shares and the level of provision must be determined simultaneously.

Lindahl proceeds to address this simultaneous determination and illustrates "the manner in which equilibrium is established" with a diagram, which is re-produced in Fig. 1.[7] As Lindahl explains, in this figure the variable on the horizontal axis

[6] To be precise, to the translation thereof in *Classics in the Theory of Public Finance*, Eds. R. Musgrave and A. Peacock, 1958.

[7] In this re-production I have left out a few markers that Lindahl uses in a part of his subsequent discussion that I am not covering.

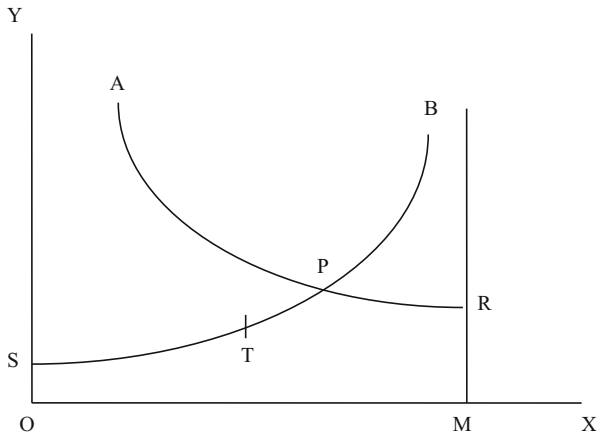


Fig. 1 Lindahl's diagram

represents "the relative share of one party (A) in total cost at various distribution ratios. At point O party A pays nothing at all towards total cost, leaving the entire burden to the other party, B. The further we move away from O, the greater becomes A's share and the smaller B's. At point M the situation is completely reversed; A carries the whole burden and B none." The variable on the vertical axis is "the amount of public expenditure which each party is prepared to sanction at the various distribution ratios." The two curves labeled A and B in the figure represent the "monetary expression of the marginal utility of total public activity for the two parties" and because "demand rises up to the point where marginal utility equals price," these curves represent "demand for public goods" as a function of the "part of public expenditure" that each party has to shoulder.

Lindahl states his equilibrium idea as follows: "The intersection point of the two curves indicates the only distribution of costs at which both parties agree on the extent of public activity." As part of his explanation, he offers "Let us suppose, for example, that the two parities initially agree to split the costs in equal parts. A provisional equilibrium will be established at point T. But only half of A's demand is satisfied and this party will insist on an expansion of public activity. Party B can agree to this only if it can secure a more favourable distribution of costs, and A will have to face the fact that it must take on a greater share of the cost burden. [...] the shift of the equilibrium position towards P continues smoothly only so long as A's growing sacrifice - and it grows in a double sense, by virtue of both the increase in public expenditure and of the increase of A's share in the cost - is more than compensated by the greater utility due to the expansion of collective activity."

Lindahl's description of equilibrium thus involves the parties weighting their demand for public good against the share of public expenditure that they have to shoulder. When considering their demand for public good, the parties take into

account their share of the cost of public good provision and how this share may have to change if they demand alternate levels of public good.

Lindahl [11], apparently far ahead of his time,[8] provides an algebraic illustration of the preceding discussion. This illustration starts as follows: "Party A contributes fraction x to the total public expenditure and party B hence $1 - x$; y is the amount of public expenditure expressed in money; $f(y)$ and $\phi(y)$ are the monetary expressions of the total utility of this expenditure for A and B respectively. Curve A then has the equation

$$f'(y) = x$$

where $f'(y)$ is the utility increment accruing to A from the last unit of money spent and x is the proportion in which A has contributed to this money unit." This shows that Lindahl envisions the parties paying shares of the costs of public good production, and that in this particular illustration the public good is measured in terms of its expenditures.

4 From Lindahl [11] to Lindahl Equilibrium

As we saw in Sect. 2, in a Lindahl equilibrium the parties base their demand for public good on personalized prices which are not necessarily a realistic measure of the cost at all levels of public good production. In the current section I explain how the literature has evolved from the ideas expressed in Lindahl [11] to the definition of Lindahl equilibrium.[9]

To help the reader to follow the flow of ideas throughout the literature that I am describing below, I include an Appendix with a reference graph that shows the relations between the papers I address. This reference graph encompasses all the papers that are relevant to the development of Lindahl equilibrium into a concept based on personalized prices. As I will explain below, this feature of Lindahl equilibrium has become established by the time of publication of Milleron [14], which is why that paper is shown as the most recent one in the reference graph, whereas the term "Lindahl equilibrium" was coined in Foley [5]. At the base of the reference graph (i.e., papers that do not reference other papers in the graph) are Lindahl [11], which is generally believed to be the source of Lindahl equilibrium, and also Bowen [1], Samuelson [19], and Novick [17]. Bowen [1] is included

[8] This is evidenced by the discussion of Novick [17] and Enke [3] in Sect. 4. Lindahl expresses that he is indebted to Knut Wicksell for the algebraic illustration.

[9] One matter that I have to deal with is that in the earlier literature references to others' work are not always explicit and when they are explicit, the references are generally listed in full in the text or in a footnote. In what follows below, I acknowledge references as precisely as I can and when a full reference is available, I acknowledge it in a modern-day format. This has changed the way these references appear in the quotes from older papers.

because Samuelson [20] draws on it for certain aspects of what is to become Lindahl equilibrium. Samuelson [19], interestingly, does not refer to Lindahl [11]; the reason for this becomes clear in Samuelson [20], where we read that he did not have access to Lindahl [11]. Novick [17] is not relevant for the development of Lindahl equilibrium per se, but is included because, as I will explain below, this paper and Enke [3] provide background information about the use of mathematics in economics that turns out to be relevant for the purpose of the current paper.

In the remainder of this section, I discuss the papers included in the reference graph in chronological order (from older to newer) and show how these papers build on one another. The discussion will reflect the gradual progression of the literature from verbal and graphical explanations of ideas to their mathematical formalizations.

4.1 Musgrave [15]

Musgrave [15] pretty closely follows the ideas in Lindahl [11], and writes about "relative distribution of tax shares", "portion of the total cost", and "percentage of the total cost". He states "a final agreement between the two is reached at a volume of the public services at which the sum of the percentage shares which both are willing to contribute equals 100 per cent of the cost of supplying these services" (pages 215 and 216). On page 217, Musgrave starts to use the terminology *price* and *per cent* interchangeably, for example when he states "B's demand price (requesting A to contribute SH per cent) falls below A's supply price (his willingness to contribute SD per cent)". Musgrave posits "The preceding exposition of the theory agrees with Lindahl's version, which both in refinement and conciseness of argument is superior to those of other authors of the school" and proceeds to discuss the "pricing process" in terms of "percentages" and "ratio of cost distribution".

Musgrave [15] verbally describes an equilibrium in which two parties are motivated by shares of the costs of public good provision, and he uses the word prices interchangeably with words like percentages, ratios, and shares. Clearly, he does not use the word price in the sense in which we would use it today—namely as a monetary amount per unit of the public good that is invariable with the level of provision of said good. This difference will become critical when other authors start formalizing the ideas because then prices will notationally have a very specific meaning.

4.2 Bowen [1]

Bowen [1] does not refer to earlier work by Lindahl or Musgrave, but I include it here because, as we will see later, Samuelson [20] draws on this paper and points

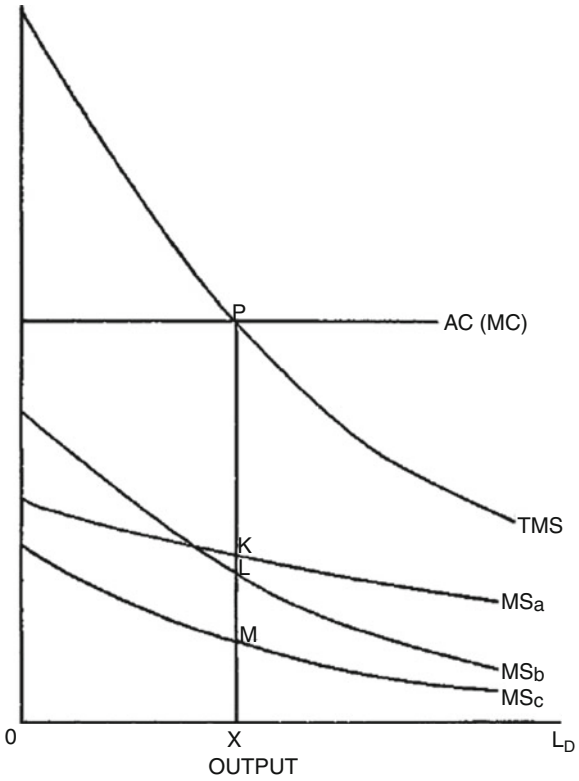


Fig. 2 Figure 1 in Bowen [1]

out that it shows a large similarity with Lindahl [11]. Bowen [1] recognizes that it is important "to establish meaningful units in which quantities of social goods may be measured". He states that one approach is to measure these "simply in terms of their money cost", but that in many cases "physical units would be preferred" and in such cases "increasing, constant, or decreasing cost may apply, whereas if cost units are used, only constant cost may apply".

Continuing, Bowen states that "assuming a "correct" distribution of income, each person's taste [for public good] can be expressed by a curve indicating the amount of money he would be willing to give up in order to have successive additional quantities made available in the community". He includes a "Figure 1 for a community which is assumed to contain three persons", a copy of which appears as Fig. 2 in the current paper. The figure shows the "curves of individual marginal substitution" (MS) for each of the three consumers between a social good (X) and "other goods (money)". For each quantity of the public good, the marginal rates of substitution of the three individuals are added to give the "curve of total marginal substitution" (TMS). Bowen likens this curve to the "familiar curve of total demand" and then states "One of the cardinal principles in determining the output

of an individual good is that price should equal cost, i.e. average cost or marginal cost, whichever is lower. This implies that the ideal output is indicated by the point of intersection between the curve of total marginal substitution and the appropriate cost curve." Bowen then simply continues to refer to this intersection point as "the optimum output of social goods" without explaining or motivating why principles for individual goods should be extended like this to social goods.

Bowen acknowledges that the cost curve may show increasing or decreasing cost, but that his Figure 1 assumes constant cost. Thus, he explicitly allows for the marginal cost of public good to not be constant, but a reader who is mainly looking at his figure may miss this point.

Bowen [1] shows similarity to Lindahl [11] with the statement that "each individual's preference will depend upon [...] the cost *to him* of different amounts of" the social good and that this "will depend partly on the total cost to the community of different amounts and partly on the contemplated distribution of that cost among different individuals" (page 34, italics in original). Bowen concludes from this that each individual will want a quantity of the social good "at which *his* marginal rate of substitution is equal to *his* marginal cost".

The similarity to Lindahl [11] ends here, however, because Bowen stipulates, for general cost cases, that "the cost must be raised by means of a tax levied upon each individual in the form of a "price" per unit of the social good", where "the price is to remain constant regardless of output" and is "equal to his marginal rate of substitution at the particular amount of the good being produced".

4.3 Samuelson [19]

Samuelson [19] does not have any explicit references, but starts with a statement "Except for Sax, Wicksell, Lindahl, and Bowen, economists have rather neglected the theory of optimal public expenditure", and also mentions "published and unpublished writings of Richard Musgrave", "theories of public finance of the Sax-Wicksell-Lindahl-Musgrave type", and "Bowen's writings of a decade ago".

Samuelson [19] contains what appears to be the first attempt to formalize a theory of public expenditure using mathematical notation. He states that "in simple regular cases" a "best state of the world" is defined mathematically by the "marginal conditions"

$$\frac{u^i_j}{u^i_r} = \frac{F_j}{F_r} \qquad (i = 1, \ldots, s;\ r, j = 1, \ldots, n) \tag{1}$$

$$\sum_{i=1}^{s} \frac{u^i_{n+j}}{u^i_r} = \frac{F_{n+j}}{F_r} \qquad (j = 1, \ldots, m;\ r = 1, \ldots, n) \tag{2}$$

$$\frac{U_i u^i_k}{U_q u^q_k} = 1 \qquad (i, q = 1, \ldots, s;\ k = 1, \ldots, n) \tag{3}$$

where $\{1, 2, \ldots, i, \ldots, s\}$ is the set of individuals, each with a utility function $u^i(X_1^i, \ldots, X_{n+m}^i)$ for consumption of "private consumption goods" $X_1, \ldots, X_n$ (with $X_j = \sum_1^s X_j^i$) and "collective consumption goods" $X_{n+1}, \ldots, X_{n+m}$ (with $X_{n+j} = X_{n+j}^i$ for every individual i), with partial derivatives $u_j^i = \frac{\partial u^i}{\partial X_j^i}$, where F models a "convex and smooth production-possibility schedule relating totals of all outputs, private and collective; or $F(X_1, \ldots, X_{n+m}) = 0$, with $F_j > 0$ and ratios F_j / F_n determinate and subject to the generalized laws of diminishing returns to scale", and where $U = U(u^1, \ldots, u^s)$ is a social welfare function with positive partial derivatives U_j. Samuelson explains that the set of conditions (2) are the new element added and that these conditions constitute "a pure theory of government expenditure on collective consumption goods". Note that Samuelson's description allows for production of public goods to not exhibit constant returns to scale and that his marginal conditions do not depend on prices.

Samuelson [19] continues by stating that "the involved optimizing equations" can be solved using "competitive market pricing" under some conditions, which include "the production functions satisfy the neoclassical assumptions of constant returns to scale" and "all goods are private". Only for cases satisfying his conditions does Samuelson [19] introduce prices into his analysis: "We can then insert between the right- and left-hand sides of (1) the equality with uniform market prices p_j / p_r and adjoin the budget equations for each individual $p_1 X_1^i + p_2 X_2^i + \cdots + p_n X_n^i = L^i$ where L^i is a lump-sum tax for each individual so selected as to lead to the "best" state of the world". Thus, we see that Samuelson [19] uses prices in his analysis, but that their use is limited to cases of private goods and constant returns to scale. Samuelson explicitly states that the use of prices cannot be extended to cases where collective consumption is not zero: "*However no decentralized pricing system can serve to determine optimally these levels of collective consumption.*" (emphasis in original).

4.4 Novick [17] and Enke [3]

There is an interesting exchange related to Samuelson [19] that may help explain how personalized per-unit prices made their way into the definition of Lindahl equilibrium in subsequent literature. The exchange concerns the rising use of mathematics in economics, which has as a side effect that "many able economists, especially the older and more experienced ones, cannot comment on some published ideas because they cannot "read" them" (quoted from Enke [3, p. 131]). The exchange starts with Novick [17], who argues that "the mathematically uninitiated jump from theory to proof to application without recognizing the intervening steps that usually must be worked out" and who calls for "a broader discussion of these limitations of the mathematical expressions currently used increasingly in the social sciences." Enke [3] states that Novick's paper "has rather unexpectedly been made the center of controversy" and writes his paper as "a rejoinder" to this controversy,

which is left unspecified and apparently takes place in the general public domain. He goes on to consider Samuelson [19] as an illustration because it "follows hard upon some of the wise precepts suggested in reply to Novick" and "Samuelson himself cites his work as an example of the uses of mathematical economics." Enke [3] states about Samuelson [19] "it is unnecessarily unintelligible to most people. Many economists, interested in public finance and welfare, will want to understand what anyone of Samuelson's reputation has to contribute. Frustration will be their lot."

Hence, at the time of Samuelson's [19] attempt to formalize a theory of public expenditures, economists in general were not very comfortable with the use of mathematics. It is therefore not hard to imagine that, somewhere along the way, someone may have picked up on Samuelson's easier-looking expressions that include prices and overlooked the fact that Samuelson intended these to be valid for only very specific cases. A study of subsequent literature demonstrates exactly what happened.

4.5 Samuelson [20]

Samuelson [20] is apparently written partly in response to the criticism raised in Enke [3] and "presents in terms of two-dimensional diagrams an essentially equivalent formulation" of the ideas in Samuelson [19]. Samuelson [20] is the first instance I find of the use of the term "public consumption good", which is defined as a good for which "each man's consumption of it [...] is related to the total [...] by a condition of *equality* rather than of summation." Samuelson then explicitly sets the consumption of public good by individuals 1 and 2 equal ($X_2^1 = X_2^2 = X_2$) and relates this to the total consumption $X_1^1 + X_1^2 = X_1$ of private good by means of a "production-possibility or opportunity-cost curve" that "relates the total productions of public and private goods in the usual familiar manner: the curve is convex from above to reflect the usual assumption of increasing relative marginal costs (or generalized diminishing returns)". A copy of Samuelson [20]'s Chart 3 (page 351) appears as Fig. 3 in the current paper. It is clear from this figure that Samuelson [20] does not assume constant returns to scale in public good production.

Samuelson [20] proceeds to derive graphically the tangency conditions that are necessary for Pareto optima[10] and the set of utility possibilities in Pareto optimal points (the curve labeled pp in Chart 4, a copy of which appears as Fig. 4 in the current paper). This together with some contours of a "social welfare function" illustrates the "best configuration for this society". Samuelson verbally explains that "this final tangency condition" has the interpretation that "The social welfare significance of a unit of any private good allocated to private individuals must at the

[10]Samuelson uses the point E and the curve $C'D'$, which represents an indifference curve for one of the consumers, for this derivation.

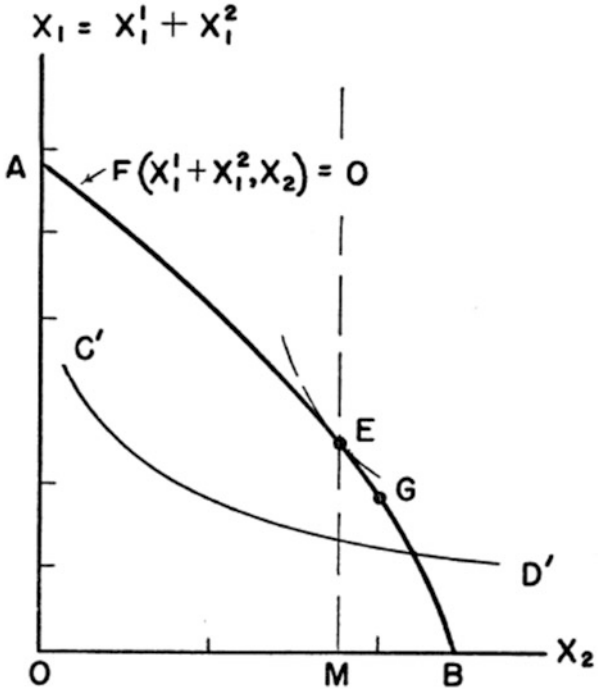


Fig. 3 Chart 3 in Samuelson [20]

margin be the same for each and every person" and "The Pareto-optimal condition, which makes relative marginal social cost equal to the sum of all persons' marginal rates of substitution, is already assured by virtue of the fact that bliss lies on the utility frontier". Note that Samuelson does not advocate for a specific Pareto optimal point, but shows how to find one that is best for society if a particular social welfare function is given. His work is related to identifying a level of public good that maximizes social welfare, and not to how a society can arrive at that level through decisions by individuals.

After completing "the graphical interpretation of my mathematical model", Samuelson [20] relates his graphical treatment to earlier work by Lindahl and Bowen, and he does so by means of Chart 5 (page 354), which I include in the current paper as Fig. 5. Samuelson [20] derives an "*MC* curve" with "*MC* measured in terms of the numeraire good", by plotting "the absolute slope" of the opportunity-cost curve "against varying amounts of the public good". Doing similarly for the individual indifference curves to obtain individuals' *MRS* curves, Samuelson then arrives at a picture that shows the addition of the *MRS* curves (added in the dimension of the private good) and he intersects the result with the *MC* curve to obtain "equilibrium". Note that Samuelson is using the word equilibrium in a context where he is characterizing a level of public good that satisfies the tangency

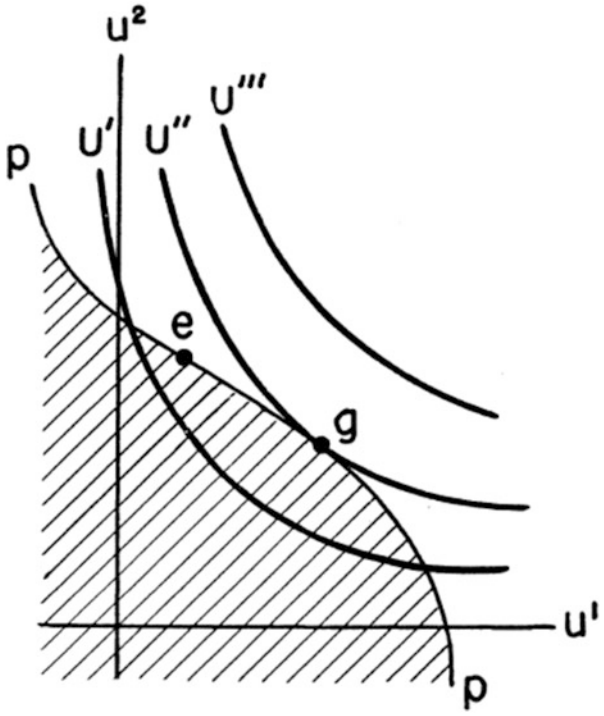


Fig. 4 Chart 4 in Samuelson [20]

conditions necessary for Pareto optima. His Chart 5 does not address the individuals' numeraire good consumption levels.

Samuelson acknowledges that "except for minor details of notation and assumption" his Chart 5 (Fig. 5 in the current paper) is identical with Figure 1 in Bowen [1] (Fig. 2 in the current paper) and goes on to say "anyone familiar with Musgrave [15] will be struck with the similarity between this Bowen type of diagram and the Lindahl 100-per-cent diagram reproduced by Musgrave [15]."[11] However, as we have seen, Figure 1 in Bowen [1] is for the special case of constant returns to scale. Moreover, the similarity between the "Bowen type of diagram" and "the Lindahl 100-per-cent diagram" is only valid in this special case, as becomes clear in the next subsection.

[11]Referring to Lindahl [11], Samuelson states (Footnote 8) that he has "not had access to this important work".

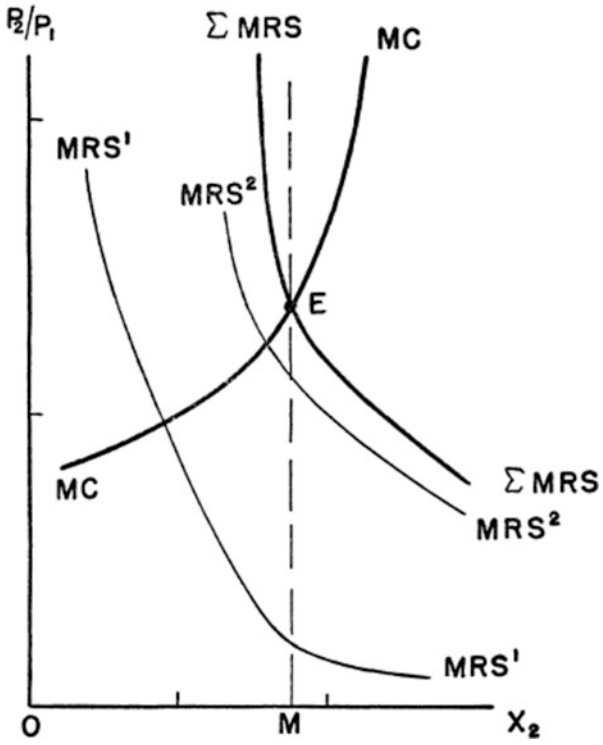


Fig. 5 Chart 5 in Samuelson [20]

4.6 Musgrave [16]

Musgrave [16] states that his discussion of Lindahl [11] goes back to Musgrave [15][12] and discusses two taxpayers who agree to "contribute certain percentages of the total cost" of "whatever volume of social goods is supplied". Musgrave illustrates his discussion with two figures captioned "Bowen model" and "Lindahl model", respectively (page 75), a re-production of which is included as Fig. 6 in the current paper.[13] Both figures have "units of social goods" on the horizontal axis, and the first figure has "combined unit price" on the vertical axis, while the second has "per cent of cost contributed" on the vertical axis. Musgrave describes how the second picture can be obtained from the first by looking at percentages of S paid by a rather than absolute amounts, which leaves the curve for a in place and basically mirrors the curve for b in the horizontal 50 % line. He states that

[12] Actually, he has the year wrong in his own reference – in footnote 1 on page 74 he says 1938 – but this is clearly a typo.

[13] I have left several lines out that are unnecessary for my purposes.

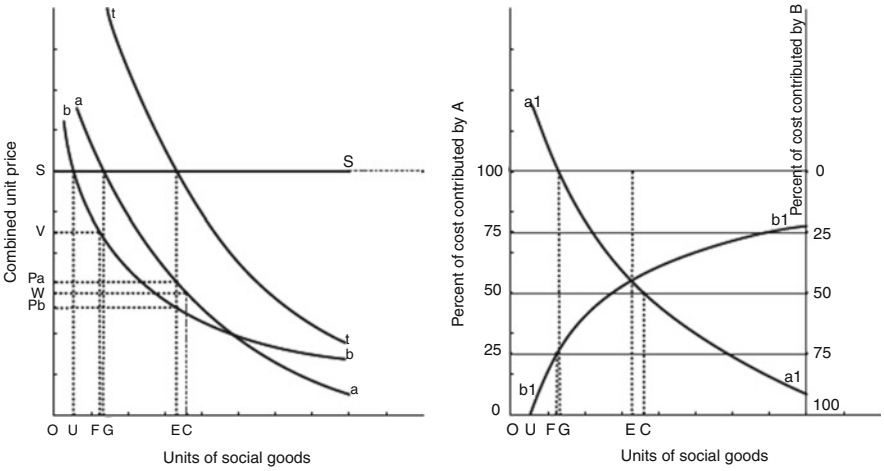


Fig. 6 Figures 4-1 and 4-2 in Musgrave [16]

"the resulting price determination" is shown in both figures, where the amount E is obtained in the graph on the left as the quantity where the sum of curves *aa* and *bb* intersects *SS*, and in the graph on the right as the quantity at which the curves *a1a1* and *b1b1* intersect. He states, apparently as an afterthought and very casually, "*SS* is the supply schedule of social goods that we assume are produced under conditions of constant cost" (page 76). In a footnote he then remarks that his figures "may be adapted to conditions of increasing cost", but he does not elaborate.

It is unclear how Musgrave intends to adapt his figures to cases of increasing cost. It does not take much effort to see that if we were to draw a graph like the one on the left in Fig. 6 but with an increasing cost function, and we were then to construct a graph like the one on the right that corresponds to this increasing-cost example, then to fit the scales of "per cent cost contributed" by A and B, we would have to interpret cost as total cost and then the two graphs may not predict the same quantity of social goods anymore. Thus, Musgrave effectively links prices and percentages of costs only in cases of constant returns to scale. However, in the discussion of Lindahl's work that follows the two figures, Musgrave continues to use the terms *price*, *per cent of the cost*, and *cost share* interchangeably, without limiting himself to cases of constant cost.

4.7 Johansen [7]

Johansen [7] "is an attempt to present Lindahl's solution in terms of more modern welfare-theoretical concepts and thereby to bring out some new aspects of the

solution" and in passing suggests that "the Lindahl solution has not been quite satisfactorily presented" in Musgrave [16].

Johansen [7] considers two parties with private consumption and "the amount of public expenditures G", all "measured in monetary units" and he assumes "prices to be fixed". He uses "utility functions [...] with private and public consumption entering side by side" as were used in Samuelson [19]. Johansen [7] follows Lindahl and has the two parties pay fractions h and $1-h$, so that the "absolute burdens levied" are hG and $(1-h)G$. It is important to note that Johansen measures public good in its expenditures and thereby limits himself to a setting where cost functions for public good exhibit constant returns to scale. Thus, he obtains budget constraints that are linear in G.

Johansen derives the Lindahl diagram using these budget constraints and indifference curves of the utility functions and states (page 349): "The solution is [...] analogous to the determination of an equilibrium in a perfectly competitive market: The equilibrium price is the price at which buyers and sellers "agree" upon the quantity to be traded, when both sides of the market consider the price as given." In his mathematical exposition of "the Lindahl solution", Johansen [7] derives that when person A maximizes his utility F_A with respect to public expenditures G, while considering A's "income" R_A and "price" h as given, the optimality condition $\frac{\partial F_A}{\partial G} = h\frac{\partial F_A}{\partial X}$ determines the amount of public expenditure wanted by person A (G_A) as a function of R_A and h. He states that this "may be compared with an ordinary demand function with h being the price" and explains that the "fraction of the total public expenditure" paid by a person depends on "the "equilibrium value" of h".

From the preceding quotes, we see that Johansen [7] describes the fraction h of public expenditures using the word price in the meaning of constant per-unit price. The two words can be used interchangeably because he measures public good in terms of its expenditures. However, it is not hard to imagine how this use of the word price may get extended to situations where public good is measured in some sort of physical quantities and the marginal costs may no longer be constant.

4.8 Foley [4]

Foley [4] acknowledges the earlier work by Lindahl, Musgrave, Samuelson, and Johansen, but he defers "all references to sources" to an appendix and thus it is not possible to tell to whom he attributes which parts of his analysis. This paper introduces and studies "a straightforward generalization of ordinary competitive equilibrium" that Foley names "public competitive equilibrium". A public competitive equilibrium consists of three elements: (a) one bundle of public goods and for each agent a bundle of private goods, (b) per-unit prices for all goods – private and public – and (c) a vector of lump-sum taxes for the agents that add up to the cost of the public-goods bundle. Foley [4] assumes that "production follows a rule of convexity" and writes that "it does not matter very much what units are used to measure the production of public goods as long as the measurement allows each

person to decide which of two bundles (including both public and private goods) he prefers". Thus Foley [4] deviates in an important way from previous literature because he has per-unit prices for public goods without necessarily having constant returns to scale in production.

4.9 Samuelson [21]

Samuelson [21] is "a brief review of the analysis" of the "unifying theory of optimal resource allocation to public or social goods and optimal distribution of tax burdens" previously presented in Samuelson [19, 20], an analysis of which "the partial equilibrium analysis of Lindahl and Bowen" is "a special case". Samuelson [21] states that with the "notable exception" of Johansen [7], "a number of the theory's essential points have [...] been misunderstood by writers in the field". The paper proceeds to describe "equilibrium" in terms of "*pseudo*-tax-prices" for the public goods, which are personalized per-unit prices that are independent of the actual levels of public goods produced. Unfortunately, any restrictions on the costs for production of public goods do not appear in the body of the paper, but are relegated to an appendix.

4.10 Foley [5]

Foley [5] is the paper in which I find the first use of the name "Lindahl equilibrium". This is a very different paper from Foley [4] in two important respects: In an apparent change of heart, Foley now defines a different equilibrium and also adds a condition of constant returns to scale. Foley [5] acknowledges "the rehabilitation and reconstruction of Lindahl's quasi-demand solution to the taxation problem", which he attributes to Johansen [7] and Samuelson [21],[14] and provides the following definition:

A *Lindahl equilibrium* with respect to $w = (w^1, \ldots, w^n)$ is a feasible allocation $(x; y^1, \ldots, y^n)$ and a price system $(p_x^1, \ldots, p_x^n; p_y) \geq 0$ such that

(a) $$\left(\sum_{i=1}^{n} p_x^i; p_y\right)\left[x; \sum_{i=1}^{n}(y^i - w^i)\right] \geq \left(\sum_{i=1}^{n} p_x^i; p_y\right)(\bar{x}; \bar{z}) \text{ for all } (\bar{x}; \bar{z}) \in Y$$

(b) $$\text{if } (\bar{x}^i; \bar{y}^i) \succ_i (x; y^i) \text{ then } p_x^i \cdot \bar{x}^i + p_y \cdot \bar{y}^i > p_x^i \cdot x + p_y \cdot y^i = p_y \cdot w^i.$$

Here, $x = (x_1, \ldots, x_m)$ denotes a vector of public goods, $y^i = (y_1^i, \ldots, y_k^i)$ denotes a vector of private goods consumed by a consumer i, z denotes a vector

[14] Foley [5] references an "unpublished manuscript" by Samuelson without a date. However, this manuscript has the same title as Samuelson [21] as referenced in Malinvaud [12].

of private goods used in production, w^i denotes initial endowments of private goods of consumer i, and Y denotes "the set of all technically possible production plans" for producing public goods from private goods. The first condition on the production set Y (B.1. on page 67) is "Y is a closed, convex cone". Thus, Foley [5]'s definition of Lindahl equilibrium is limited to public goods whose production technologies exhibit constant returns to scale.

4.11 Milleron [14]

Milleron [14] is a survey article with some "original contribution", and it takes elements from many of the papers that I have covered up to now. Discussing Samuelson [19, 20], Milleron asserts that the main result of those papers was to show that "with any Pareto optimal situation may be associated a system of "prices", [...] the price paid for the public goods being "*personalized*." These personalized prices [...] may be interpreted as a contribution of each agent to the production of each public good. The sum of contributions is then equal, for each public good, to the production price of this good." Milleron [14] proceeds by deriving "Samuelson prices" under "a rather general set of assumptions". These assumptions, however, do not include constant returns to scale in production, which was a condition Samuelson was careful to include when invoking the existence of prices for public goods.

Milleron [14] refers to Johansen [8] for a discussion of the concept of Lindahl equilibrium, and states "Lindahl's idea was that, for a given production price of a public good, it is meaningful to define personalized prices such that the sum of these personalized prices is equal to the production price. Thus, it is possible to define the "demand" for public good by each consumer as a function of the corresponding personalized price". On page 439, Milleron [14] provides a definition of Lindahl equilibrium that includes personalized per-unit prices.

I conclude that at this point in the literature, Lindahl equilibrium has become established as a concept based on personalized per-unit prices for public goods.

5 Lindahl [11] and Ratio Equilibrium

In this section, I describe the relation between Lindahl [11] and ratio equilibrium. I start by providing a definition of ratio equilibrium, then discuss the axiomatic link between the ideas expressed in Lindahl [11] and ratio equilibrium, and conclude this section with a discussion of relations between ratio equilibrium and Lindahl equilibrium.

5.1 Ratio Equilibrium

A ratio equilibrium, first defined in Kaneko [9], consists of a set of personalized ratios – one for each consumer – and for each consumer a consumption bundle of public and private good amounts. For each consumer i, her ratio r_i determines a budget set as follows. If consumer i wants to consume a specific amount of the public good, then the amount of private good she can consume is diminished by her share r_i of the cost of production of her demanded level of public good. In equilibrium, each consumer i consumes a utility-maximizing bundle of public good and private good within her budget set. In addition, all consumers must choose the same amount of the public good and the ratios of all consumers must add to 1, so that the costs of public good production are covered. Formally, a *ratio equilibrium* consists of a vector of ratios $\mathbf{r}^* = (r_i^*)_{i \in N}$ and consumption bundles $(x^*, y_i^*)_{i \in N}$ such that

$$\sum_{i \in N} r_i^* = 1,$$

$$x^* \text{ is a solution to } \max_x u_i(x, w_i - r_i^* c(x))$$

for each $i \in N$, and

$$y_i^* + r_i^* c(x^*) = w_i$$

for each $i \in N$.

I illustrate ratio equilibrium in the following example.

Example 3 Ratio equilibrium in an economy with decreasing returns to scale.

Consider the 2-consumer public good economy with $c(x) = x^2$, $N = \{1, 2\}$, $w_1 = 4$, $w_2 = 6$, $u_1(x, y_1) = x + y_1$, and $u_2(x, y_2) = 3x + y_2$. Using substitution, we write consumer 1's utility-maximization problem as

$$\max_x \; x + 4 - r_1^* \left(x^2\right),$$

from which it is easily derived that $x^* = 1/(2r_1^*)$. From consumer 2's utility-maximization problem we similarly find that $x^* = 3/(2r_2^*)$. In equilibrium, these amounts have to be equal, which together with $r_1^* + r_2^* = 1$ leads to $r_1^* = 1/4$ and $r_2^* = 3/4$. Thus, we find that $x^* = 2$, $y_1^* = 3$, and $y_2^* = 3$.

Kaneko [9] defines ratio equilibrium because "the concept of Lindahl equilibrium has a difficulty in normative meanings" and because "the core never coincides with the Lindahl equilibria". Kaneko refers to Foley [5] and Milleron [14], but not to Lindahl [11] or any of the other papers that I have covered in Sect. 4, and includes no indication that he is aware that the ratio equilibrium that he defines is a formalization of the ideas presented in Lindahl [11]. Kaneko [9] does, however, include a lemma

that states that when cost functions are linear, there is an equivalence between ratio equilibria and Lindahl equilibria.[15]

5.2 The Axiomatic Link Between Lindahl [11] and Ratio Equilibrium

In van den Nouweland et al. [23], we use the axiomatic method to study ways of determining public good levels and private good consumption by individuals in pure public good economies as described in Sect. 2. While Lindahl [11] does not contain an explicit definition of an equilibrium concept, it is easy to capture the ideas expressed in that paper in axioms. Namely, we require that a group of agents takes the shares of total costs paid by other agents as given when making decisions on how much to demand of a public good and how to cover the remaining costs of its production. Based on this requirement, we define two related axioms, consistency and converse consistency, that relate solutions in smaller economies (i.e., those with fewer consumers) to solutions in larger economies. We show that, together with a simple individual rationality axiom, these axioms that capture the cost-share idea expressed in Lindahl [11] determine a unique equilibrium concept under the very general conditions of strictly increasing utility for consumption of private and public good and non-decreasing costs for public good production in terms of private good. Interestingly, we find that the unique equilibrium concept that satisfies these axioms based on the ideas in Lindahl [11] is not Lindahl equilibrium, but ratio equilibrium.

The axiomatic work in van den Nouweland et al. [23] strongly suggests that ratio equilibrium is a better fit with Lindahl [11] than Lindahl equilibrium is, because ratio equilibrium displays characteristics as described in Lindahl [11], whereas Lindahl equilibrium does not.

5.3 Relations Between Ratio Equilibrium and Lindahl Equilibrium

Ito and Kaneko [6] consider the relationship between Lindahl equilibrium and ratio equilibrium. They do so by looking at similarities and differences between the allocations – levels of public good and each consumer's consumption of private good – that are supported by the two equilibrium concepts. They show that, under a certain condition, every ratio equilibrium allocation can be obtained as a Lindahl equilibrium allocation by varying the re-distribution of the profits from public goods production among the consumers. The condition that they need in their proof is a

[15] Kaneko [9] casually refers to Foley [5] for validation of this result.

weak one, namely that when all money in the economy is used to produce public goods, then every individual has a lower utility than she would have when she just consumed her initial endowment of money. Ito and Kaneko [6] also demonstrate that the possibility of varying the re-distribution of the profits from public goods production among the consumers in a Lindahl equilibrium leads to many more allocations being supported by Lindahl equilibrium than by ratio equilibrium.

Thus, under a mild condition on an economy, the set of ratio equilibrium allocations is a strict subset of the set of Lindahl equilibrium allocations. We see an illustration of this when comparing Examples 2 and 3. Both examples look at the same economy, and for this economy we find exactly one ratio equilibrium allocation, but many Lindahl equilibrium allocations. Also, the ratio equilibrium allocation is obtained as a Lindahl equilibrium allocation with distribution rule $d = (1/4, 3/4)$.

Ito and Kaneko [6] also study what happens to the predictions of the two equilibrium concepts under "cost-linearizing transformations" of units of measurement of the public goods. The motivation for doing this is two-fold. First, Lindahl equilibrium and ratio equilibrium give the same predictions when public good is measured in terms of its costs, because then a fixed share of the costs is equivalent to a personalized price per unit (of cost). Second, starting from any economy, cost functions can be transformed into linear functions by "measuring the public good in terms of minimal costs required for their production". Thus, it is a desirable property that transforming an economy so that public good is measured in terms of its cost, does not change the equilibrium other than to account for the changed units of measurement. Ito and Kaneko [6] find that Lindahl equilibrium is not necessarily invariant under such cost-linearizing transformation, whereas ratio equilibrium always is.

I use the economy in Examples 2 and 3 to illustrate the cost-linearizing transformation of an economy and that this may change the Lindahl equilibrium allocations.

Example 4 Cost-linearizing transformation of an economy with decreasing returns to scale.

Consider the economy in Examples 2 and 3. Measuring the public good in terms of its costs, we denote public good by $X := x^2$. The cost function and the utility functions of the consumers have to be changed accordingly: $c(X) = X$, $u_1(X, y_1) = \sqrt{X} + y_1$, and $u_2(X, y_2) = 3\sqrt{X} + y_2$.

To find all the Lindahl equilibrium allocations of this economy, note that the profit-maximization problem of the producer only has a non-zero solution if $p_1^* + p_2^* = 1$ and that the maximum profit of the producer equals 0. Using methods similar to those in Example 1, it is then easily derived that there is a unique Lindahl equilibrium allocation, with $X^* = 4$, $y_1^* = 3$, and $y_2^* = 3$.

Thus, the cost-linearization of the economy makes the infinite set of Lindahl equilibrium allocations that we found in Example 2 shrink to a single allocation. It is easily verified using methods similar to those in Example 3 that this single allocation is also the unique ratio equilibrium allocation of the cost-linearized economy. And, this allocation coincides with the ratio equilibrium allocation that we found in Example 3, because $(x^*)^2 = X^*$.

6 Conclusions

As I have shown, due to some duplicitous use of the word *price*, uneasiness with mathematical formalizations, and carelessness with conditions of constant returns to scale, somewhere between Lindahl [11] and Milleron [14], *Lindahl equilibrium* became known as a concept based on *personalized prices*. However, van den Nouweland et al. [23] demonstrate that Kaneko's *ratio equilibrium*, which is based on *personalized shares* of the cost of public good production, embodies better the ideas presented in Lindahl [11]. The fact that, as demonstrated in Ito and Kaneko [6], the two equilibrium concepts support similar allocations notwithstanding, the two concepts are very different in method. Because personalized prices linearize consumers' payments for public goods, Lindahl equilibrium needs to include profit-maximization requirements and profits need to be somehow re-distributed among consumers. Ratio equilibrium distributes the costs of public good production according to personalized shares and thus automatically results in zero profits.

Acknowledgements I am grateful to Nick Baigent for suggesting to me that I write this paper, which started as a simple project explaining the discrepancy between the concept known as Lindahl equilibrium and the ideas expressed in Lindahl [11], but ended up being a detective story. I am indebted to Peter Lambert, Nicholas Sly, and two anonymous referees for their comments which helped me improve the paper, and to Myrna Wooders for advocating the axiomatic technique to study equilibrium concepts in (local) public good economies.

Appendix: The Reference Graph

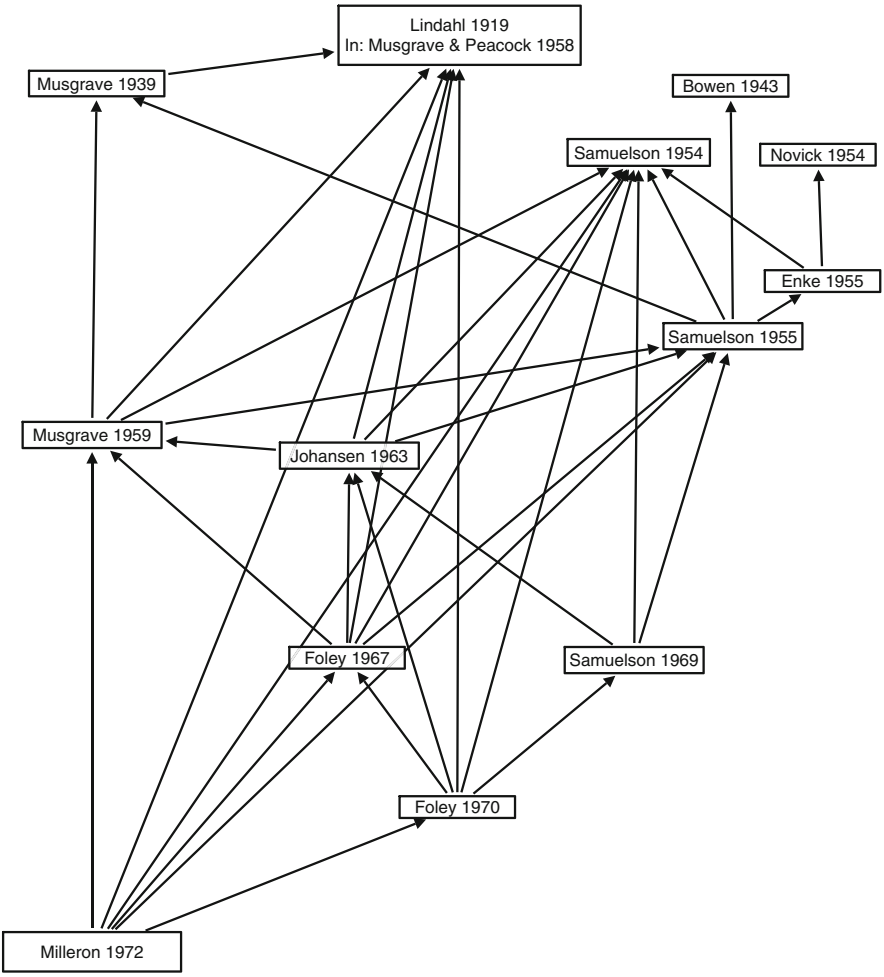


Fig. 7 Reference graph

References

1. Bowen H (1943) The interpretation of voting in the allocation of economic resources. Q J Econ 58(1):27–48
2. Coase R (1960) The problem of social cost. J Law Econ 3:1–44
3. Enke S (1955) More on the misuse of mathematics in economics: a rejoinder. Rev Econ Stat 37(2):131–133
4. Foley D (1967) Resource allocation and the public sector. Yale Econ Essays 7(1):45–98
5. Foley D (1970) Lindahl's solution and the core of an economy with public goods. Econometrica 38(1):66–72
6. Ito Y, Kaneko M (1981) Linearization of cost functions in public goods economies. Econ Stud Q 32(3):237–246
7. Johansen L (1963) Some notes on the Lindahl theory of determination of public expenditures. Int Econ Rev 4(3):346–358
8. Johansen L (1965) Public economics. North-Holland, Amsterdam
9. Kaneko M (1977) The ratio equilibrium and a voting game in a public goods economy. J Econ Theory 16:123–136
10. Kreps D (2013) Microeconomic foundations I: choice and competitive markets. Princeton University Press, Princeton
11. Lindahl E (1919) Just taxation - a positive solution. Translated from German (Die Gerechtigkeit der Besteuerung, Lund 1919, Part I, Chap. 4, pp 85–98: Positive Lösung.") by E. Henderson. In: Musgrave R, Peacock A (eds) Classics in the theory of public finance (1958). Macmillan, London
12. Malivaud E (1971) A planning approach to the public good problem. Swed J Econ 73(1):96–112
13. Mas-Colell A, Whinston M, Green J (1995) Microeconomic theory. Oxford University Press, New York
14. Milleron J-C (1972) Theory of value with public goods: a survey article. J Econ Theory 5:419–477
15. Musgrave R (1939) The voluntary exchange theory of public economy. Q J Econ 53(2):213–237
16. Musgrave R (1959) The theory of public finance: a study in public economy. McGraw-Hill, New York
17. Novick D (1954) Mathematics: logic, quantity, and method. Rev Econ Stat 36(4):357–358
18. Oakland W (1987) Theory of public goods. In: Auerbach A, Feldstein M (eds) Handbook of public economics, vol II, Chap. 9. North-Holland, Amsterdam, pp 485–535
19. Samuelson P (1954) The pure theory of public expenditure. Rev Econ Stat 36(4):387–389
20. Samuelson P (1955) Diagrammatic exposition of a theory of public expenditure. Rev Econ Stat 37(4):350–356
21. Samuelson P (1969) Pure theory of public expenditure and taxation. In: Margolis J, Guitton H (eds) Public economics: an analysis of public production and consumption and their relations to the private sector. Proceedings of a conference held by the International Economic Association, Chap. 5. Macmillan, London, pp 98–123
22. Silvestre J (2003) Wicksell, Lindahl and the theory of public goods. Scand J Econ 105(4):527–553
23. van den Nouweland A, Tijs S, Wooders M (2002) Axiomatization of ratio equilibria in public good economies. Soc Choice Welf 19:627–636

Part IV
An Interview with Nick Baigent

An Interview with Nick Baigent

Constanze Binder, Miriam Teschl, and Yongsheng Xu

The interview was conducted by Constanze Binder (**CB**) Miriam Teschl (**MT**) and Yongsheng Xu (**YX**) via email over a period of a few weeks in the Summer/Fall of 2014.

CB & MT & YX: A few years ago, you gave an interview in "The Reasoner".[1] In that interview, you said you became interested in economics because as a teenager you found Samuelson's famous textbook.What exactly was it in or about this book that "hooked" you to economics?

NB: *It was my first exposure to analytical thinking and I was fascinated by the way points were made by deductions from clear and explicit assumptions.*

CB & MT & YX: How would you describe the main themes that interested you in your own research? Have you been able (or are you able) to answer some of your original questions that motivated you to do that kind of research? How did you choose the questions/problems you worked on?

[1] Dietrich, Franz (2010), 'Interview with Nick Baigent', *The Reasoner* 4(8), 118–120.

C. Binder (✉)
Faculty of Philosophy, Erasmus University Rotterdam, Campus Woudestein, Postbus 1738, 3000 DR Rotterdam, The Netherlands
e-mail: binder@fwb.eur.nl

M. Teschl
Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Centre de la Vieille Charité, 2, rue de la Charité, 13002 Marseille, France
e-mail: miriam.teschl@ehess.fr

Y. Xu
Department of Economics, Georgia State University, Atlanta, GA, USA
e-mail: yxu3@gsu.edu

C. Binder et al. (eds.), *Individual and Collective Choice and Social Welfare*, Studies in Choice and Welfare, DOI 10.1007/978-3-662-46439-7_19

NB: *These questions seem to make assumptions about the experience of research that were not true in my case. The starting point should probably be my argumentative and skeptical predisposition. These attitudes accompanied my reading of the literature and the way I reacted to what I read. But there was no grand question or theme with which I came to research. I am glad that I was not handicapped by such unhelpful attitudes.*
Of course, at a very early stage I really did not know what made for a good research topic and, worse, I did not know that I did not know. However, I was very lucky to encounter several people who did understand what made for a good research topic. Chief among these was Paramesh Ray, Pham Chi Thanh, the late Howard Petith and Wulf Gaertner. Later, I learned a lot from many others particularly Kotaro Suzumura and Prasanta Pattanaik.
Without realizing it, this has something in common with the role of "Fields" in doctoral programs in the USA, namely a critical reading of the literature and the opportunity to discuss it critically with people of sound research judgment. This is far less common in Europe, at least in my experience. In recent years in Europe, I was faced with students wanting me to be their supervisor, whose starting point was a very firm view about a solution to, in their view, a pressing global problem. Sometimes they seemed to know not only the problem but the answer too, and it was not clear to me what was left to do. But the key point is that "solving" a global practical problem is neither necessary nor sufficient for a problem to have intellectual and therefore research interest. I am aware that I am out of step with prevailing attitudes.
So to return to my experience of research, a critical reading and thinking in a general area would sometimes lead either to something that seemed unsatisfactory in the literature, perhaps because I did not understand it sufficiently. Typically, I would find this disturbing and doing research was my way to change this uncomfortable state of affairs. Topological social choice was certainly a topic on which I ended up doing research as a way to understand what was going on. I was struck at the time by how few in my peer group seemed to understand this demanding literature. Of course, when I began my long struggle with this topic, I was at Essex University where Norman Schofield was already eminent.
So this was the phenomenology of my early research experience. It felt more like the topics I worked on chose me and not that I chose them.
CB & MT & YX: Could you elaborate a little bit on using clear assumptions and deductive reasoning to understand the muddy waters of reality?
NB: *Let a muddy river with leaves floating along on it denote the world "as it is". I do not think there are any claims about social, economic and political life as it is in which I have much confidence. I realize this is an extreme epistemological stance with which many, perhaps most, economists would disagree. But given this view, how then do I see economic theory? One way I see it is similar to a cartoon which does not claim to be empirically accurate but which nevertheless may make a point about the world in a very powerful way. How do cartoons achieve this? One way is by abstraction and exaggeration, just like theoretical models.*

Let me offer an illustration, not from social choice theory. In the Diamond-Mirlees theory of optimal commodity taxation, the following questions are posed: In the absence of lump sum taxation, should all commodities be taxed at the same rate? The elegant Diamond-Mirlees[2] *model nicely shows that the answer is no. I very much doubt that reducing abstractions and exaggerations in their model will ever change the answer to this question. Indeed, my intuition is that the introduction of 'more reality' would provide additional reasons for optimal commodity taxes to depart from uniformity.*

Some may wish that "realism" played a greater role than it does in the argument I have just given. But what I think is mistaken is that good points cannot be made without a high degree of realism. In lectures, I used to say that some models in economics make points in much the same way as cartoons do, by abstraction and exaggeration. The same is true for many art forms including opera, ballet, painting, sculpture, puppetry and mime. Among the most effective theatres is that of Noh, with its highly formalized gestures. Departures from realism can be very effective in making points about reality.

Given how often students raise questions about models lacking realism, it might be a good idea for teachers to have a glove puppet or even a couple of finger puppets and give a little performance that demonstrates the point just made.

CB & MT & YX: What reading in economic theory have you most enjoyed?

NB: *A couple of caveats before my answer. What I enjoyed most is not necessarily what I think is the most important. Also, my ranking is not only very incomplete but also unstable in that it varies wildly even in the course of a single conversation. I should add that included among the papers that I have most enjoyed reading are many by authors in this volume.*

However, one paper that is often most enjoyable is Yakar Kannai's paper in Econometrica, 1970. It is not a seminal paper. The seminal paper was Auman's paper in Econometrica, 1964, but I did not appreciate it at first since I did not understand quite how to view a continuum of agents. After reading Kannai I did and it enabled me to appreciate Auman's work and the literature building on it.

CB & MT & YX: Is there any reading outside of economics from which you think a practicing economist could benefit?

NB: *Caveats similar to those for the previous question apply also to this question as well as an additional one. I also have rational choice theory very much in mind, though I think my answer is probably relevant even outside of economics. I have thought a lot about the answer to this question and my thinking started with Lee Anne Fuji's book, "Killing Neighbors: Webs of violence in Rwanda", (Cornell University Press, 2009). This led me to think about rationality within a framework that was new to me, namely Scripts, Directors and Actors. It seems to me that there are analogies to these concepts that are relevant to rational choice. But Lee Anne Fuji's book was only the start of my current journey.*

[2]Diamond, P.A. and J.A. Mirlees (1971), "Optimal Taxation and Public Production II: Tax Rules", *American Economic Review*, 61, 261–278.

I am very tempted to suggest Peter Brook's, "The Empty Space" (Penguin, 1968). However, my choice is the great classic of the theatre, Constantin Stanislavky's, "Creating a Role" (Taylor Francis, 1989). It seems to me that the choices made by directors and actors, given a script, are of interest well beyond dramatic art and run well into rational choice theory. Of course, writers make choices in creating a script; Directors make choices about staging that script; and actors make choices about characters within that script. It seems to me that there are close analogies in economic, social and political life. After all: "All the World's a stage . . . "

CB & MT & YX: You have crossed the disciplinary boundaries by intense interactions with philosophers and lately also collaborating with psychologists in your empirical work on violence[3] In your experience, what are/is the main (a) benefits/advantages, (b) challenges and (c) possible pitfalls of crossing disciplinary boundaries as an economist in academia today?

NB: *I do not think there are any disciplinary boundaries, at least intellectual ones. Until relatively recently there were no university departments in the modern sense based on disciplinary boundaries. There are of course reasons for organizing universities into subject based departments to do with administration, hiring, equal opportunities, tenure and job markets. But I see no good intellectual reason to be constrained by such boundaries and I see more universities developing research group structures along-side departmental structures. The Choice Group at the London School of Economics is a highly effective example.*

Actually, my experience of crossing disciplinary boundaries goes back a long way, to my student days in the Department of Political Economy as the Economics Department was then called at UCL. Even in those days, I was a 'philosophy groupie'. Later, at Tulane, I was in the Murphy Institute of Political Economy in which there were economists, philosophers and political scientists. And I am now very happy indeed in the Choice Group at the London School of Economics.

CB & MT & YX: As supervisor or mentor, you inspired a large number of students to become academics. What are your main goals as (i) a teacher, (ii) a supervisor and (iii) a mentor?

NB: *For the purpose of this question, I would not distinguish between (i) and (ii), so I will try to answer them together.*

One goal dominates all the others and it is as overwhelming as it is vague. It is to open the door to "the life of the mind". Readers of Hermann Hesse's The Glass Bead Game will know exactly what I mean. My mother attempted to encourage me to practice playing a musical instrument by telling me that if I succeeded, it would be something I could always enjoy and I view teaching in a similar way. In teaching, my hope is to help students appreciate the joy and deep satisfaction of thinking. Of course, there are other more worldly teaching goals as well, but fortunately and controversially, I do not think they conflict with developing an appreciation of the life of the mind.

[3] Baigent, N. (2013), 'Total Violence', *Mimeo*

(iii) My mentoring grew out of my own great good luck to have been mentored by the late Paramesh Ray, who had taught me as an undergraduate at University College London and became one of my closest friends. Several times I came very close to failing to reach my academic goals for reasons that were largely not to do with ability. It was John Spraos who accepted me into the undergraduate program at University College, London. He later told me that he viewed me as quite a considerable gamble. Then, it was mainly Paramesh as friend and mentor who was responsible for me getting to the point at which I had the option of working as an academic.

To do good research one must have certain sorts of attitudes and values. Some lucky ones get these as a result of coming from an academic family and may be quite unaware that they have anything remarkable that some less fortunate colleagues do not have. Some of these others are at good universities with vibrant research communities and their socialization gives them these essential attitudes and values. But, especially for research in theory, there are universities where students have the ability to do research, but are at risk of failing because they lack appropriate attitudes and values. This is unfair and socially wasteful. My mentoring has largely been aimed to counter this. In case this sounds a bit lofty, I have been lucky enough to like nearly all of my supervisees and enjoy spending time with them in an informal way and perhaps what you think of as mentoring is not more than enjoying chatting over a coffee, lunch or a glass of something.

CB & MT & YX: You mention attitudes and values. What are they and can you say some more about them?

NB: *I already gave one example in my answer to question 2. In my experience, if a student claims to "know" what they want to work on before starting research and not having done "Fields" or something similar, supervision and mentoring coincide to change this attitude. It can be challenging if this attitude is held intransigently. But there are many other destructive attitudes and values with which I have struggled in supervision and mentoring.*

I was struck by the difference in what I was asked by supervisees in Austria compared with the USA. Not only did students there ask a lot more questions, but what they expected of me in general and in responding to questions in particular, could hardly have been more different from my experience elsewhere.

To illustrate this, I remember receiving an email from a student asking something while a colleague from the USA was visiting. My American colleague was extremely shocked at the question and said: "They should not ask you that; they should figure that out themselves." There were many other attitudes that made supervision require an enormous additional amount of mentoring time and effort simply in order to reach the starting point for serious research.

The point here is that in programs at good research universities, students know how to relate to their supervisors and there is little need for mentoring. How? Largely from socialization. In settings where this does not happen, my experience has been that students ask questions for which it is not in their interest to ask, given their objectives. However, because of the absence of role models and reference points, it can be very difficult for them to see it. Indeed, some of my most difficult

encounters with graduate students in universities that lack a genuine research ethos have involved issues such as this.
I have recommended students/supervisees in these situations read "Zen in the Art of Archery" by Eugen Herrigel (Penguin, 2004), though I never received any feedback about whether anyone did actually read the book and, if so, if it was helpful. It deals with the twin themes of accepting initial ignorance without a framework of attitudes and values that could help escape from this ignorance. What follows from this is the necessity of something like blind trust in whoever the student has chosen to lead them towards enlightenment.
Another attitude concerns spreading risks. At some point in acquiring the tools to do research, particularly in the early stages of research, most students need to spend huge amounts of time on it and focus exclusively on it. If you are in a good doctoral program, you will soon realize this, largely from socialization. But if you are not so well placed, and your supervisor and/or mentor suggest that you completely prioritize research focused activity, why should you believe them? Perhaps all your fellow students are dividing their time and effort among several different things and simply do not understand the wisdom of putting all your eggs in one basket. Why should you accept your supervisor's advice if all around you suggests otherwise? But the fact is that in such settings, if you do not absolutely prioritize research pretty much exclusively, you are most unlikely to succeed.
Since I had little formal or regular supervision myself, I know this all too well having learned it the hard way. However, I did have one advantage over students with whom I have worked in recent years. In my early years my academic environment was always outward looking and I understood that my research needed to be respected by an international peer group. In later years however, I also experienced life in a university that seemed very inward looking to me. At all levels, from the top down, the greatest concern was with what colleagues at the same university thought and not with what an international peer group thought. Such attitudes make it very difficult for students and teachers alike to succeed at research.
CB & MT & YX: Could you elaborate on what makes a good research climate and whether there are big differences in the US and in Europe according to your experience?
NB: *Of course, there are very good doctoral programs in Europe, though it seems to me that most of the ones I know are American style programs. There are also places in Europe with a great research climate – the Choice Group at the London School of Economics where I am happily located is certainly one of the best I have encountered. As to what makes for a good research climate, it would take many pages, which I assume I do not have, to comment non-trivially on this. Remember that universities with good research climates have developed over a long period, and attempts to do it quickly even if well-funded, have not always succeeded. It is a very difficult thing to achieve a good research climate by design. Also, good research climates are obvious when you are in one, but difficult to describe in detail.*
However, here is a quick list of what I think is important: Either an understanding or at least an acceptance of the priority of research in university administrations; absence of the influence of rank or hierarchy in day to day communication; a lot of

informal interaction – coffee breaks, lunch, bar and dinner. It strikes me that in the best research communities that I have experienced, discussion and comment is not restrained by any form of social or even professional hierarchy or by wondering if one will be judged in future by what one has said in the past. The real question is what can be done to ensure a good research atmosphere? But this is not the same question.

Let me give an example involving Yongsheng, with his permission. I remember him coming into my office at Tulane one day and he seemed very excited about some point he had been thinking about. I don't remember exactly what it was about, but I remember that I did not share his view. Let us say that he expressed his opposition to my arguments with more vigor than is usual, before storming out of my office. I remember feeling very pleased at the intensity with which he was engaging in social choice theory and when he returned ten minutes later feeling the need to apologize, I assured him that no apology was necessary and, in fact, I was delighted. Now, there are many places where students would never express themselves in this way in case the listener imposed a sanction at a later date. Their reticence, while understandable, is not conducive to a good research atmosphere.

CB & MT & YX: Turning to your own research, looking back what 'made you' a social choice theorist? What fascinates you about social choice theory?

NB: *The first key event in this process was being thrown out of Woking County Grammar School where the Head Teacher had told me that in wanting to study and eventually do research in economics, I was aiming way too high. (I had given little reason to any teachers at the school to think otherwise and, in any case, the school did not offer economics as a subject.) So instead of completing the final two years of school, I was immensely lucky to have attended what is one of the most wonderful educational institutions I have ever experienced, now known as Guildford College. It was there that I took my first course in economics superbly taught by a young graduate fresh from the London School of Economics, Mr. Barret, and he taught in a fairly analytical way which appealed to me. I also studied the British Constitution and Law which led to a serious interest in Political Philosophy and Jurisprudence. I also taught myself some mathematics before arriving at university. At University College London, I did not encounter social choice theory until studying for my M.Sc.(Econ.), where again I was taught by two, very different, but exceptionally inspiring teachers, Maurice McManus and Christopher Archibald. Both included Arrow's theorem in their teaching and both had very different views about it.*

But if I had to single out one key encounter, it was being taught and then mentored by Paramesh Ray. As I was struggling to find my feet in research in the late 1960's and early 1970's, Paramesh was working on what I still regard as one of the nicest papers in social choice theory. It is on the Independence of Irrelevant Alternatives and was published in Econometrica, 1973.[4] *Sadly, the confusion that he addressed definitively can still be found all too frequently. While I had heard Amartya Sen give papers on Benefit-Cost Analysis (at Nuffield College Oxford) and on the Internal*

[4]Paramesh Ray (1973), 'Independence of Irrelevant Alternatives', *Econometrica*, 41(5), 987–991.

Rate of Return (at UCL), it was through Paramesh that I first met him. This was a half hour session in his office at the LSE on the economics of the performing arts. As I was leaving, I remarked that I had read most of his 1970 book and liked it very much. Like many at the time, this was a huge influence in attracting me to Social Choice Theory. After this, I pretty much switched my efforts from the economics of the performing arts to social choice theory and by the time of the famous Caen conference that led to the founding of the journal, "Social Choice and Welfare", I was well and truly hooked – and I still am.

I find it very hard to say what exactly fascinated me about social choice theory. I rather think that it was not social choice theory itself that fascinated me, at least in my early research. Rather what fascinated me was that I might be able to make a point in any area of economics, and it happened to be social choice theory.

CB & MT & YX: You mentioned that you switched your interests from the economics of performing arts to social choice theory. Can you elaborate a bit more on this "switch" of your research interest? Was there any particular "click" that pulled the trigger?

NB: *There are several issues in the Economics of the Arts, particularly in Arts Education, that I think are very much social choice issues. For example, for some issues in the Economics of the Arts, assuming that preferences are independent of social outcomes does not make much sense, and this is probably true more generally in the Economics of Education. While supervising Benjamin Lane, Ben and I talked a lot about doing a paper on the Economics of the Arts without the usual assumptions about preferences, or indeed without any role for preferences. I still think about this issue.*

Thus, moving from the Economics of the Arts to social choice theory was a very small step, or so it seemed and I retain an interest in the Economics of the Arts.

So, I did not switch my interests in any major way. I already listed key events and encounters along my path to social choice theory. The real point is that in the very short list of areas in which I might have attempted to do research, social choice theory was the first where I thought I could make a point. Fascination is not a sufficient reason to write in an area – you must have a point to make and, by this, I don't just mean a theorem you can prove. For example, in topological social choice, my main and only real point was that is it not clear what point is made in that literature.

CB & MT & YX: What were, do you think, the great achievements of social choice theory (apart from typical examples such as Arrow's Impossibility Theorem)? What is the contribution of social choice theory today? What do you think are the main questions of the new generation of (a) social choice theorists, (b) welfare economists?

NB: *First, I do not see any point in trying to distinguish between social choice theory and welfare economics and I suspect it is impossible to do so coherently. Perhaps it can be argued that social choice theory addresses foundational issues of welfare economics, though that is certainly not all that social choice theory does and not all social choice theory does that.*

In asking about the contribution of social choice theory today, I take it that you intend to ask about current and perhaps recent work. If so, forgive me for not attempting an answer. The reason is that it really takes some time to appreciate what is significant and what is not. I remember Amartya Sen saying, in his keynote lecture at the Social Choice and Welfare Meetings in Caen, that it took 20 years before a clear understanding was reached of what was important about Arrow's theorem – it took me slightly longer than this! So, if it does take this long to appreciate significance, as I think it does, I would rather not attempt to answer this question about current research. Contrast the literature of the 1950's and 1960's with the Golden Age of social choice theory that flourished in the 1970's.

By the way, I have mainly worked on issues 20 years or so after they were first raised.

You give Arrow's theorem as one of the obvious great achievements of Social Choice Theory and I would say that its achievement is two-fold. Not only is the result itself profound in a formal sense. It also changed the way we think about social welfare and welfare judgments. It also changed the way we think about Utilitarianism in political philosophy and the way we think about consensus and national interest in political science. The same could be said of the following theorems: Sen's theorem on the Impossibility of Unanimity and Individual Rights and the Theorems of Gibbard and Hurwicz. All of these contributions were seminal.

CB & MT & YX: One needs a lot of courage to treat a question that has been discussed in the literature 20 years ago! So would you say that you are encouraging students and researchers not to be afraid to tackle "old" questions if they think they have something to say about it?

NB: *Absolutely. Research must make a point and its vintage is less important than its substance. Of course, it may be harder to motivate and interest readers in "old" points, but if it is a good point then it is possible.*

CB & MT & YX: These days, one has very much the impression that one has to be "up to date" with methods and questions and often research may rather appear as a competition against time with respect to who is first out with a new variant of the dictator game experiment or the variation of a parameter in a model. Actually, many complain that courses such as the history of economic thought are not taught any more – again, it is our impression that even as a doctoral student, one often does not have any longer the time to look into historical and conceptual developments. Would you agree with this picture about current (job-market) pressures and their impact on the questions (especially young) scholars work on? If so, what would be your "piece of advice" to young researchers?

NB: *I really cannot say anything about current job market pressures since I have already been out of that market for too long to comment. But I do remember similar concerns being expressed very many years ago. From the luxury of retirement my view again starts from the priority of making a good point in research. Of course, research work should consider whether or not that point can be better made using the latest fads and fashions, but it would be unwise to contrive some way to include them in a paper. If there is a reasonable expectation that the latest approach or*

method would be present in a paper, then the introduction can always explain why it is not.

After many journals "moved against" social choice theory in the late 1970's, it was widely thought that it would be very difficult to publish social choice theory papers. Indeed, I remember some distinguished social choice theorists either refusing to supervise doctoral students in social choice theory or at least issuing dire warnings. I was one of the latter, as some of you may remember.

I say this after having swum against the current myself more than once. In choosing a dissertation area at UCL, John Spraos rightly advised against highly theoretical work in view of the risks. I also worked on Merit Wants at a time when the subject was much less open to such ideas than it is now. I must say though, in my case it would have been completely inconsistent with my reasons for becoming an academic to avoid making what I thought were reasonable research points for job market reasons. But I was unemployed for one year!

Finally, I do regret the lack of opportunities for students to study the history of economic thought. I was extremely fortunate that history of economic thought (Aristotle to Keynes, if I remember correctly) was a compulsory subject at UCL, wonderfully taught by Marion Bowley. Another compulsory course was the "Development of Economic Analysis", which began a mere 150 or so years ago, and gave an excellent sense of how research develops. It is striking to me that in many Philosophy Departments, this is the standard way to teach the subject.

CB & MT & YX: Let's turn to another research area of yours. In some of your work on individual choice theory you argued for the need to look behind the "veil of preferences" and questioned the acceptance of consistency axioms (such as condition Alpha) employed in rational choice theory on their face value. Could you tell us a bit more about what led you to do research in these fields and what you considered to be important about these questions?

NB: *Several episodes in my life, as well as reading the literature outside of economics, nudged me towards questioning whether a theory of rational choice could simply start with such "given preferences" and proceed to behavior via optimization. I will limit myself to just two examples of such episodes, though many more could be given.*

The first arises from the fact that my school education gave me no appreciation of classical music, literature, art, or poetry and I eventually resented that. So in my early 20's I tried to develop a preference for classical music and this made, and continues to make, huge contributions to my enjoyment of life. Later, I did the same for painting with similar results. My efforts at developing a preference for literature and poetry are still very incomplete, though I am still trying.

The other episode was my experience of psychotherapy in my mid 20's. I had "group therapy" for 4 years followed later by 2 years of individual therapy. Both were successful in raising my long term well-being, however that is defined. But how do I view the decision to enter these therapies? Was it a decision to change the set of outcomes to include ones that were higher in my ranking than those present without therapy, and then to choose one of them? I certainly deliberated long and hard about the decision, but neither ex ante nor ex post did the decision seem

to be rationalizable in this way and one of the therapist's own characterizations made much more sense to me. (She was an eminent therapist at a famous London Teaching hospital.) My decision to enter therapy was a decision to change some mental (emotional) relationships with events and in relation to others, and it was obvious that this would lead to changes in my preferences and choice behavior.

I could give many more examples in which attempting to develop preferences, sometimes in known directions and sometimes in unknowable directions, have been pivotal in changing my life. Were my decisions rational? This question begs more questions than there are answers in the literature. But it is not surprising that my experience of life led me to consider what may be behind the preference veil, and not presume ab initio that whatever is behind that veil can be reduced or summarized by an All Things Considered preference of the kind that is key to the traditional theory of rational choice.

By the way, I also think that a decision to do a dissertation in choice theory shares some key characteristics with a decision to undertake therapy or to enjoy classical music. At least that is the impression I have from many of my supervisees. I can remember warning prospective supervisees about this

So, key episodes in my life that had lifelong effects played a role in disposing me to go behind that "Veil of Preference".

Finally, you ask: ". . . what (I) considered to be important about these questions?" A narrow answer to this question would involve repeating the introduction to my "Veil of Preference" paper which anyone interested could read. But in some wider sense, I cannot offer much at all since it is not at all clear where the literature will lead. My hunch, for what little it may be worth, is that it may make a difference both to our understanding of choice behavior and welfare judgments. But this is only a hunch.

I should add that I have been very lucky to have so many friends, many of whom are authors of papers in this volume, whose views differ from mine in varying degrees, particularly Prasanta Pattanaik, Remzi Sanver, Kotaro Suzumura and Yongsheng Xu. Of course, an immense debt in this area is owed to Amartya Sen.

CB & MT & YX: What do you think about the criticism of rational choice theory today in the light of experimental and behavioral evidence? How do you think rational choice theory should be seen? Normatively, in the sense of what we mean when we say that a person acts rationally? Or is it an exploration of the implications of the empirically testable hypothesis?

NB: *I am not sure what criticisms you have in mind. So just let me offer a few remarks. Knowing what people actually choose and even what they would choose is neither necessary nor sufficient for determining the rationality of choices. However, experimental and behavioral economics have both contributed to a questioning of the concept of rationality which I welcome. You ask how I think rational choice theory should be seen. I am clear about how I wish it not to be seen. I would not like it to be seen as limited to preference maximization, especially not in its revealed preference form. Also I would not like it seen as exclusively normative or as exclusively an exploration of testable implications.*

After all, rational choice theory in economics is a tool, though in other areas it plays a more fundamental role. Mark Johnson once offered a view about technical progress using the example of a sword. In the early days of swords, they could be used for clearing a path through a forest, cutting wood for fire, killing for food, defense or pillage and personal grooming in the form of cutting one's hair and shaving. Over time, specialized equipment was developed for each of these activities so that, for example, you can shave with a safety razor but it is not very good for anything else.

Mark's observation suggests that it should not be surprising if standard rational choice theory is superseded by advances in experimental and other behavioral approaches. I must say I have always been very skeptical that standard rational choice theory will ever provide a satisfactory general explanation or understanding of actual behavior. However, my answer to your previous question reveals my view that rational choice theory should not be limited to standard rational choice theory and greater success in explaining behavior cannot yet be ruled out.

CB & MT & YX: Following the current economic crises, the standard economics curriculum has been criticized in many different universities. What would you change about the classical economics curriculum as it is taught today?

NB: *I think the economics curriculum should be determined mainly by research. This is because I see teaching as communicating research after much filtering and evaluation by the research community. Of course if recent events lead to research developments which eventually find their way into the curriculum, that is fine. This is how environmental economics became a standard option in the curriculum as well as a standard example of market failure in required courses in microeconomics.*

So I think that curricula development is best led by research. This view is derived from a view about what universities are, and should be, for. Actually, I am saddened that many developments in universities often fail to be justified in terms of a clearer view about their main purpose. It seems to me that research, which itself requires some careful definition, is constitutive of the concept of university, and is therefore one of its defining characteristics. Otherwise, it would be difficult to differentiate them from all sorts of other institutions that are clearly not universities. This view of what constitutes a university has clear implications for the role of research and its implications for curriculum development.

There are many who, as you say, argue that current events should enter curricula in economics before much or any research has been done. I wonder what they think the purpose of a university is and how they think universities differ from other institutions?

On newer areas of research such as experimental and behavioral economics, I see no reason not to offer them as options and their eventual success with the research communities will rightly determine whether or not they become required courses.

CB & MT & YX: The role of economics in public debate on and public formation of economic policies has been rather controversial. What's your view on the connection between economics and public policy in general, and welfare economics and public policy in particular?

NB: *First of all, I do not think it is a requirement of research that it directly informs public policy formation or debate, at least in any direct way. I certainly do not think there are any "results" in economic theory that are "directly usable". Rather, I think the starting point for thinking about this is what economic theory and welfare economics is best at doing, something that I have already spelled out together with a striking implication for commodity tax structure in my answer to question three. But I also think that economic theorists could be helpful in establishing more conceptual clarity. For example, the OECD manual for Benefit-Cost Analysis by Little and Mirless is explicit that what they offer is derivable from a Utilitarian Social Welfare Function. I would like to see some practical principles of project appraisal derived from other ethical criteria. Indeed, my new interest in aggregate violence provides another example for which I find myself increasingly interacting with people close to the practical edge of policy and implementation. In this area, it is not at all clear what is meant by or even could be meant by claims that the level of violence has increased or decreased or would increase or decrease as a result of some policy intervention.*

CB & MT & YX: Nick, it would be great to go on asking you questions, but of course we must stop at some point. We thank you a lot for this very interesting interview and are looking forward to reading and discussing your work on violence!

NB: *Thank you so much for getting me to reflect on the issues you raise. I am also touched by the generosity of all contributors to this volume. Every contribution contains at least one author, and some several, with whom I have enjoyed the warmest friendship and intellectual interaction over many years. My life has been hugely richer as a result and I am very grateful.*