



ERASMUS JOURNAL FOR PHILOSOPHY AND ECONOMICS
VOLUME 14, ISSUE 2, WINTER 2021

The Erasmus Journal for Philosophy and Economics (EJPE) is a peer-reviewed bi-annual academic journal supported by the Erasmus Institute for Philosophy and Economics, Erasmus School of Philosophy, Erasmus University Rotterdam. EJPE publishes research on the methodology of economics, history of economic thought, ethics and economics, and the conceptual analysis of inter-disciplinary work relating economics to other fields. EJPE is an open-access journal. For additional information, see our website: <http://ejpe.org>.

EDITORS

Måns Abrahamson
Annalisa Costella
Savriël Dillingh
Akshath Jitendranath
Marina Uzunova
Erica Celine Yu

ADVISORY BOARD

Erik Angner, Roger Backhouse, Constanze Binder, Marcel Boumans,
Richard Bradley, Matthew Braham, Nancy D. Cartwright, Christopher
Clarke, David Colander, John B. Davis, Sheila C. Dow, Francesco Guala,
Till Grüne-Yanoff, D. Wade Hands, Martin van Hees, Conrad Heilmann,
Frank Hindriks, Geoffrey Hodgson, Elias Khalil, Arjo Klamer,
Alessandro Lanteri, Aki Lehtinen, Uskali Mäki, Caterina Marchionni,
Deirdre N. McCloskey, Mozaffar Qizilbash, Julian Reiss, Ingrid Robeyns,
Olivier Roy, Malcolm Rutherford, Margaret Schabas, Eric Schliesser,
Esther-Mirjam Sent, Robert Sugden, Alex Voorhoeve,
Jack Vromen, Nicholas Vrousalis.

EDITORIAL ASSISTANTS

James Grayot, Bronagh Dunne

ERASMUS JOURNAL FOR PHILOSOPHY AND ECONOMICS
VOLUME 14, ISSUE 2, WINTER 2021

TABLE OF CONTENTS

ARTICLES

- Choosing Less over More Money:
The Love of Praiseworthiness and the
Dread of Blameworthiness in One-Player Games
NINA SERDAREVIC [pp. 1-24]
- Social Contract, Extended Goodness, and Moral Disagreement
CYRIL HÉDOIN [pp. 25-52]
- Integrated Moral Agency and the
Practical Phenomenon of Moral Diversity
MICHAEL MOEHLER [pp. 53-76]

ARTICLE SYMPOSIUM on “Narrow Identities”

- The Paths to Narrow Identities
JEAN-PAUL CARVALHO [pp. 77-86]
- Deepening and Widening Social Identity Analysis in Economics
JOHN B. DAVIS [pp. 87-98]
- Social Identities: Narrow and Broad,
Exclusive and Inclusive, Firm and Fuzzy
PETER FINKE [pp. 99-105]
- Group Membership or Identity?
MIRIAM TESCHL [pp. 106-114]
- Narrow Identities Revisited
PARTHA DASGUPTA AND SANJEEV GOYAL [pp. 115-122]

INTERVIEW

- Grounding Equal Freedom:
An Interview with *IAN CARTER* [pp. 123-156]

CRITICAL COMMENTS

- Can One Both Contribute to and Benefit from Herd Immunity?
LUCIE WHITE [pp. 157-164]
- Vaccine Refusal Is Still Not Free Riding: A Reply
ETHAN BRADLEY AND MARK NAVIN [pp. 165-169]
- The Different Facets of Injustice:
A Critique of Nancy Folbre's 'Manifold Exploitations'
VIVEK CHIBBER AND ROBERTO VENEZIANI [pp. 170-184]

BOOK REVIEWS

- Review of Till Düppe and Ivan Boldyrev's (eds.)
Economic Knowledge in Socialism, 1945-89
MARTA PODEMSKA-MIKLUCH [pp. 185-190]
- Michel S. Zouboulakis' *The Varieties of Economic Rationality:
From Adam Smith to Contemporary
Behavioural and Evolutionary Economics*
YAM MAAYAN [pp. 191-195]

PHD THESIS SUMMARIES

- A Tale Between Finance and Economics:
Four Essays on the History and Methodology
of the Efficient Market Hypothesis
THOMAS DELCEY [pp. 196-201]
- Otto Neurath and Ludwig von Mises:
Philosophy, Politics, and Economics in
Viennese Late Enlightenment
ALEXANDER LINSBICHLER [pp. 202-208]

Choosing Less over More Money: The Love of Praiseworthiness and the Dread of Blameworthiness in One-Player Games

NINA SERDAREVIC

Norwegian School of Economics

Abstract: Why choose less money over more when no one is watching? A central tenet of economics is that this behaviour can be explained by intrinsic motivation. But what does intrinsic motivation entail? What encourages it? This paper answers these questions through a Smithian lens: moral motivation includes not only a naturally strong love of praise and dread of blame but also a natural, and stronger, love of being worthy of praise and dread of being worthy of blame, even if neither is necessarily given. I rely on quantitative and qualitative data from economic experiments to illustrate this claim. While the current scholarship on Smith has applied his theory to situations in which our actions either evoke reactions from others or have monetary consequences for them, I extend his insights to receiver games (Tjøtta 2019) and dice-rolling games (Fischbacher and Föllmi-Heusi 2013) aimed at eliciting self-regarding concerns, that is, actions affecting the interests of only ourselves. I argue that these games accentuate the strength of the love of praiseworthiness in guiding behaviour, emphasising its immediate reference to others and foundation in intentions along with outcomes.

Keywords: experiments, moral judgement, non-strategic games, incentives

JEL Classification: B12, B15

I. INTRODUCTION

Maria Pia Paganelli ends the chapter “Smithian Answers to Some Experimental Puzzles” with an observation and an encouragement: “Adam

AUTHOR’S NOTE: I am grateful to the editors and two anonymous referees for insightful comments and suggestions that have helped improve this paper considerably. I thank participants of the International Adam Smith Society at Chapman University in Los Angeles (2019), University of Wisconsin-Madison (2021), Maria Pia Paganelli, Jo Thori Lind, and seminar participants at the University of Bergen for useful comments. I thank Sigve Tjøtta for valuable suggestions and for generously sharing the experimental data with me.

Smith is increasingly being read by experimental, behavioural and neuro-economists. He still has a lot to offer all of us” (2009, 22). I agree. So far, the Smithian love of praise/praiseworthiness and dread of blame/blameworthiness has been cited to explain behaviour in experimental games such as the ultimatum game, dictator game, trust game, and prisoner’s dilemma game (Ashraf, Camerer, and Loewenstein 2005; Brown 2011; Meardon and Ortmann 1996; Paganelli 2009; Smith and Wilson 2019; Young 2009). A key aim of many experimental games has been to rule out selfish reasons for a variety of other-regarding behaviours, allowing the researcher to elicit subjects’ intrinsic motivations.¹ While the Smithian love of praiseworthiness and dread of blameworthiness could certainly be the strongest intrinsic motives in these settings, distinguishing them from extrinsic motivations, such as the mere love of praise and dread of blame, is empirically and theoretically challenging, if not impossible. With this as a backdrop, one could claim that the love of praiseworthiness and praise are merely two different names for the common dichotomy between extrinsic social motives and intrinsic moral motives—concepts that are borrowed from social psychology and applied to economics (Bénabou and Tirole 2003; Bénabou and Tirole 2006; Frey and Oberholzer-Gee 1997; Gneezy and Rustichini 2000; Scitovsky 1976). This claim would not be entirely wrong.

At first, the interwoven relationship of our moral motivations accords with Smith’s own discussion in *The Theory of Moral Sentiments* (1759).² We desire not only to act according to what results in actual praise and avoids blame from others but also have a natural love of being worthy of praise and fear of being worthy of blame, even if neither can be given. Smith rightfully notes that these two principles “resemble one another” and are “often blended with one another” (*TMS*, III.ii.2, 114). But there exist two palpable differences between Smith’s account and how economists usually view motivational concepts: Smith’s argument is not based on a dichotomy, either when it comes to the substance of the love of praise and praiseworthiness, or what makes a praiseworthy character. Asserting that “in every well-formed mind this second desire [the love of

¹ Remic (2021) offers an important and interesting discussion of the concept and definitions of intrinsic motivation and how they have been used by economists, emphasising the challenge of importing competing psychological theories of intrinsic motivations into economics.

² This and all subsequent references to *The Theory of Moral Sentiments*, abbreviated as ‘*TMS*’, will be to the Glasgow edition (Smith [1759] 1982). References include, in this order, part, section (if applicable, in lowercase roman numerals), chapter and paragraph (both in Arabic numerals).

praiseworthiness] seems to be the strongest of the two" (*TMS*, III.ii.7, 117) did not prevent him from observing a necessary interdependence between the two loves. Our intrinsic moral motivations do not exist in a vacuum, so to speak. Our love for being worthy of praise and fear of being worthy of blame are influenced by a variety of external factors interacting in concert. What is more, the love of praiseworthiness and dread of blameworthiness have qualities that transcend (monetary) outcomes. In addition to pleasurable and less pleasurable outcomes, intentions are what manifest respect in others and, in turn, ourselves.

Taking all of the above together, regardless of how hard the experimentalist tries to create a non-social situation, it will necessarily entail a reference to others.³ Acknowledging the influence of this external component when discussing intrinsic motivations is significant in order to explain how we have learned to become aware of undeserved praise and incapable of being truly satisfied with it. But are there some decision environments that evoke such redirected judgements more than others, emphasising the strength of the socially constructed love and dread in guiding behaviour? As Paganelli (2009) points out, compared to the ultimatum game, the dictator game is a good candidate to elicit such judgements. However, while this game is non-strategic in the game-theoretical sense where actual praise and blame cannot occur from another person, it does not give rise to entirely self-directed moral judgements, as one's actions do in fact have monetary consequences for another person. A growing literature has shown that, the fact that the dictator's decision affects others, is sufficient to compel subjects to be other-regarding and restrict narrow self-interest (Cappelen et al. 2017; Dana, Weber, and Kuang 2007; Krupka and Weber 2013). Behaviour in the dictator game does not, so the argument goes, necessarily reflect solely intrinsic motivation but may also be driven by extrinsic motivation.

In this paper, I extend the application of Smith's theory of moral motivations to receiver games (Tjøtta 2019) and dice-rolling games (Fischbacher and Föllmi-Heusi 2013). Importantly, I do not want to suggest that these games isolate the love of praiseworthiness and dread of

³ In social science, a social relation or social interaction is any relationship between two or more individuals. See, for instance, Rummel (1976) for a comprehensive discussion of social interaction and behaviour. In game theory, a game usually consists of at least two players where one player's payoff is contingent on the strategy implemented by the other player. In one-player games, such as the receiver and dice-rolling games, this strategic component is absent, and subjects make a decision that will only affect themselves in terms of material payoff.

blameworthiness. Rather, I argue that they emphasise how socially rooted our intrinsic motivations really are. The features of one-player games, as seen from the perspective of economics and game theory, do not explicitly involve external rewards or costs. The setting is very simple: subjects are asked to choose between receiving more or less money. The desire for praise (positive payoff) and the fear of blame (negative payoff) from others is absent by experimental design. Moreover, actions have monetary consequences only for subjects themselves, as opposed to the ultimatum and dictator games, in which subjects decide on payoff allocations affecting both themselves and another person (Berg, Dickhaut, and McCabe 1995; Forsythe et al. 1994; Güth and Tietz 1990). The results show that even in these seemingly non-social situations, subjects commonly choose less money over more (Abeler, Becker, and Falk 2014; Tjøtta 2019; Utikal and Fischbacher 2013). In many ways, these games resemble Smith's notion of a "solitary place" (*TMS*, III.1.3, 110) in which individuals lack the social mirror of what are the objects of praise and blame, allowing them to fully endorse their self-love, often understood in economics to mean choosing more money over less. But what these games ultimately enable is the elicitation of Smith's self-directed process of moral judgement. An internal assessment of praiseworthiness and blameworthiness, the habit experimental subjects have acquired in redirecting judgments toward their own conscience—to the "well-formed mind" (*TMS*, III.2.7, 117)—promoting behaviour consistent with a love of praiseworthiness that, in Smith's world, has social roots.

To help flesh out my arguments, I pay particular attention to the receiver game due to its attractive design encompassing both quantitative and qualitative data. In doing so, I extend Tjøtta's (2019) discussion of the relevance and importance of Smithian insights for this class of games. I argue that an application of Smith's theory to these games adds at least two points of reflection regarding how we theoretically model human nature and how we empirically interpret it. First, through interacting with others and seeking praise and avoiding blame from the "man without", we gradually learn, through experience, how to turn the lens inward to the "man within" (*TMS*, III.2.32, 130). Thus, what experimentalists refer to as intrinsic motivations to explain why subjects choose less money bear necessary connections to the external world in Smith's model. It is our desire to be approved of by others that allows us to learn to view ourselves from without, which in turn lets us see the difference between something being praised and something being worthy of praise. Second,

the determination of what actions are praised and praiseworthy or blamed and blameworthy cannot be inferred only through analysing monetary outcomes. Smith warns us of making such shortcuts and offers a more nuanced picture; people are neither altruistic saints nor self-interested sinners. With their sociality comes the importance of intentions and deservingness, in addition to actions and outcomes. This means that choosing less and choosing more money in an economic experiment may indeed be different in the monetary outcome space, but these two actions need not be based on entirely different motivations—both may be understood as being encouraged by a love of praiseworthiness or a dread of blameworthiness.

To explain subjects' motivations and why they sometimes choose less money over more even in one-player games, the remedy in economics has typically been to alternate preference formulations: one merely re-specifies the utility function to include different other-regarding motivations or to reflect an intrinsic preference for less money. However, behaviour in these games cannot be explained by social preference models, as decisions lack explicit consequences for other experimental subjects—there is not another utility function to take into account. We are left with explanations advancing intrinsic motivation and built-in aversion concepts (Bénabou and Tirole 2003; Dufwenberg and Dufwenberg 2018; Kajackaite and Gneezy 2017; Romaniuc 2017). Such ad hoc conceptualisations not only violate one of the core economic assumptions of payoff-maximising agents but also contribute to explaining “all apparent contradictions” that Nobel laureate Gary Becker (1976, 5) warned about. Becker argued that “the assumption of stable preferences [...] prevents the analyst from succumbing to the temptation of simply postulating the required shift in preferences to ‘explain’ all apparent contradictions to his predictions” (5).

In what follows, I present the ultimatum game and dictator game, showing how Smith's theory has been applied to these games by the scholarship applying his insights to economic decision-making. I proceed to introduce the receiver game (Tjøtta, 2019) and the dice-rolling game (Fischbacher and Föllmi-Heusi 2013). To lay the foundation for further discussion of the role of the love of praiseworthiness and the dread of blameworthiness in one-player games, I analyse answers to open-ended questions from Tjøtta's (2019) receiver game. The results from the qualitative analysis strengthen Smith's claim that the love of praiseworthiness symbolises the love of approval from our *socially* acquired character. Importantly, this love extends beyond monetary outcomes: the role of

character, deservingness, and other-regarding concerns are motivations reported not only by subjects who choose less money but also by a substantial minority of those who choose more. The paper ends with concluding remarks.

II. EXISTING APPLICATIONS OF SMITH'S MORAL THEORY

II.I. *The Ultimatum Game*

In the ultimatum game, introduced by Güth, Schmittberger, and Schwarze (1982), an anonymous person named the *proposer* (Person A) is endowed with an amount of money (\$10) and has to decide how much to keep. What is not kept is offered to an anonymous *responder* (Person B). The responder has to either accept or reject this offer. If the responder accepts the offer, the money is shared according to the proposer's initial offer (x , $10-x$). However, if the responder rejects the offer, both players receive nothing (0, 0), as illustrated in Figure 1. The prediction is, according to game theory, a sub-game perfect Nash equilibrium in which the proposer gives as little as possible to the responder and the responder accepts any positive amount.

The results of the ultimatum game show that individuals participating in these experimental games frequently violate the equilibrium predictions (Camerer and Thaler 1995; Güth and Tietz 1990; Roth et al. 1991). The proposers offer approximately 40% of their endowment, whereas about half of the responders reject what they perceive to be unfair offers in which they would receive less than 30% of the total sum (Tisserand 2014). Hence, neither the proposer nor the responder acts to maximise material self-interest.

The scholarship on Smith has offered additional points of view that go beyond monetary outcomes, thereby enriching the decision process presented in Figure 1 (Paganelli 2009; Smith and Wilson 2018). In the Smithian sense, outcomes are secondary to the conduct governing actions; we judge others, and we know others are judging us. When determining the praise and blame of an action, we focus on why the action occurred in the first place or, as Smith puts it, the "sentiment or affection of the heart from which any action proceeds" (*TMS*, I.i.3.5, 18). We judge an action or reaction according to the cause that gave occasion to it and the consequences it produces.

Judgements that focus on causes are what Smith terms judgments of propriety and impropriety. After entering the actor's situation, we judge whether actions and reactions are appropriate to their circumstances.

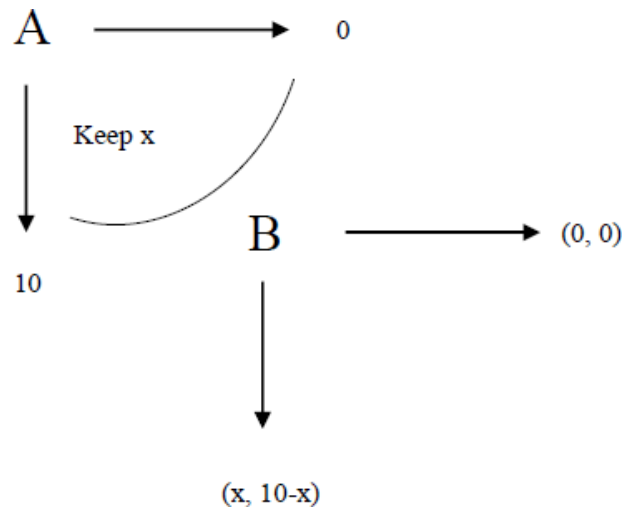


Figure 1: The Ultimatum Game

Judgments that focus on consequences are termed judgements of merit and demerit; one enters the situation of those who benefit (or are harmed) by that action, judging whether beneficial or harmful effects are proper in evoking either gratitude or resentment. Smith further maintains that we rule out the reactions of those who have a personal stake in what is happening, as that would influence moral judgement. Here, the impartial spectator enters the scene, constituting the conscience and setting an impartial (as possible) standard for what is generally deemed worthy of approval and disapproval.

To explain the responder's behaviour (Person B), the scholarship points to the relevance of reciprocity, both positive and negative (Hoffman, McCabe, and Smith 2008; Paganelli 2009; Young 2009). Positive reciprocity is present when someone reciprocates a cooperative action with gratitude, or in the economic sense, a positive monetary payoff. Negative reciprocity, or what Smith refers to as resentment, is the responder's and impartial spectator's willingness to punish non-cooperation in social exchange with a negative payoff. The responder may be willing to forego whatever was offered by Person A out of resentment because "to us, [...] that action must as surely appear to deserve punishment, which every body who hears of it is angry with, and upon that account rejoices to see punished" (*TMS*, II.i.2.4, 70).

Now, turning to the proposer's behaviour (Person A), it is plausible that they make few low offers because of their habit of seeking common ground with others. Motivated by the incentive of being granted praise and praiseworthiness and avoiding blame and blameworthiness, by both person B and the impartial spectator, the proposer uses the expected

approval or disapproval as an indicator of whether sentiments are appropriate to their causes and the merit and demerit of the consequences produced. Smith explains that “we are pleased when they approve of our figure, and are disobliged when they seem to be disgusted. We become anxious to know how far our appearance deserves either their blame or approbation” (*TMS*, III.i.4, 111). In order to achieve mutual agreement, Person A moderates behaviour according to standards that are expected to constitute appropriate behaviour and what “would be our own [conception], if we were in his case” (*TMS*, I.i.2, 9).

II.II. The Dictator Game

To probe further into what really motivates the proposer’s behaviour, experimental economists compared the ultimatum game to a new game: the dictator game (Forsythe et al. 1994; Kahneman, Knetsch, and Thaler 1986). While the ultimatum game has been an important instrument for eliciting people’s preferences for fairness and reciprocity as seen from the responder’s point of view, it suffers from a strategic confound when it comes to eliciting the genuine motives of the proposer. The act of kindness is strategic if the proposer shares money simply to *appear* generous in order to avoid rejection and blame, leaving both the proposer and recipient with no money.

In the dictator game, the *sender* (Person A) is endowed with \$10 and has to decide how much of this money to keep. The rest goes to an anonymous *recipient* (Person B). As Figure 2 shows, in contrast to the ultimatum game, the recipient of the money in the dictator game cannot reject or accept any offer made by the sender. He or she is a passive receiver who must accept whatever the sender does not keep (x , $10-x$). The actual results of the dictator game differ greatly from these predictions. While the offers are certainly lower than in the ultimatum game, subjects still continue to allocate about 20–25% of their endowment to a random anonymous recipient (see Engel (2011) for a meta-study of dictator games).

Insights from Smith’s moral theory have also been applied to the dictator game. In particular, Paganelli’s (2009) interpretation of Smith leads her to argue that behaviour in this game emphasises that resentment does not come only from actual others, such as Person B, but the impartial “man within the breast” (*TMS*, III.3.24, 146). When actual spectators are absent (or present, but partial and in need of correcting), Person A becomes the impartial spectator of his or her own conduct and scrutinises actions and reactions according to impartial standards of moral

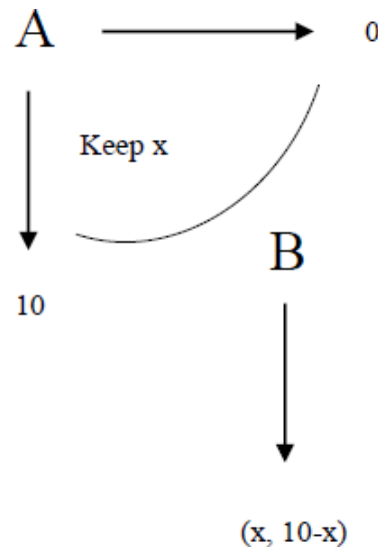


Figure 2: The Dictator Game

judgement. The spectator judges whether actions are proper (improper) to their circumstances and, after entering the situation of Person B, whether actions have merit (demerit)—whether gratitude or resentment is felt towards Person A.

Conceived this way, the behaviour of subjects in the dictator game could be explained by responding to the call of an imagined impartial spectator who “immediately calls to us, that we value ourselves too much and other people too little, and that, by doing so, we render ourselves the proper object of the contempt and indignation of our brethren” (*TMS*, III.3.5, 138). In finalising the discussion of how Smithian insights can be applied to the dictator game, Paganelli (2009) turns her attention to the importance of the moral conscience. She explains that the love of being worthy of praise and dread of being worthy of blame motivates the dictator to do the right thing, because he does not want to be rendered the proper object of hatred in the eyes of his conscience. Positing further that “the fairness observed in the experimental results may indeed have little to do with self-regarding preferences” (Paganelli 2009, 16). Hence, it is not our love of being praised that makes us behave in a praiseworthy manner, nor is it the dread of blame that motivates us to avoid it. Rather, we wish to be the proper object of praise and avoid being the proper object of blame.

III. THE RECEIVER GAME AND THE DICE-ROLLING GAME

I think that the strength and significance of Paganelli’s (2009) argument can be better illuminated by another class of games in which distri-

butional preferences (the allocation of payoffs between *self* and *other*) cannot explain the results. To encourage an entirely self-directed process of moral judgement and tease out the strength of the love of praiseworthiness and the dread of blameworthiness, one would ideally create a decision-situation resembling Smith's notion of a 'solitary place' in which experimental subjects must rely on their socially acquired conscience to make decisions with consequences affecting only themselves.

Taking this into account and stripping the dictator game of everyone except Person A, Tjøtta (2019) conducts a modified version of the dictator game in a series of anonymous receiver games, redirecting judgments of propriety (impropriety) and merit (demerit) even more towards one's conscience. By removing Person B, the decision-maker is *both* the acting agent and the agent acted upon. Thus, there is no other person present to either accept or reject an offer, as in the ultimatum game. Moreover, no one is affected by the actions of Person A, as in the dictator game.

In the receiver game, the decision-maker is asked to determine her own payoff by simply choosing how much to keep of the money that is received for partaking in the experiment, as Figure 3 illustrates. Person A can choose to keep between \$0 and \$10. Assuming that subjects are able to distinguish between situations in which reciprocity may be beneficial and situations in which there are no external benefits or costs for choosing more money over less, they should choose to keep the \$10.

Tjøtta (2019) starts by presenting the results of an experiment in which a substantial minority, 28.6% of 91 participants, decided to receive less money over more. In another experiment, even a majority chose to keep less money over more. In total, the results from seven receiver game experiments conducted with both student and representative samples corroborate this result: on average, one-third of a total of 3,503 individuals who participated in these experiments chose to keep less money instead of more.⁴

The dice-rolling game is a related yet slightly different game in which subjects also determine their own payoff. Person A is asked to roll a die and report a number they want to determine their payoff. In other words, one has the opportunity to misreport the true number and earn more money (Fischbacher and Föllmi-Heusi 2013). This is because the higher the number reported, the more money they choose to receive. Given that

⁴ In a follow-up study, Serdarevic and Tjøtta (2021) show that this finding is robust across five countries: France, Germany, United States, Croatia and the United Kingdom. Approximately 28% of subjects choose to receive less than the payoff-maximising option.

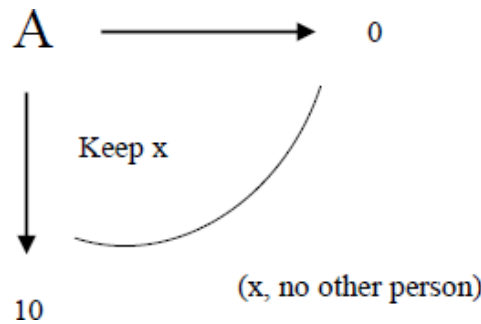


Figure 3: The Receiver Game

subjects in this game are completely anonymous to other subjects and to the experimenter, they are expected to report obtaining a higher number than they actually rolled as a way to maximise their payoff. Here too, there is no Person B, and misreporting cannot be identified at the individual level by the experimenter. Contrary to the economics prediction assuming that subjects will misreport when given the opportunity, a vast literature has shown that subjects do not take advantage of the opportunity to act in a fully self-interested fashion (Abeler, Nosenzo, and Raymond 2019; Gächter and Schulz 2016). In fact, subjects in the dice-rolling game forgo on average about three-quarters of the potential gains. Notably, there are also subjects who on average report numbers lower than they actually obtained, imposing a monetary disadvantage to themselves without improving the payoff of anyone else (Abeler, Becker, and Falk 2014; Utikal and Fischbacher 2013).

IV. THE LOVE OF PRAISEWORTHINESS AND THE DREAD OF BLAME-WORTHINESS

In so far as the experimenter's goal is to elicit subjects' genuine intrinsic motives, this is likely to be achieved by the receiver game and the dice-rolling games. However, to answer *what* this genuine motivation is comprised of and *what* encourages its existence, Smith would ask further leading questions: "What so great happiness as to be beloved, and to know that we deserve to be beloved? What so great misery as to be hated, and to know that we deserve to be hated?" (*TMS*, III.I.6.7, 113).

In asking these questions, Smith does two things. First, he reiterates that a central premise of this theory is that we want to share feelings with the people around us. He reminds us of the standards according to which we can satisfy the "original desire to please [...] our brethren" (*TMS*, III.2.6, 116) in order to be beloved and avoid being hated. Clearly, without the

natural love of praise and blame, we would risk failing to pass judgement on our own conduct as seen through the eyes of others.

Second, in the course of articulating these questions, Smith introduces another desire, expanding the definition of motivation for why we exercise self-command and dampen our self-regarding concerns: the love of praiseworthiness and dread of blameworthiness. Unlike propriety, which is generated by our sympathy and approval with and from others, the love of praiseworthiness exists independently from any actual acknowledgment of it. It provides us with the means to distinguish what is praised from what should be praised, as well as the genuine incentive to *want* to make this distinction in the first place.

In his own words:

Man naturally desires, not only praise, but praise-worthiness; or to be that thing which, though it should be praised by nobody, is, however, the natural and proper object of praise. He dreads, not only blame, but blame-worthiness; or to be that thing which, though it should be blamed by nobody, is, however, the natural and proper object of blame. (*TMS*, III.2.1, 113)

While praise and blame express the actual sentiments with regard to others' and our own conduct, praiseworthiness and blameworthiness express what these sentiments naturally *should* be (Griswold Jr 1999).⁵ Smith's empirical project of what impartial spectators (real or imaginary) will praise and blame is enriched with an additional layer of what should be praised and blamed if they were being better spectators. This love of praiseworthiness represents the natural desire of rendering ourselves the proper objects of praise and gratitude. Even if actual praise or blame is given, it provides, according to Smith, little pleasure if it is not accompanied by praiseworthiness. Unwarranted praise satisfies only "the weakest and most superficial of mankind" (*TMS*, III.ii.7, 117). The highest source of satisfaction comes from acting and reacting in ways we know to be praiseworthy. Tranquillity arises when we know ourselves to be worthy of praise, irrespective of whether it is actually being given:

⁵ Forman-Barzilai (2010, 18) suggests that one of the objectives of Smith's constant revisions of *The Theory of Moral Sentiments* was "to assert the independence of conscience" of external influences; the independence of the impartial spectator secures that there are no biases in his or her moral judgment. Müller (1993, 100) agrees and argues that Smith presents a "a theory of the development of conscience through internalization of social norms, as well as a theory of how the morally developed individual is able to ascend from moral conformity to moral autonomy".

The jurisdiction of the man without, is founded altogether in the desire of actual praise, and in the aversion to actual blame. The jurisdiction of the man within, is founded altogether in the desire of praise-worthiness, and in the aversion to blame-worthiness; in the desire of possessing those qualities, and performing those actions, which we love and admire in other people; and in the dread of possessing those qualities. (*TMS*, III.2.32, 130)

As this quote illustrates, Smith not only refines the definition of intrinsic motivation, but he also offers insights how this motivation evolves, reminding us that the love of praiseworthiness is the love of warranted praise—there is and always will be an ‘extrinsic’ element present. That we wish to conduct ourselves to satisfy the love of praiseworthiness does not mean we have built-in knowledge about what is deemed worthy of praise and what is worthy of blame in different situations. Smith explains that this natural incentive, the voice of conscience, is perfected and cultivated through “slow, gradual and progressive work” (*TMS*, VI.iii.25, 247) and our experiences with the ‘man without’. He continues to assert that “virtue is not said to be amiable, or to be meritorious, because it is the object of its own love, or of its own gratitude; but because it excites those sentiments in other men” (*TMS*, III.I.7, 113). By this reasoning, depending less and less on the praise and blame with reference to actual others and more on the deservingness of praise and blame of the ‘man within’ allows us to become more autonomous in our moral judgments (Evensky 2005) and to learn to recognise deserved praise and how to mitigate the excess of undeserved praise (Hanley 2009).⁶

But what is the natural and proper object of praise and blame in one-player games? In the dice-rolling game, subjects who are misreporting in order to receive more money are violating a relatively clearly defined norm; the shared perception that honesty is the most appropriate action (Lois and Wessa 2021; Serdarevic 2021). By eliciting norms in the receiver game, however, Tjøtta (2019) reveals that only a minority of actual spectators deemed keeping the entire endowment as very socially inappropriate. Hence, even if subjects had chosen to keep more money, this would not necessarily have resulted in more disapproval from others. To

⁶ Sivertsen (2017) offers an interesting discussion on how the love of praiseworthiness may have evolved and contrasts his view with Hanley (2009), claiming that the love of praiseworthiness is a love redirected in the sense that our desire to be approved by others teaches us to view ourselves as others see us, how they *would* judge us had they been better informed, and how they *should* judge us as impartial spectators. A similar argument is offered by Uyl and Griswold Jr (1996).

understand what judgements could be at play in the receiver game, it is useful to consider Smith's criteria of how we judge the 'sentiments of mankind, with regard to the merit or demerit of actions':

Whatever praise or blame can be due to any action, must belong either, first, to the intention or affection of the heart, from which it proceeds; or, secondly, to the external action or movement of the body, which this affection gives occasion to; or, lastly, to the good or bad consequences, which actually, and in fact, proceed from it. These three different things constitute the whole nature and circumstances of the action, and must be the foundation of whatever quality can belong to it. (*TMS*, II.iii.intro.1, 92)

Firstly, as Smith asserts, what qualities must belong to an action, what is praiseworthy and what is blameworthy, can indeed be judged according to the outcomes that action produces. If we understand the love of praiseworthiness as the incentive to make a sacrifice or, as Hanley (2009) puts it, letting go of familiar pleasures, then choosing less money could certainly satisfy this criterion. When subjects are placed in a decision situation constructed to resemble "some solitary place, without any communication with his species" (*TMS*, III.1.3, 110), they bring with them the acquired "habit of conceiving" the approbation that should come from praiseworthy conduct, even if "admirers may neither be very numerous nor very loud in their applauses" (*TMS*, VI.iii.31, 253). This habit comes to represent a higher tribunal that makes up the motivation to exercise self-command, restraining them from doing something that might tamper with how they view their own character. Being spectators of themselves as acting agents, a substantial minority of subjects in the receiver game (22.6%) indeed seem to be able to let go of familiar pleasures, such as money, that are a driving force in many lives.

But monetary outcomes are not the end of the story in judging the praiseworthy and blameworthy qualities of an action. Moving forward, Smith emphasises the importance of intentions, arguing that:

The two last of these three circumstances [external action and consequences] cannot be the foundation of any praise and blame, is abundantly evident [...] the only consequences for which he can be answerable, or by which he can deserve either approbation or disapprobation of any kind, are those which are somewhat intended. (*TMS*, II.iii.intro.3, 93)

In the continuation of his argument, Smith thus carefully reminds us that intentions are what is truly laudable or blameable. Clearly, actual consequences which happen to proceed from any action have a very great effect upon our sentiments, but actions and outcomes must be judged in relation to intentions. Notably, we do not only reveal our intentions to others, like in two-person games, in the hope of recompense and acclamation, but also to ourselves, in the hope of self-applause because “no action can properly be called virtuous, which is not accompanied with the sentiment of self-approbation” (*TMS*, III.6.13, 177). By seeing ourselves from without, we are able to predict the judgements of others in our imagination. We moderate our self-interest because we know that we would be loved by others and, indirectly, by ourselves. If we follow Smith and Wilson (2019) in viewing intentions in experimental games as the alternative cost of the action taken, subjects choosing less money over more are paying a higher monetary cost, revealing less self-interested motives than those who choose more. An additional way to reveal subjects’ intentions is to ask *what* motivated their choice.⁷

IV.1. Analysis of the Receiver Game

Tjøtta (2019) analyses subjects’ answers in the receiver game, showing that many mention reasons consistent with non-distributive norms as an explanation for receiving less money (see his Table B1 page 75 for transcripts). I obtained the experimental data from Tjøtta for the purpose of analysing the qualitative data, paying particular attention to whether subjects mention reasons that pertain to moral character and how they judge deservingness in the receiver game setting.⁸ The centrality of these concepts flows from the idea that praiseworthiness is encouraged and supported by the habit of self-evaluation and self-approbation. Praiseworthiness supposes the human ability to imagine what deserves approval,

⁷ Many, myself included, have tended to ask subjects to explain their choices once their behaviour deviates from how we commonly think about motivation: more money being more desirable than less. But from Smith’s theory it becomes clear that the *why* question transcends outcomes, highlighting the importance of intentions, something to which scholars are increasingly paying attention. In addition to Tjøtta (2019), see, for instance, Capizzani et al. (2017) and Aguiar, Branas-Garza, and Miller (2008), who incorporate qualitative data to analyse behaviour in the ultimatum game and moral motivations for subjects’ giving behaviour in the dictator game, respectively.

⁸ To avoid biasing my interpretation, a research assistant categorised the answers from Tjøtta’s (2019) Experiment 6, in which a representative sample of the Norwegian population chose whether to receive more or less money before answering what motivated their choice. The coder only saw the text answers, not the choices made by subjects. This was done to minimise attribution bias whereby the coder assigns intentions to the subjects’ answers based on knowledge about their monetary choice.

without regard for what would actually be approved by someone. Furthermore, evaluating the deservingness of what is bestowed upon us—humbling down the elevation of mind when brought about from groundless acclamation or reproach—is what motivates us to continue seeking worthiness itself. Of course, the text analysis will not reveal whether the subjects' behaviour was grounded in the love of praiseworthiness or fear of blameworthiness. Still, one can examine whether they use reasons related to their character, esteem, or deservingness when asked to explain their behaviour.

A total of 1019 subjects were informed that:⁹

As a participant in the Norwegian Citizen Panel, you are being included in a drawing for an extra monetary prize. If you win, you can choose to receive 1000 kroner or 1800 kroner. Which would you choose? Please tick one of the options:

Yes, please, I would like to receive 1000 kroner

Yes, please, I would like to receive 1800 kroner

Table 1 in Panel A shows that among those who chose less, three predominant categories stand out: reasoning about one's character (9%), deservingness (21%), and other-regarding concerns (21%).¹⁰ The first group was concerned about whether they view themselves as morally upright people, mentioning reasons such as virtue, humility, and modesty for receiving less money. The second group reasons whether the amount chosen (forgone) was deserved given the particular game-setting, arguing that they had not done enough work to deserve more money. The latter group reported that they intended to give the money to charity and that the remaining amount could go to someone who was struggling.

Turning the focus to subjects choosing more in Table 1, Panel B reveals an interesting pattern: subjects choosing more money also offer reasons that are consistent with concerns for character, deservingness, and others, albeit with a somewhat different image according to which they

⁹ See page 74 in Tjøtta (2019) for experimental instructions. Note that 1000 Norwegian kroner corresponded to 115 USD at the time of the experiment.

¹⁰ Some subjects' answers were removed from the data for anonymity reasons, leaving a text-analysis sample of 927 answers. Following Tjøtta (2019), I use 214 (93.0%) answers of those who chose less and 713 (90.4%) of those who chose more. Subjects whose answers could not be classified or consisted of multiple motivations were categorised as 'combination'. Twenty percent among those who chose less fell within this category, while this was the case for 7% of those choosing more. For simplicity, these categories are not depicted in Table 1.

A. MOTIVATION	N	LESS MONEY EXAMPLES
Character	19/214	I am modest. I am not greedy. I don't like to be greedy. I like to view myself as unique. Humility. I do not like greed-culture. Showing moderation. Modesty. Virtue. I am showing virtue. I do not want to be demanding. I am not motivated by money. Defiant.
Deservingness	46/214	The amount is large enough for this kind of participation. One should do this without getting paid. I do this voluntarily and do not need to get paid. My participation is not worth that much. A lot of money for such little effort. 1800 is too much money for 20 minutes. I do not deserve more for this.
Other	44/214	I will give to charity. Buy something nice for my wife. There are other people who need it more than me. The remaining 800 could be given to someone who is struggling. Can be used for good causes. I want to give the money to a charity, i.e., help to Syria. Give to charity. Others need it much more than me. Let the rest go to charity. The rest can go to Doctors without Borders.
Experimenter	7/214	Research is expensive. Perhaps you can use the remaining money for a 'research pot'. So that the Norwegian Citizen Panel can use the money for something else. More money for future research. Less costs of research. Money can be used for some research.
Value of money	29/214	Money is not everything in this world. Other things than money matter too. I have enough money. Money is not everything although we do depend on money. One does not always need to have the most. I do not need more money.
Misunderstand	25/214	Higher probability of winning. I thought the chances were higher. Possibility of winning is higher. More people can get 1000, chances to receive increase. Maybe more people can be drawn if I choose less. So that more people can win. I don't want it all. Maybe more people can win if I don't take it all.
B. MOTIVATION	N	MORE MONEY EXAMPLES
Character	39/713	Not a saint, need the money. Mostly greed. I am just being sincere. I do not want to pretend to be modest. Being modest is not a virtue here. Pure selfishness. Because of my greed. I just want to be honest. Honesty. I do not want to be falsely modest. The more you have, the more you will want.
Deservingness	39/713	I deserve this. My effort is worth that much. Deservingness. It takes time to answer these questions. I assume that others would do it, but I really think this should be done for free. Because of my willingness to contribute. This is the price for my work. My time costs that much. I am a student with no income so I feel I deserve it.
Other	65/713	Will donate part of it. Will give to Amnesty. Useful in the family budget. To share it with my grandkids. Buying shoes for my kid. Could benefit my family. I will give my children extra money for their education. Will share it with the missus. Can be paid forward to Doctors without Borders. Give to cancer research. Give to the local football club.
Experimenter	1/713	I am a student with a loan, I do not think the extra 800 would harm the finances of Norwegian Citizen Panel much.
Value of money	509/713	I really need the money. Cash is king. 1800 is 80% more than 1000. Money is good and more money is better. The size of the sum. More is more useful. Higher sum. Because it is more money. It was a better offer and 1800 is more valuable to me. The more money the merrier. Simple: 1800 is more worth than 1000. Money is freedom.
Misunderstand	10/713	I do not get the question. Because people think they have a higher chance of winning. Probably a trick question with respect to taxation.

Table 1: Classification of randomly chosen open-ended answers in the receiver game

view themselves and the context of the receiver game.¹¹ Answers related to subjects' character (5%) emphasise that being modest is not necessarily a virtue in this setting, but that they indeed view themselves as honest. Choosing more is the sincere action and they do not want to be falsely modest. In terms of deservingness, subjects argue that this is the price for their attendance and that they therefore deserve the highest payment (5%). Approximately 9% of those who choose more state that they intend to share the money with others such as charities, grandchildren, friends, and family members.

Clearly, subjects could have preferences towards the experimenter: choosing less money means more money left for the experimenter. The reality is that in any experiment, there is at least one person conducting the experiment, making this person a spectator to take into account. Consider the ultimatum game in which a receiver chooses to reject the proposer's offer. This results in neither party receiving any money. In practice, this would mean that the entire endowment is left to the experimenter. As with the receiver game, there is no information where the remaining money goes. If subjects are systematically affected by whether their choices earn them the approval of the experimenter, then this could compromise the interpretation of many experiments. Three percent of subjects who chose less money mentioned research. One percent choosing more did the same. In economics, concerns for the experimenter are interpreted as a challenge, as all researchers ideally want to reveal the true preferences of their subjects and not what they think the researcher(s) wants (Zizzo 2010).¹² Finally, the receiver game is about receiving money. Subjects choosing less seem to reason about the value of

¹¹ Most situations that involve communication enable people to engage in cheap talk whereby they make unverifiable statements about private information and future action out of concern for appearances (Crawford and Sobel 1982). While this may be a challenge in general when it comes to interpreting text-answers, one could be particularly worried that those who choose more money and say that they intend to share it with others, do this to rationalise their choices to appear other-regarding or, in Smith's words, "appear to be fit for society" (*TMS*, III.2.7, 117). Still, even this would reveal that subjects are aware of some general rule of conduct that governs this particular situation and their need to maintain conformity between their action and the seeking of praise and praiseworthiness. Recall that Smith is open about the fact that we are not making judgements based on principles a "perfect being would approve of" (*TMS*, II.i.5.10, 77) and that moral judgements will never be perfectly impartial beyond any doubt. We gradually learn how to turn the lens inward and become better and more impartial in our self-evaluation.

¹² Frank (1998) shows that burning money in front of the subjects instead of letting the experimenter keep it makes no difference to the subjects' behaviour. Chlaß and Moffatt (2017) find that, if anything, concerns for the experimenter are negatively related to generosity, meaning that that subjects would choose more money if the experimenter's role was influencing their behaviour.

money differently than those who choose more: whereas 14% of subjects who choose less argue that more money is not necessarily always better and that other things in life also matter (see category ‘Value of money’ in Table 1), 71% of those who choose more simply argue that more money is more useful, revealing no further motivation.

It is also worth mentioning that the simplicity of the receiver game may make it easy to misunderstand its rules. Typically, in economic experiments, subject comprehension is about identifying the selfish best response and the social welfare-maximizing strategies (Bartke et al. 2019). Of those who chose more, 1% were classified as misunderstandings; 12% of those who chose less gave answers that could be consistent with misunderstanding (i) the probability of being drawn to receive the money, and (ii) that choosing less did not increase the probability of someone else receiving it. Some subjects believe that taking less increases their own chances of being drawn for the money, while others think that choosing less would increase the chances of someone else in the experiment. Notably, this latter reason of misunderstanding is related, yet different from the ‘Others’ category reported in Table 1. The main difference being that subjects who are classified as having other-regarding concerns state that they intend to personally give the money to others or that the remaining money should go to someone who needs it more. They do not use the probability of being drawn as a reason for choosing less—which is independent of subject’s choices in the receiver game.

V. CONCLUDING REMARKS

Simple questions like ‘Do you want more or less money?’ are not always accompanied by simple answers. While I agree with the economics literature that choosing less over more may be consistent with intrinsic motivations, this explanation invites deeper questions: What is this intrinsic motivation comprised of? Why act on what you think is right even when it does not get you praise from other people? Why avoid acting on what you think is blameworthy even when no one is watching? Is the love of praiseworthiness and fear of blameworthiness in experimental games only determined by monetary outcomes? Presented in these terms, I have argued that there is still need for a theoretical account that can explain the process through which we learn to seek worthiness and act consistent with it.

Smith offers such a theoretical account. To know what is praised and blamed is a first step on the road toward virtue, but to be virtuous we

need something more that motivates us to enforce these standards upon ourselves in different contexts. My reading of Smith prompts me to believe that this ‘something’ is the love of praiseworthiness and dread of blameworthiness. This additional component Smith adds to his analysis is important in order to understand the development and role of the moral conscience. Having an interest in others and fearing their resentment cannot alone explain the evolution of other-regarding behaviour. Rather, the evolution of moral agency depends on another incentive, an inner strength to do what one perceives to be right.

In this paper, I have argued that this inner strength is likely to be highlighted in the receiver game. Having established that a substantial minority of subjects in one-player games choose less money, I proceeded to show that, for the most part, the motivations provided by those who chose less differs from those who chose more. The majority of those who chose less offered reasons related to some sense of self-satisfaction and inward tranquillity—feelings associated with the knowledge that they have acted in a worthy manner even in the absence of actual spectators. The majority of those who chose more simply argued that more reward is better and less is worse. Still, the analysis challenges common dichotomous perceptions of what motivates behaviour in economics experiments, where those who choose less are often portrayed as altruistic or motivated by genuine moral concerns. Those who choose more as non-cooperative or having selfish motivations. Applications of Smith’s theory to understand subjects’ open-ended answers in a one-player game refines and nuances these perceptions—those who choose more also engage in the same self-approbation process as those who choose less. In fact, a substantial minority evaluate their own character and deservingness, and reveal intentions that take into account others’ well-being in addition to their own.

Notably, and as already indicated in the introduction, the relevance of an inner strength to do what one perceives is right is not new to economics (Bénabou and Tirole 2003; Bénabou and Tirole 2006; Fehr and Schmidt 1999). Borrowing from social psychology, economists have defined extrinsic motivations as pure externally motivated rewards such as money and praise, at one extreme. For instance, we wish to avoid rejection by responders in the ultimatum game; we do not want to lose money and the praise from the responder. However, this type of motivation does not make us pursue an action for its own sake or value. At the other extreme, intrinsic incentives have been introduced as a residual motivation, giving

enjoyment and utility for its own sake, and often viewed independent (negatively correlated) of financial incentives (Remic 2021). Still, this interpretation only seemingly delineates the underlying process that drives this impulse.

Smith's moral theory, as I read it, deepens and complements such intrinsic motivation explanations, but is grounded in a rich theoretical system that focuses on the *social process*, on the evolvment of "the excellent and praise-worthy character, the character which is the natural object of esteem, honour, and approbation" (*TMS*, VII.I.2, 265). My argument echoes Bénabou and Tirole (2006) who similarly look to Smith and self-image concerns to shed light on the channels and mechanisms involved in sustaining and inhibiting intrinsic motivations, emphasising the role of deservingness and self-evaluation through the eyes of other fair and impartial spectators. I think Smith's theory of motivation is a source of insightful alternative interpretations and explanations; what experimentalists have to date termed 'intrinsic costs,' 'guilt aversion,' 'greed aversion' and 'intrinsic honesty' as explanations to forgone money in one-player games could be related to what Smith refers to as the self-directed remorse that arises when we know ourselves to be blameworthy. Similarly, self-directed gratitude arises when we know ourselves to be praiseworthy.

As mentioned by Paganelli (2009) in the opening paragraph, Smith can indeed contribute to illuminating yet another experimental puzzle.

REFERENCES

- Abeler, Johannes, Anke Becker, and Armin Falk. 2014. "Representative Evidence on Lying Costs." *Journal of Public Economics* 113: 96–104.
- Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for Truth-Telling." *Econometrica* 87 (4): 1115–1153.
- Aguiar, Fernando, Pablo Branas-Garza, and Luis. M. Miller. 2008. "Moral Distance in Dictator Games." *Judgment and Decision Making* 3(4): 344–354.
- Ashraf, Nava, Colin F. Camerer, and George Loewenstein. 2005. "Adam Smith, Behavioral Economist." *Journal of Economic Perspectives* 19 (3): 131–145.
- Bartke, Simon, Steven J. Bosworth, Dennis J. Snower, and Gabriele. Chierchia. 2019. "Motives and Comprehension in a Public Goods Game with Induced Emotions." *Theory and Decision* 86 (2): 205–238.
- Becker, Garry. S. 1976. *The Economic Approach to Human Behavior*. Chicago, IL: University of Chicago Press.
- Bénabou, Roland and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *The Review of Economic Studies* 70 (3): 489–520.
- Bénabou, Roland and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–1678.

- Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, reciprocity, and social history." *Games and Economic Behavior* 10 (1): 122-142.
- Brown, Vivienne. 2011. "Intersubjectivity, the Theory of Moral Sentiments and the Prisoners' Dilemma." *Adam Smith Review* 6: 172-190.
- Camerer, Conlin. F. and Richard H. Thaler. 1995. "Anomalies: Ultimatums, Dictators and Manners." *Journal of Economic Perspectives* 9 (2): 209-219.
- Capizzani, Mario, Luigi Mittone, Andrew Musau, and Antonino Vaccaro. 2017. "Anticipated Communication in the Ultimatum Game." *Games* 8 (3): 29.
- Cappelen, Alexander W., Trond Halvorsen, Erik Ø. Sørensen, and Bertil Tungodden. 2017. "Face-Saving or Fair-Minded: What Motivates Moral Behavior?" *Journal of the European Economic Association* 15 (3): 540-557.
- Chlaß, Nadine and Peter G. Moffatt. 2017. "Giving in Dictator Games: Experimenter Demand Effect or Preference over the Rules of the Game?" Jena Economic Research Paper No. 2012-044. Friedrich Schiller University and the Max Planck Institute of Economics, Jena.
- Crawford, Vincent. P. and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica: Journal of the Econometric Society* 50 (6): 1431-1451.
- Dana, Jason, Roberto A. Weber, and Jason X. Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33 (1): 67-80.
- Dufwenberg, Martin and Martin A. Dufwenberg. 2018. "Lies in Disguise: A Theoretical Analysis of Cheating." *Journal of Economic Theory* 175: 248-264.
- Engel, Christoph. 2011. "Dictator Games: A Meta Study." *Experimental Economics* 14 (4): 583-610.
- Evensky, Jerry. 2005. "Adam Smith's Theory of Moral Sentiments: On Morals and Why They Matter to a Liberal Society of Free People and Free Markets." *Journal of Economic Perspectives* 19 (3): 109-130.
- Fehr, Ernst. and Klaus. M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics* 114 (3): 817-868.
- Fischbacher, Urs and Franzisk Föllmi-Heusi. 2013. "Lies in Disguise: An Experimental Study on Cheating." *Journal of the European Economic Association* 11 (3): 525-547.
- Forman-Barzilai, Fonna. 2010. *Adam Smith and the Circles of Sympathy: Cosmopolitanism and Moral Theory*. Cambridge: Cambridge University Press.
- Forsythe, Robert, Joel L. Horowitz, Nathan E. Savin, and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior* 6 (3): 347-369.
- Frank, Björn 1998. "Good News for Experimenters: Subjects do Not Care About Your Welfare." *Economics Letters* 61 (2): 171-174.
- Frey, Bruno. S. and Felix Oberholzer-Gee. 1997. "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out." *The American Economic Review* 87 (4): 746-755.
- Gächter, Simon and Jonathan F. Schulz. 2016. "Intrinsic Honesty and the Prevalence of Rule Violations Across Societies." *Nature* 531 (7595): 496-499.
- Gneezy, Uri and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." *The Quarterly Journal of Economics* 115 (3): 791-810.
- Griswold Jr, Charles L. 1999. *Adam Smith and the Virtues of Enlightenment*. New York, NY: Cambridge University Press.

- Güth, Werner, Rolf Schmittberger, and Bernerd Schwarze. 1982. An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization* 3 (4): 367–388.
- Güth, Werner and Reinhard Tietz. 1990. “Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results.” *Journal of Economic Psychology* 11 (3): 417–449.
- Hanley, Ryan P. 2009. *Adam Smith and the Character of Virtue*. New York, NY: Cambridge University Press.
- Hoffman, Elizabeth, Kevin McCabe, and Vernon Smith. 2008. “Reciprocity in Ultimatum and Dictator Games: An Introduction.” In *Handbook of Experimental Economics Results*, edited by Charles R. Plott, and Vernon L. Smith, 411–416. Amsterdam: Elsevier.
- Kahneman, Daniel, Jack, L. Knetsch, and Richard Thaler. 1986. “Fairness as a Constraint on Profit Seeking: Entitlements in the Market.” *The American Economic Review* 76 (4): 728–741.
- Kajackaite, Agne and Uri Gneezy. 2017. “Incentives and Cheating.” *Games and Economic Behavior* 102 (1): 433–444.
- Krupka, Erin L. and Roberto A. Weber. 2013. “Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?” *Journal of the European Economic Association* 11 (3): 495–524.
- Lois, Gianni and Michele Wessa. 2021. “Honest Mistake or Perhaps Not: The Role of Descriptive and Injunctive Norms on the Magnitude of Dishonesty.” *Journal of Behavioral Decision Making* 34 (1): 20–34.
- Meardon, Stephen J. and Andreas Ortmann. 1996. “Self-Command in Adam Smith’s Theory of Moral sentiments: A Game-Theoretic Reinterpretation.” *Rationality and Society* 8 (1): 57–80.
- Müller, Jerry Z. 1993. *Adam Smith in His Time and Ours: Designing the Decent Society*. Princeton, 1 – 263, NJ: Princeton University Press.
- Paganelli, Maria P. (2009). Smithian Answers to Some Puzzling Results in the Experimental Literature. In *Elgar Companion to Adam Smith*, edited by Jeffery T. Young, 181–192. Cheltenham: Edward Elgar
- Remic, Blaž. 2021. “Three Accounts of Intrinsic Motivation in Economics: A Pragmatic Choice?” *Journal of Economic Methodology*, 1–16.
- Romaniuc, Rustam. 2017. “Intrinsic Motivation in Economics: A History.” *Journal of Behavioral and Experimental Economics* 67: 56–64.
- Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. “Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study.” *The American Economic Review* 81 (5):1068–1095.
- Rummel, Rudolph J. 1976. “Social Behavior and Interaction.” In *Understanding Conflict and War, Volume 2: The Conflict Helix*. Beverly Hills, CA: Sage Publications.
- Scitovsky, Tibor. 1976. *The Joyless Economy: An Inquiry into Human Satisfaction and Consumer Dissatisfaction*. London: Oxford University Press.
- Serdarevic, Nina. 2021. “Licence to Lie and the Social (In)Appropriateness of Lying.” *Economics Letters* 199 (4): 1–5.
- Serdarevic, Nina and Sigve Tjøtta. 2021. “A Cross-National Study on the Receiver Game.” SSRN Working Paper No. 3963932. SSRN, Rochester, NY.
- Sivertsen, Sveinung S. 2017. “Love Redirected: On Adam Smith’s Love of Praiseworthiness.” *Journal of Scottish Philosophy* 15 (1): 101–123.

- Smith, Adam. 1982. *The Theory of Moral Sentiments*. Edited by D.D Raphael and A.L Macfie. The Glasgow Edition of the Works and Correspondence of Adam Smith: 1 - 422, Liberty Fund, Charmel, Indiana.
- Smith, Vernon L. and Bart J. Wilson. 2018. "Equilibrium Play in Voluntary Ultimatum Games: Beneficence Cannot be Extorted." *Games and Economic Behavior* 109: 452-464.
- Smith, Vernon L. and Bart J. Wilson. 2019. *Humanomics: Moral Sentiments and the Wealth of Nations for the Twenty-first Century*. Cambridge Studies in Economics, Choice, and Society. Cambridge: Cambridge University Press.
- Tisserand, Jean-Christian. 2014. "Ultimatum Game: A Meta-Analysis of the Past Three Decade of Experimental Research." In *Proceedings of International Academic Conference, Antibes, France, 13*, 609-609. Prague: International Institute of Social and Economic Sciences.
- Tjøtta, Sigve. 2019. "More or Less Money? An Experimental Study on Receiving Money." *Journal of Behavioral and Experimental Economics* 80: 67-79.
- Utikal, Verena and Urs Fischbacher. 2013. "Disadvantageous Lies in Individual Decisions" *Journal of Economic Behavior & Organization* 85: 108-111.
- Uyl, Douglas. J. Den. and Charles L. Griswold Jr. 1996. "Adam Smith on Friendship and Love." *The Review of Metaphysics* 49 (3): 609-637.
- Young, Jeffery T. 2009. *Elgar Companion to Adam Smith*. Cheltenham: Edward Elgar.
- Zizzo, Daniel J. 2010. "Experimenter Demand Effects in Economic Experiments." *Experimental Economics* 13 (1): 75-98.

Nina Serdarevic is a researcher in behavioural economics and a member of the FAIR Insight Team at the Centre for Applied Research (the Norwegian School of Economics) in Norway. Her research interests lie in the intersection of experimental economics and moral and political philosophy. Currently, her research focuses on the representativeness of elite preferences and views on inequality and redistribution.
Contact e-mail: <nina.serdarevic@snf.no>

Social Contract, Extended Goodness, and Moral Disagreement

CYRIL HÉDOIN

University of Reims Champagne-Ardenne

Abstract: This article discusses the role played by interpersonal comparisons (of utility or goodness) in matters of justice and equity. The role of such interpersonal comparisons has initially been made explicit in the context of social choice theory through the concept of extended preferences. Social choice theorists have generally claimed that extended preferences should be taken as being uniform across a population. Three related claims are made within this perspective. First, though it is sometimes opposed to social choice theory, the social contract approach may also consider the possibility of interpersonal comparisons. This is due to the fact that justice principles may be partially justified on a teleological basis. Second, searching for the uniformity of interpersonal comparisons is both hopeless and useless. In particular, moral disagreement does not originate in the absence of such uniformity. Third, interpersonal comparisons should be accounted for both in social choice and social contract theories in terms of sympathetic identification based on reciprocal respect and tolerance, where each person's conception of the good partially takes care of others' good. From the moral point of view, any person's conception of the good should thus be 'extended' to others' personal conceptions. This extension is, however, limited due to the inherent limitations in sympathetic identification and is a long way from guaranteeing the uniformity assumed by social choice theorists.

Keywords: social contract theory, social choice theory, extended preferences, interpersonal comparisons, teleological justification

JEL Classification: D63

I. INTRODUCTION

The social choice approach and social contract theory are two broad traditions that aim at reflecting on possible ways for overcoming the

AUTHOR'S NOTE: This article has been presented at the "Social Justice in a Complex World" international workshop, held in Reims in November 2019. It has benefited from the comments of the participants whom I thank. I also thank two anonymous referees who have provided detailed and sharp comments on a previous version. All remaining mistakes are my own.

plurality of judgments and viewpoints to establish a social or collective agreement over rules and choices. These traditions have been traditionally opposed with respect to moral issues, especially those concerned with justice and equity (Gaus 2011; Sen 2017). This paper addresses an aspect on which they may be thought to be in opposition, i.e., the respective role played and the form taken by interpersonal comparisons (of utility, of goodness) in these two traditions.

Interpersonal comparisons of utility have a long and controversial history in normative economics. Their rejection in the 1930s by influential economists on the ground that they rely on unscientific value judgments is directly responsible for Arrow's (1963) impossibility result in social choice theory. They have been subsequently rehabilitated by Sen (1970) and others. One motivation for this rehabilitation was related to the consideration of issues related to equity and justice: once social choice theory is used to account for moral principles and doctrines such as utilitarianism or Rawls' difference principle, escaping interpersonal comparisons is no longer possible; there must be some ways through which utilities (thought to correspond to individual welfare or any other morally relevant metric) can be compared. The concept of *extended preferences* is the main device by which interpersonal comparisons have been made meaningful within a social choice framework. They correspond to a binary preference relation between pairs of variables (x, i) where x refers to a social alternative or position and i to a personal identity. Then, an extended preference indicates whether one prefers to be individual i in social alternative x or individual j in social alternative y . Interestingly, there are several instances in the literature of 'hybrid' accounts mixing an explicit social choice framework with a broad contractualist commitment over moral matters. Because they use social choice theory as their formal roots, these accounts have tended to rely on the concept of extended preferences. This creates a significant constraint on deriving moral conclusions, i.e., that individuals across a society share the same set of extended preferences. However, arguments justifying such a uniformity assumption are left wanting.

From this perspective, this paper makes three related claims about the comparative status of interpersonal comparisons in social choice and social contract theories. First, I argue that while the concept of extended preferences should be dispensed with altogether, social contract theorists might also consider the need for an appropriate account of the way members of a society settle over a common conception of goodness. This is

due to the fact that some contractualist accounts may partially rely on a teleological form of justification for a moral code. Second, even though establishing such a common conception entails making interpersonal comparisons of goodness possible, searching for the uniformity of interpersonal comparisons is both hopeless and useless. In particular, moral disagreement does not originate primarily in the absence of such uniformity. Third, interpersonal comparisons should be accounted for both in social choice and social contract theories in terms of sympathetic identification based on reciprocal respect and tolerance, where each person's conception of the good partially takes care of others' good. From the moral point of view, any person's conception of the good should thus be 'extended' to others' personal conceptions. This extension is, however, limited due to the inherent limitations in sympathetic identification and is a long way from guaranteeing the uniformity assumed by social choice theorists.

As a result, this article develops a rationale for what can be called a (partially) *teleological contractualism*. This rationale helps to show that the opposition between social choice and social contract approaches to justice and equity issues is less strong than it is generally thought. I proceed through a comparison with other contractualisms. Interestingly, although John Rawls (1971) has developed a thoughtful criticism of teleological accounts of justice, it appears that Rawlsian contractualism also has a teleological dimension that materializes in Rawls' thin theory of the good and concept of primary goods. Gerald Gaus' version of contractualism initially also relied on a teleological form of justification (Gaus 1990), but has more recently given up any reference to the good (Gaus 2012). I shall suggest, however, that teleological contractualism is better equipped to deal with the problem that is at the core of both Rawlsian and Gaussian contractualism, namely the problem of moral disagreement.

The rest of the paper is organized as follows. Section II briefly presents the concept of extended preferences as developed within social choice theory and the problem of their non-uniformity across a population. Section III presents an argument to the effect that social contract theory, even if taken to constitute an alternative to the social choice approach, may take advantage of considering the role of interpersonal comparisons. This '(partially) teleological contractualism' goes further than Rawls' use of a thin theory of the good. Section IV develops an account of 'extended goodness' and argues that interpersonal comparisons should be accounted for in terms of sympathetic identification. Section V

concludes, reflecting on the fact that extended goodness judgments are unlikely to be uniform and complete. While this a source of moral disagreement, it is, however, not the only one.

II. EXTENDED PREFERENCES IN SOCIAL CHOICE THEORY

The status of interpersonal comparisons of utility in welfare economics has been controversial at least since Lionel Robbins' (1938) claim that such comparisons necessarily involve unscientific value judgments. The ensuing rejection of interpersonal comparisons has considerably restricted the range of welfare criteria available to assess states of affairs. Arrow's (1963) concept of 'social welfare function' defined as a function from a vector of individual ordinal rankings to a social preference ordering effectively implies the impossibility of making interpersonal comparisons. The exclusion of interpersonal comparisons within the Arrowian framework directly leads to the infamous impossibility result that marked the birth of social choice theory: there is no social welfare function satisfying both a Paretian and a non-dictatorship condition that is defined on any vector of individual rankings and that orders two social states only as a function of the individual orderings of these two states.

Social choice theorists have started to reconsider the role and the legitimacy of interpersonal comparisons for two related reasons. On the one hand, allowing for ordinal or even cardinal comparability has proved sufficient at the formal level to avoid Arrow's impossibility result. This research program, opened by Sen (1970), has established that using a broader framework than Arrow's social welfare function allows for a richer informational basis of social choice. On the other hand, as explicitly stated by Sen (1970) and Arrow (1978), while a general theory about collective decision, social choice theory partially overlaps with theories of justice. In particular, social choice may have to select alternative distributions (of welfare, of satisfaction), and hence, "serves the same function as the principle of distributive justice and might be identified with it" (Arrow 1978, 223). It is unclear, however, how social choice could produce an evaluation in terms of distributive justice while prohibiting any kind of interpersonal comparison.

From a purely technical perspective, allowing for ordinal or cardinal interpersonal comparability in a social choice framework is unproblematic. This is easily achieved within the 'social welfare functional' and 'welfarist' approaches of social evaluation that has been dominant since the

1970s.¹ Indeed, within this framework, measurability and comparability assumptions are fully accounted for by the uniqueness properties of the utility functions serving as inputs for the social evaluation (Weymark 2016).² However, this approach is silent with respect to the source of the information allowing for the different kinds of interpersonal comparisons: the possibility of interpersonal comparisons is stipulated rather than demonstrated. The concept of extended preferences is constitutive of a whole methodological and theoretical account for making interpersonal comparisons. If successful, it would provide social choice theorists with a way to justify the comparability assumptions made within a social choice framework. Unsurprisingly, this account has been advocated by welfare economists and social choice theorists such as Suppes (1966), Sen (1970), Harsanyi (1977), and Arrow (1978) interested in issues related to social justice. It is still regarded as the main way to give meaning to interpersonal comparisons and to make them eventually operational.³

As the name indicates, extended preferences are based on the preference concept that is at the core of modern economics, both positive and normative. I will ignore here the debates surrounding both the definition and the measure of preferences. What matters here is that generally economists regard preference satisfaction either as a proxy for or as constitutive of an agent's welfare. Formally, preferences correspond to a set of binary relations R_i where xR_iy means that individual i weakly prefers social alternative x to social alternative y .⁴ These binary relations are generally assumed to be reflexive, complete, and transitive, and thus to define

¹ 'Welfarist' is an ambiguous term. It is understood here in the sense of 'formal welfarism' as characterized by Fleurbaey (2003), i.e., a formal approach that makes the social ordering fully dependent on the individual utility functions of the agents constituting the relevant population. This approach is, however, silent regarding the substantive interpretation of the utility functions, e.g., whether they represent happiness or preference satisfaction. Formal welfarism should not be conflated with 'real welfarism', a substantive moral doctrine which is the target of Sen's criticism in several articles (e.g., Sen 1979).

² Take the two following examples. The Rawlsian maximin criterion requires ordinal measurability and full comparability. It is obtained if each individual preference ordering is represented by a utility function unique up to any *common* monotonic positive transformation. On the other hand, cardinal measurability and full comparability is required to define a prioritarian social welfare function. It implies that individual utility functions are unique up to any *common* affine positive transformation.

³ The main alternative is constituted by money-metric approaches which only allow for indirect and essentially ordinal interpersonal comparisons. In a nutshell, they consist in determining individuals' willingness to pay for achieving a state of affairs or in identifying income equivalents and then using these measures as proxies for individuals' preferences or welfare.

⁴ The corresponding relations of strict preference P_i and indifference I_i are defined in terms of R_i : xP_iy if and only if xR_iy and not yR_ix ; xI_iy if and only if both xR_iy and yR_ix .

(pre-)orderings of social alternatives. In some cases, additional assumptions can be made. A continuity condition allows each ordering R_i to be represented by a set of utility functions u_i , all positive monotonic transformations of each other. Moreover, if the relations R_i are also defined over probabilistic distributions (i.e., lotteries or prospects) of social states and satisfy a sure-thing or independence axiom, then the functions u_i are cardinal, i.e., they represent the same ordering up to all positive affine transformations of each other.⁵ Mathematically speaking, extended preferences will be defined by the same set of properties. They correspond to an ordering and may be represented by a set of ordinal or cardinal utility functions. The difference is the domain over which these relations are defined. Denote X the set of social alternatives to be evaluated and ranked. Any social alternative $x \in X$ is an exhaustive description of everything that is relevant from a normative point of view, including possibly wealth distribution, health states, happiness levels, and so on. Denote N the set of individuals figuring in the relevant population—I will assume here that N is fixed, i.e., we ignore population issues in collective choices. Each individual $i \in N$ is endowed with a preference ordering R_i over X . I will assume that each ordering can be represented by a set of utility functions u_i but put aside for the moment the question of whether additional assumptions are relevant, especially about the cardinality of the functions u_i . A classical social choice exercise is to determine restrictions on the set of possible social orderings R^* given any profile $\{u_i\}_{i \in N}$ of utility functions. As an illustration, we may think it relevant to impose a weak Pareto condition such that if $u_i(x) > u_i(y)$ for all $i \in N$, then xR^*y . As I have stated the problem, the social choice is also restricted by the relative ‘thinness’ of the informational basis. Indeed, because each function u_i in any profile $\{u_i\}_{i \in N}$ is unique up to any positive monotonic transformation, we are free to use instead, for any person i , any function $v_i = f_i(u_i)$ such that $v_i(x) \geq v_i(y)$ if and only if $u_i(x) \geq u_i(y)$. Because there is no need to apply the same transformation to all individuals, utilities are obviously non-comparable. As I indicate above, this restricts the range of possible social welfare functions.

⁵ Roughly, the independence condition states that an agent weakly prefers a lottery L over a lottery L' , if and only if she prefers the ‘compound’ lottery M over the ‘compound’ lottery M' , where M and M' are formed by the same probabilistic distribution of L and L'' on the one hand and L' and L'' on the other hand, with L'' any lottery. Independence then guarantees that the preference relation over any pair of lotteries depends only on the ‘non-constant’ part of these lotteries.

An extended preference relation R_i^E is (minimally) defined over the Cartesian product $X \times N$, i.e., over all pairs of social alternatives and individuals. Hence, the statement $(x, i)R_k^E(y, j)$ may be read as ‘individual k weakly prefers to be individual i in social state x than individual j in social state y ’. An alternative and—as it will appear—more satisfactory reading is ‘individual k judges as good or better to be individual i in social state x than individual j in social state y ’. I shall, however, leave this interpretative issue for the next sections. In any case, the point of defining extended preference relations is that they offer a basis to make interpersonal comparisons of utilities. If, above of the fact of defining orderings, the relations R_i^E are continuous, then they can be represented by sets of utility functions u_i^E such that $u_k^E(x, i) \geq u_k^E(y, j)$ if and only if $(x, i)R_k^E(y, j)$. Hence, from individual k ’s point of view, individuals i ’s and j ’s utilities can be compared, ordinally at least. We may go farther and assume that the extended preference relations are not only defined over $X \times N$ but also over $\Delta(X \times N)$, i.e., the set of all probabilistic distributions of pairs of social alternatives and individuals. Call any such probabilistic distribution $L \in \Delta(X \times N)$ an *extended lottery*. If, in addition to the preceding conditions, each R_i^E also satisfies an independence requirement, then they can be represented by utility functions from which *cardinal* interpersonal comparisons can be derived. As an illustration, suppose individual k has to compare two lotteries L and L' . The former corresponds to an equiprobable distribution of (x, i) and (y, j) and the latter to an equiprobable distribution of (w, i) and (z, j) . Now, if LR_k^EL' , then that implies $u_k^E(x, i) + u_k^E(y, j) \geq u_k^E(w, i) + u_k^E(z, j)$ and therefore $u_k^E(x, i) - u_k^E(w, i) \geq u_k^E(z, j) - u_k^E(y, j)$, i.e., the utility *difference* between (x, i) and (w, i) is higher than or equal to the utility *difference* between (z, j) and (y, j) .⁶ This quantitative information notably opens the door for a myriad of utilitarian-based social welfare functions.

The formalism of the preceding paragraphs has left two related questions unanswered. The first concerns the meaning of the binary relations R_i^E and more generally how extended preferences should be interpreted. The second is about the extent to which extended preferences can be expected to be uniform across a whole population. The two issues are dependent since the answer to the first one will presumably make a difference with respect to the answer to the second issue.

⁶ Note that the fact that the two pairs of ‘extended alternatives’ involve the same persons is irrelevant. We could substitute any persons i' and j' for i and j in (w, i) and (z, j) without that making any difference.

Regarding the interpretational issue, virtually all social choice theorists have suggested that extended preferences are obtained through a process of *empathetic* identification. It is especially clear in Harsanyi's writings:⁷

Value judgments in social welfare [...] may still be interpreted as an expression of what sort of society one would prefer if one had an equal chance to be 'put in the place of' of any member of the society. (Harsanyi 1953, 435)

We have assumed that *i* will attempt to assess these utilities $u_j(x)$ by some process of *imaginative empathy*, i.e. by imagining himself to be *put in the place of individual j* in social situation *x*. This must obviously involve his imagining himself to be placed in individual *j*'s *objective position*, i.e. to be placed in the objective conditions (e.g. income, wealth, consumption level, state of health, social position) that *j* would face in social situation *x*. But it must also involve assessing these objective conditions in terms of *j*'s own *subjective attitudes* and *personal preferences* (as expressed by *j*'s own utility function u_j). (Harsanyi 1977, 51–52; emphasis in original)

Empathetic identification, or 'imaginative empathy', is achieved by putting oneself in others' shoes, i.e., by identifying oneself with all the objective and subjective features constitutive of other individuals' social position, and personal identities. This interpretation almost requires accepting what can be called a 'sovereignty principle' according to which individual *i*'s extended preferences must respect individual *j*'s preferences over any pair of social alternatives, i.e., $R_i^E = R_j$ over the restricted domain $X \times \{j\}$. As noted by Mongin (2001) and other commentators, the interpretation of extended preferences in terms of empathetic identification does not save this account from difficult ambiguities. Two are worth noting. A first difficulty is related to the concept of preferences. Welfare economists have generally identified preferences and welfare on the basis of a—mostly intuitive—consumer sovereignty principle. It has been pointed out many times, however, that preference satisfaction cannot be constitutive of welfare, especially if preferences are understood in terms of actual or hypothetical choices (e.g., Hausman and McPherson 2006; Sen

⁷ As it may create some confusion with the terminology I am using in this paper, it is worth noting that several authors, such as Arrow (1978) and Sen (1970), use the term 'extended sympathy' to refer rather to the identification mechanism underlying extended preferences. But in this context, they use the term 'sympathy' in its old meaning, which effectively makes it synonymous with the modern meaning of empathy.

1973). Harsanyi (1996) himself recognized that not all preferences are conducive of personal welfare and that antisocial and self-detrimental preferences should be ignored by the welfare analysis. This casts doubts on the normative strength of the sovereignty principle highlighted above. The second difficulty is even more significant. The empathetic reading of extended preferences seems to indicate that an individual i evaluating social alternatives from j 's point of view should completely identify with j 's preferences. However, if asked to compare two extended alternatives featuring two different individuals j and k , it is not clear how i should proceed. Endorsing successively j 's and k 's preferences is presumably not sufficient because, as such, it does not make them comparable. Hence, it seems clear that extended preferences cannot merely replicate each individual's preferences. They are the preferences of the individual who is making an assessment between two extended alternatives. Therefore, it is not possible to avoid the following question: Where do extended preferences come from?

The difficulties that surface with respect to the second issue regarding the uniformity of extended preferences across a population are directly related to the impossibility to answer this question in a satisfactory way. As a first step, it is useful to remark that the uniformity of extended preferences (formally, $R_i^E = R_j^E$ for all pairs of individuals i and j in the population) has been sometimes assumed to derive interesting formal results from a social choice perspective.⁸ What is at stakes here, however, is not its theoretical usefulness but its normative relevance. The reason why uniformity is generally thought to be required is that, in its absence, individuals' comparative assessments of extended alternatives would differ, leading to several extended utility functions u_i^E . But then, we would be back to square one as we would be devoid of any way to compare these extended utility functions: each individual would make her own interpersonal comparisons, but disagreement over the right way to make them would ensue. If it is assumed that a *collective* choice cannot be the choice of one person, there are only two possibilities: either everyone agrees on a common standard to compare utilities or only social welfare functions not requiring interpersonal comparisons are acceptable. However, though several arguments have been offered in the social literature to ground the uniformity principle (e.g., Arrow 1978; Binmore 1998; Harsanyi 1977),

⁸ For instance, Sen (1970, Chapter 9*) makes the uniformity assumption to establish a formal relationship between Suppes' grading principles of justice and aggregative and Rawlsian social welfare functions.

none of them has generally been regarded as convincing for several reasons.⁹ The problems with the uniformity principle emphasize that the quest of grounding interpersonal comparisons and extended preferences from an objective, valueless point of view is hopeless. Disagreement over the standards for making interpersonal comparisons seems inescapable, putting the social choice approach to equity and justice issues under pressure.

III. INTERPERSONAL COMPARISONS AND ‘TELEOLOGICAL CONTRACTUALISM’

The discussion of the previous section therefore calls for an essentially negative conclusion: there seems to be no convincing argument establishing that extended preferences must be uniform while keeping their normative meaning. This makes the prospects of a pure social choice approach to equity and justice unlikely, because without interpersonal comparisons almost nothing can be said about equity and justice within this framework. This conclusion may leave social contract theorists indifferent: so much the worse for the social choice approach, but as a rival tradition social contract theory is unaffected.

I shall argue in this section that this conclusion is too quick. The idea that social contract theorists may spare themselves the need to deal with interpersonal comparisons is due to the conflation of two related but still distinct philosophical underpinnings of the social contract tradition. On the one hand, the social contract tradition is constituted by a (set of) first-order moral doctrine(s) that are loosely referred to as ‘contractualism’. Though there is no widely agreed definition, contractualism may be characterized as the general view that morality is based on contract or agreement (Ashford and Mulgan 2018; Gauthier 1977). On the other hand, the social contract tradition may be viewed as a (set of) normative account(s)

⁹ Harsanyi’s so-called causal argument is the one that has been given the most attention and has attracted most of the criticisms. The most forceful criticism comes from Broome (1993) who shows that Harsanyi’s argument confuses the cause for preferences (the $C(i)$ variables) with the content of preferences. While impartial observers may agree over the causes for different individuals’ assessments of social alternatives, this does not logically imply that they must agree over their impartial assessments of extended alternatives. Mongin (2001) also highlights that Harsanyi’s argument would at best support the claim that all impartial observers agree over their *predictions* of which extended alternative brings the highest degree of preference satisfaction. Finally, Pattanaik (1968) observes that there is absolutely no reason to expect that impartial observers’ attitude toward risk should be identical. As the utility values of extended alternatives are directly derived from preferences over *lotteries*, that implies that two impartial observers may ascribe different utility values to extended alternatives, even if the causal argument is true.

of *justification*, especially of moral justification. Most social contract accounts will then be characterized as subscribing to a ‘deontological’ account of justification.¹⁰ Each of these labels has its natural nemesis, ‘consequentialism’ and ‘teleology’ respectively. I will not have much to say about the contractualism/consequentialism opposition, except that social choice theory is generally classified within the consequentialist category. This is not only or foremost due to the fact that social choice theorists tend to study equity and justice issues in terms of *choice* rather than in terms of *contract* (or another procedure leading to an agreement).¹¹ It is mostly related to the fact that within social choice theory the evaluation of collective choices is fully dependent on a more or less broad characterization of *outcomes* rather than on the fact of instantiating or following rules and principles. This point is, however, not relevant for the deontological/teleological distinction. I shall indeed argue that even in the realm of contractualism as a first-order moral doctrine, a teleological form of justification may be relevant and that this calls for the possibility of making comparative judgments of goodness. In this regard, Gaus—following Sandel (2010, 3)—characterizes teleology as:

A form of justification in which first principles are derived in a way that presupposes final human purposes and ends. [...] Principles of right, or public morality, are justified through appeal to values of actual people to whom the morality is to apply. (Gaus 1990, 331)

¹⁰ To my knowledge, the explicit distinction between contractualism as a first-order moral doctrine and deontology as a theory of moral justification is due to Sandel (2010). I shall point out that the characterization of the deontology/teleology distinction as competing accounts of justification I use in the text is not the most common in the literature. In particular, it is clearly not how Rawls defined these terms (see Freeman 1994). Consequentialism and teleology indeed tend to be used as quasi-synonymous (though, according to Freeman, ‘mixed’ forms of consequentialism are more appropriately seen as belonging to deontology (2007, Chapter 3)). In the same way, almost all forms of contractualism are generally regarded as belonging to the domain of deontology. Though the terminology may be unorthodox, I still think that distinguishing between normative accounts of justification (Sandel’s and my deontology/teleology distinction) on the one hand, and first-order moral theories (the contractualism/consequentialism distinction) on the other hand, is enlightening because it helps to separate two related but still different issues: first, what is it for an act or a state of affairs to be good or right? Second, what is the relationship between the good and the right in the justificatory endeavor?

¹¹ On the choice/contract distinction, see Hampton (1980). Though her article focuses on Rawls and the possible interpretations of his theory of justice in terms of the choice paradigm or the contract paradigm, she also notes that Harsanyi’s ‘contractualism’ belongs to the choice paradigm. As noted by Gaus and Thrasher (2015), the fact that Rawls’ contractualism depends on a choice from an ‘Archimedean point’ rather than a proper contract is not peculiar to it, as other forms of contractualism like Gauthier’s actually share this same feature.

As we shall see below, teleological justification can take many forms. By contrast, deontology is then a mode of justification of first-order normative principles that does not presuppose ultimate human purposes and ends, i.e., no definite conception of the good.

The deontological underpinnings of the social contract approach are well exemplified by Rawls (1971) in his theory of justice. Rawls repeatedly emphasizes that a theory of justice must recognize the priority of the 'right' over the 'good'. That is, principles of justice determining what is right and just in the society can and must be determined independently of any conception of goodness and thus of any view regarding what a 'good' life is. In this sense, the right not only has priority over the good, the former also constrains the latter: only conceptions of goodness compatible with the chosen principles of justice will be able to prosper within the society. On the other hand, this 'axiological neutrality' characteristic of a deontological theory of justice is also deemed to be compatible and even to favor a 'reasonable' moral pluralism. This feature is especially developed by Rawls in his later writings which emphasize that his theory is foremost 'political' rather than 'moral' (Rawls 1993, 2001). In a 'well-ordered society', an overlapping consensus will prevail through which citizens affirm a unique political conception of justice, while entertaining conflicting religious, philosophical, and moral views. In particular, the agreed upon principles of justice are endorsed *from within* competing comprehensive moral doctrines:

We say that in a well-ordered society the political conception is affirmed by what we refer to as a reasonable overlapping consensus. By this we mean that the political conception is supported by the reasonable though opposing religious, philosophical, and moral doctrines that gain a significant body of adherents and endure over time from one generation to the next. This is, I believe, the most reasonable basis of political and social unity available to citizens of a democratic society. (Rawls 2001, 32)

There is absolutely no doubt that the priority of the right over the good is an enduring feature of Rawls' account of justice, thus establishing its deontological character. That said, Rawls' later writings also indicate a growing concern for establishing that his political theory of justice is compatible with the *fact* of moral pluralism that is constitutive of modern societies. It is true that that this concern is essentially due to empirical, rather than normative reasons. Moral pluralism and hence moral disagreement are facts with which we have to live and any practically relevant

theory of justice must not only recognize it but also be compatible with it. With respect to the logic of justification, the right still comes first: we must determine and justify political principles of justice without appealing to any feature of a comprehensive doctrine, being moral, religious or anything else. But from an empirical point of view, principles of justice cannot be but supported from within comprehensive doctrines, hence the requirement for a well-ordered society of establishing an overlapping consensus.

The growing recognition by Rawls of the practical importance of reasonable pluralism and the related turn of his justice as fairness account from a 'metaphysical' or 'comprehensive' to a 'political' understanding also involves a change in the interpretation of his thin theory of the good. Contra Sandel (2010) and communitarian critics of Rawls, Rawls' contractualism in *A Theory of Justice* is not fully deontological as it builds on a conception of goodness as rationality that is attributed to the parties in the original position. As Rawls (1971, 396) made it clear, this notion of goodness *precedes* the establishment of the principles of justice by individuals put behind a veil of ignorance. Its role is to ground the assumptions made about the primary goods that all individuals are assumed to pursue to realize their rational plans and *full* conceptions of the good. As a consequence, though 'thin', this account of the good introduces a teleological feature in the otherwise deontological and constructivist procedure of justification developed by Rawls. The thin theory of the good is pivotal in Rawls' account at least at two levels. First, it allows for the creation of an index of primary goods which itself provides a basis for interpersonal comparisons (Rawls 1971, 92). The latter are indeed required to make the first part of Rawls' second principle of justice (the difference principle) operational and meaningful. Second, it plays an essential role in Rawls' complicated account of stability developed in Part III of *Theory of Justice*. Rawls' congruence argument between the right and the good indeed builds on the claim that maintaining a sense of justice is a good in the sense of the thin theory. Rawls however, later rejected the congruence argument as being incompatible with a political theory of justice and accounted for stability in terms of an overlapping consensus. The interpretation of the thin theory of the good has subsequently changed. Goodness as rationality is no longer a plausible account of a person's objective good. It rather refers to a pluralist conception of value compatible with political principles of justice as agreed between persons that mutually regard themselves as free and equal citizens (Freeman 2007, 97–98).

This brief survey of the evolution of Rawls' contractualism indicates that teleological considerations were present from the start to justify and operationalize principles of justice in the context of reasonable pluralism. It also serves as an intermediary step to establish the *normative* relevance of a teleological form of justification for contractualism. Rawls emphasized the fact of moral pluralism essentially for empirical and political reasons. I now want to argue that contractualists may want to go further and deal with interpersonal comparisons in the context of teleological justification for more foundational reasons. Rawls' contractualism has at least three components that account for the secondary role played by teleological justification: first, a particular kind of constructivism that materializes through the original position; second, the specific account of justice consisting in its two principles; third, its 'formal' conception of the person. Modifying one or several of these components may make room for a more important role for teleological justification in contractualism.¹² In the following, I will focus on the last component, but first I comment on the former two.

The derivation of the two principles of justice through the device of the original position is characterized in terms of a 'procedure of construction' establishing a link between a particular (political) conception of the concept and principles of justice (Rawls 1980, 304). This construction builds on the conception of moral persons as free and equal citizens who, because of this very conception, are committed to search for principles of justice while ignoring their own conception of the good—except for the thin theory of the good which justifies the reasoning in terms of primary good. Now, there is room for disagreement regarding the content of the principles of justice, something which Rawls increasingly emphasized in the last part of his career. Constructivism is also itself not the only way to justify principles of justice in a contractualist framework. Even within constructivism, we may imagine a different procedure of construction where persons are aware of their conceptions of the good. Ultimately, I submit that the issue of justification is tightly related to the *theory of the person* that one sees as appropriate. To defend a full theory of this kind is of course a daunting task that I cannot undertake here. Let me, however, sketch an argument indicating that even within a contractualist

¹² These three components are of course tightly related in Rawls' writings and so, changing one may affect the other two. Rawls' article on Kantian constructivism (Rawls 1980) provides the clearest statement of the relationship between his conception of the person and the constructivist nature of his contractualism.

account of morality there is scope for a theory of the person, one which demands a more significant role for teleological justification.

Though somewhat rough, we may distinguish between two broad accounts of the person in a contractualist perspective. A first account characterizes personhood independently of the ends, goods, and values that particular persons may contingently endorse or pursue. This ‘formal view’ attributes to persons agency powers and capacities that make them rational and reasonable beings, but takes no stance with respect to how these powers and capacities are actually used. This is not to mean that individuals do not endorse values or do not pursue ends, but rather that what makes these individuals foremost moral persons endowed with a particular normative status is not their values or ends, but their *ability* to pursue ends and endorse values. This formal view finds notably its expression in Rawls’ *Theory of Justice*. The normative relevance of the original position is fully dependent on the possibility of decoupling persons from their contingent conceptions of the good life. This postulate remains at the core of Rawls’ later writings. For instance, Rawls repeatedly emphasizes that his conception of the person as citizens is a ‘political’ one that gives persons two “moral powers”:

- i. One such power is the capacity for a sense of justice: it is the capacity to understand, to apply, and to act from (and not merely in accordance with) principles of political justice that specify the fair terms of cooperation.
- ii. The other moral power is a capacity for a conception of the good: it is the capacity to have, to revise, and rationally to pursue a conception of the good. (Rawls 2001, 18–19)

As it is well known, Rawls also emphasizes the ‘separateness of persons’ as a normatively relevant feature. This separateness is precisely grounded in the fact that each person has their *own* capacity for a sense of justice and for a conception of the good. But this separated identity is not grounded on the *content* of this sense of justice and this conception of the good.

Sandel (2010), among others, has emphasized that such a formal view of the person is almost needed by deontologists. This is by decoupling persons from their ends and values that the claim of the priority of the right obtains its normative force. As I have argued, that does not imply, at least in the case Rawls’ contractualism, a complete lack of teleological elements. However, there is at least an alternative plausible view of the person that puts the priority claim in jeopardy. What can be called the

‘substantive view’ conceives the person, both in her rational agency and her identity, as being fundamentally constituted by the values she endorses and the ends she pursues. Another way to characterize this view is that what makes persons morally separated and relevant is that values and ends are *theirs*: they are able to act and to justify their actions on the basis of values that they recognize as their own. Obviously, the substantive view also regards Rawls’ moral powers as normatively important. But it goes further: we *also* give importance to persons as bearers of particular values constitutive of a plurality of conceptions of goodness. On the substantive view, the rationality of persons cannot be characterized independently of the values that justify their intentional attitudes. Moreover, the identity of persons is tightly related to their (possibly) evolving conception of goodness. This implies that the *content* of conceptions of the good cannot be arbitrary from a normative perspective: what makes a person a rational or reasonable being is their ability to act on the basis of *some* values; what makes a person a continuous being is their ability to endorse a *particular* conception of goodness. That implies that some persons may lose their particular normative status if their conceptions of goodness are constituted by what is judged to be unacceptable values or if it is impossible to ascribe to them some continuous and consistent conception of the good.¹³

The formal and the substantive views have quite different implications regarding the relationship between what can be called the ‘personal point of view’ and the ‘moral/political point of view’. On both accounts, a person’s personal point of view is constituted by their conception of the good. But as far as the moral/political point of view is concerned, the formal view insists that the choice of principles of justice should be detached from particular conceptions of the good going beyond the thin theory, resulting in the priority of the right over the good. The substantive view not only indicates that the moral/political should not be expected to be completely independent from conceptions of the good, but that it ought not to be. The point is not only that, as a matter of practical rationality, it is impossible to choose a set of political principles in *complete ignorance* of one’s own values; it is also that such a choice would be normatively irrelevant. Now, what sets the moral/political point of view apart is that when reflecting on the appropriate principles that should regulate

¹³ That would not mean that these persons have no normative importance, but rather that this importance would be grounded on a different normative reason (e.g., as sensible beings capable of enduring pain).

a society, each person must acknowledge that she has to find an agreement with other persons bearing their own conceptions of the good. That implies at least two things: first, a substantive assessment of competing conceptions of the good may be needed; second, establishing a minimal common conception of the good or a compromise between competing conceptions may be required by invoking values (e.g., equality, impartiality) that may not transpire in personal conceptions.

Hence, *on the substantive view of personhood*, agreement over a social contract will require one form or another of *teleological justification* building on a compromise between competing conceptions of the good and/or on a common conception of the good. As pointed out by Gaus (1990), this teleology will still be constrained by deontological principles and values that may find their justification in the formal view of personhood.¹⁴ But the point is that, under this conception of personhood, the social contract cannot avoid any form of teleological justification. Once this point is granted, another question arises: How could this justification be achieved? This question marks a point of departure in the social contract tradition between *contractarian* and *contractualist* (in a narrow sense) approaches. The former argues that the social contract is ultimately the product of a compromise between competing conceptions of goodness. Significant instances of the contractarian approach such as Gauthier's (1987) make use of axiomatic bargaining theory to ground the compromise on bargaining principles. Interestingly, they tend to use bargaining solutions that eschew the need to make interpersonal comparisons of goodness. The contractarian approach is, however, notoriously controversial, and I will not discuss it there. Apart from the contractarian approach to teleological justification in terms of compromise, another possibility is a contractualist approach working through the identification of a common conception of goodness. The next section establishes that such an identification must rely on interpersonal comparisons and develops a particular account in terms of 'extended goodness'.

IV. SYMPATHETIC IDENTIFICATION AND EXTENDED CONCEPTIONS OF THE GOOD

Reflecting on the role of teleological justification in a contractualist framework, Gaus (1990) appeals to Gauthier's (1987) contractarianism

¹⁴ As I have pointed out, the substantive view encompasses the formal view, i.e., regarding a person as being constituted by her conception of goodness obviously implies the consideration that a person has the *ability* to form a conception of goodness.

which, with respect to justification, has two key features: on the one hand, values appealed to are *agent-relative values*, on the other hand the justification takes the form of a *compromise* between competing conceptions of goodness. As I indicate in the preceding section, this compromise is shaped by a bargain that—at least in Gauthier’s case—can be characterized through axiomatic bargaining game theory. Quite a different form of teleological justification is provided by some variants of utilitarianism, including Harsanyi’s. These variants appeal to *agent-neutral* values from which a *community of valuing* is derived. The teleological justification of this brand of utilitarianism can be seen as an instance of Nagel’s (1989) ‘view from nowhere’ account, according to which the objectivity and neutrality of the moral point of view is achieved by adopting a perspective free from any individual contingencies and idiosyncrasies. Harsanyi’s impartial observer theorem is indeed an instance of such an account: by ignoring who and where they will end up in the society, individuals behind the veil of ignorance are forced to adopt a set of values that, though they reflect the values (i.e., preferences) of everyone, are the values of nobody in particular. These values are captured by the observers’ extended preferences. As we have seen, Harsanyi claims—wrongly—that these extended preferences must be uniform across the population, thus achieving a community of valuing.

I shall make the case for another form of teleological justification. While the values constitutive of the persons’ conceptions of goodness are the ones appealed to in the justificatory endeavor, I am agnostic with respect to their agent-relativity or neutrality.¹⁵ Presumably, individuals’ conceptions of the good will essentially be constituted of (prudential and non-prudential) values providing agent-relative reasons for action (e.g., personal welfare, dignity, autonomy). Yet, we cannot exclude the possibility that people value things leading to agent-neutral reasons for action (e.g., overall welfare). Whether or not such values should be considered in the justificatory endeavor is an issue that I leave aside. On the other hand, neither Gauthier’s compromise nor Harsanyi’s community of valuing are satisfactory views of teleological justification. The former because it makes the resulting compromise depend on bargaining factors whose normative relevance are doubtful from the moral point of view; the latter because it depends on the mistaken assertion that extended preferences must be uniform. Instead, I suggest that teleological justification should

¹⁵ For a discussion of the distinction between agent-neutral and agent-relative values and reasons, see Parfit (1984).

proceed on the basis of the identification of a *minimal* common conception of the good. In other words, a form of community of valuing is needed but it cannot be expected to be a complete overlapping of personal conceptions of goodness. Crucially, this minimal common conception of the good requires the use of interpersonal comparisons. The rest of this section is dedicated to developing an account of interpersonal comparisons of goodness in this context.

The formal framework of extended preferences presented in section II will be useful here. However, while the formalism remains, its interpretation must be quite radically changed. First, in the framework of extended preferences, the binary relations R_i and R_i^E and the related utility functions u_i and u_i^E were thought to correspond to and represent, respectively, a *preference* relation. This implies a commitment to preferentialism which may be regarded as problematic given the already mentioned difficulties surrounding the preference concept in a normative context. We may, however, ignore such commitment by taking the binary relations and the functions to capture personal conceptions of goodness. To avoid any confusion, I will denote B_i the binary relation ‘as better or as good as’ and g_i the related ‘goodness function’ of individual i . This entails a set of assumptions about the formal properties of personal conceptions of goodness. In essence, I am assuming that individuals can rank social alternatives on the basis of their conceptions of goodness. Call it the ‘ordering property of goodness’. This first change entails a second one: we no longer deal with extended preferences but rather with ‘extended goodness’, or more precisely *extended personal conceptions of the good*. They are captured by the binary relations B_i^E and the ‘extended goodness functions’ g_i^E . As in the case of extended preferences, the relations B_i^E are defined over the Cartesian product $X \times N$. It remains to establish the precise meaning of these relations. A third and final change is the nature of the process through which extended conceptions of goodness are arrived at by individuals. As I note in section II, social choice theorists have interpreted extended preferences in terms of empathetic identification. I shall suggest instead that extended personal conceptions of the good are formed through a process of *sympathetic identification*. I will detail and argue for these three major changes, starting with the last one.

As I explain in section II, empathetic identification consists in taking another person’s point of view, i.e., to adopt all her affective and evaluative states and attributes in some given situation. This approach has been argued to be problematic, in particular with respect to the very feasibility

of empathetic identification in some cases. Adler (2014) makes this objection, which he labels the ‘essential-attribute problem’. This problem can be briefly put in the following way. Suppose that any social alternative $x \in X$ includes information about individuals’ attributes that, under some theory of personal identity, may be referred to as ‘essential’, i.e., as being constitutive of a person’s identity. For instance, plausible theories of personal identity may hold that gender or memories are such essential attributes.¹⁶ Now, the essential-attribute problem is simply the point that when social alternatives include such information, it might be impossible—in quasi-phenomenal terms—for an observer i to assess an extended alternative (x, j) because ‘being j ’ depends on attributes that i cannot even imagine owning. In other words, empathetic identification confronts the problem that two persons i and j may not be able to fully *share* their respective points of view. Adler makes this objection in a preferentialist framework, but it is not hard to see that it applies at least equally strongly here. In the preceding section, I remarked that under a substantive view of personhood, one’s conception of the good is constitutive of one’s personal identity. That does not mean that conceptions of the good are not shareable in total, but that imaginative projections of the kind ‘I assess social alternative x taking j ’s values as mine’ may in some cases be simply impossible.

Instead of empathetic projection, Adler (2014) suggests a ‘sympathy-based conception of extended preferences’. Following Darwall (2002), Adler defines sympathy as an attitude of care and concern for others. In particular, to fully sympathize with someone is to be “motivated to pursue what you believe lies in the interests of the sympathy target” (Adler 2014, 146). This definition implies an important connection between sympathy and *welfare*. Take the specific case where person i is asked to take a moral point of view (i.e., to endorse the position of an ‘impartial observer’) and to compare two extended social alternatives (x, i) and (y, i) . i is thus asked to ‘self-sympathize’ with herself. Under Adler’s and Darwall’s account, that means that i must assess each extended alternative uniquely in terms of her self-interest, i.e., her welfare. Similarly, when comparing two extended alternatives concerning the *same* person j who is not her, the observer who fully sympathizes should base her evaluation entirely on j ’s welfare. Finally, when comparing two extended alternatives

¹⁶ Alternatively, we may assume that these essential attributes are attached to the components of the individuals set N . What should be avoided is to postulate that essential attributes can be found both in X and N , as that would make some extended alternatives (x, i) implausible if not metaphysically impossible.

concerning two different persons j and k , the sympathetic observer should form a welfare judgment and determine whose welfare is higher. This approach has two advantages but also two implications that may be regarded as undesirable. The first advantage is that the sympathy-based account of extended preferences eschews a general problem with the ‘view from nowhere’ approach to morality and more generally to normativity. It is clear here that the observer’s welfare judgment cannot be based but on her *own conception of welfare as part of the good*. Because presumably most if not all individuals value welfare, sympathetic identification is at least well defined. This leads to the second advantage: this account is not confronted to the essential-attribute problem. The sympathetic observer is not asked to put herself in others’ shoes but to make comparative welfare assessments *from her own perspective*. Shareability is thus not an issue.

Two other implications should also be considered as they may be judged problematic. On the one hand, there is obviously no guarantee that all observers will agree on their comparative welfare judgments. This is due to several factors: they may not have all relevant information at their disposal and some may be better informed than others, they may not share the same conception of welfare, or they may disagree about the welfare-effects of certain essential attributes related to personal identity. We can assume that perfectly informed observers may eliminate the first factor; but the second and third ones seem unavoidable. However, as neither we (nor Adler for that matter) are looking for uniformity of extended preferences, this problem is not relevant here. The second potentially problematic implication is that the sympathetic approach gives too much priority to welfare over any other value (prudential or not prudential) when persons take the moral point of view. There is no doubt that welfare is an important value and that any normative endeavor has to acknowledge its importance. But from the perspective of moral and political philosophical theories, it may be argued that teleological justification cannot be based *fully* on it.¹⁷ Take the following plausible case: Emma is a young adult who takes pride in being an ecological activist, even if participating in street protests may occasionally result in her being arrested or hurt. John, a forty-something bank employee and father may judge, in wholehearted sympathy with Emma, that Emma would be better—in

¹⁷ It should be acknowledged that Adler (2014) develops his account mostly as a contribution to welfare economics, not to moral and political philosophy. The objection discussed in the text cannot thus be directly addressed to him.

terms of welfare—in a social alternative x where she renounces her ecological activism than in a social alternative y where she ends up being severely hurt. Suppose that, from Emma's conception of goodness, the converse judgment applies. Formally, we thus have $yB_{Emma}x$ but $(x, Emma)B_{John}^E(y, Emma)$. It is clearly debatable that John's judgment should have priority over Emma's. The point is that, except for the mostly unlikely case where Emma does not value welfare at all, her goodness judgment already encapsulates a tradeoff between welfare and other values, including non-prudential ones. With regard to providing Emma (among others) a teleological justification to a set of normative assessments about what is good, just, permissible, or obligatory at the collective level, it is not clear that John or other observers should be authorized to simply disregard Emma's tradeoff. There are only two possibilities here: either it is established that Emma is *wrong* in her goodness judgment, either because she is not correctly informed or she has not properly reasoned about the issue, or sympathetic identification should not simply consist in taking care of others' welfare but also of other's good as a whole. The former option leads to paternalistic considerations which should not be straightforwardly rejected on 'naïve' libertarian grounds. But whatever one may think of paternalism, the latter option should also be given due consideration.

The revision of the concept of sympathetic identification in terms of taking care of others' good entails what I call a concept of 'extended goodness', or more precisely, of *extended personal conceptions of the good*. Consider the meaning of the statement $(x, j)B_i^E(y, j)$. Literally, it reads as 'i judges x to be better than or as good as y for j '. I have just argued that i 's goodness judgment cannot be made only on the basis of a concern for j 's welfare. Other values should presumably also be taken into consideration. But the statement remains ambiguous: Is i judging that x is better than or as good as y for j on the basis of i 's conception of the good? Or is i fully endorsing j 's conception of the good in making his goodness judgment? I would argue that neither is acceptable. The former interpretation may be compatible with sympathetic identification but only on the formal view of personhood where personal identity and personal conceptions of the good are decoupled. But under the substantive view, the revised concept of sympathetic identification renders it implausible. The latter interpretation is more plausible but would imply that i invariably defers to j 's conception of goodness. Even if it is accepted, this interpretation is not available in comparative judgments involving two different persons, i.e.,

statements of the form $(x, j)B_i^E(y, k)$. I thus submit a third interpretation: in forming her extended goodness judgment through sympathetic identification, the observer should take others' conceptions of the good *as far as they include values that the observer herself endorses in her own conception*. The observer and other individuals may still differ with respect to the *weight* they give to these values, but they agree that these values matter. Thus, they could and should be appealed to in the justificatory endeavor. In the case of statements of the kind $(x, j)B_i^E(y, j)$, i 's extended goodness judgment should thus be based on those values that have a positive weight in i 's and j 's respective conceptions of goodness. Obviously, j 's personal goodness judgment and i 's extended goodness judgment may differ, in particular if their respective conceptions of goodness do not fully overlap. The same idea holds, perhaps even more convincingly, in the case of a statement of the form $(x, j)B_i^E(y, k)$. Unless j and k share the same personal conceptions of the good, i as an observer may have to ignore some values. The observer is more likely to justify her judgment to j and k by grounding it on values that j and k actually share. At the same time, because i proceeds through sympathetic rather than empathetic identification, she may consider taking into account only those values that she is herself actually endorsing.

Consider a population of n persons, each endowed with a personal conception of the good. These n individuals also are n potential observers who may form extensive goodness judgments. Suppose that we may identify a set V of (prudential and non-prudential) values v that are weighted positively by at least one person. Denote V^* the (maximal) subset of V such that any $v \in V^*$ is common to *all* personal conceptions of the good. V^* forms the basis on which extended goodness judgments can take shape, though values which are only partially common (i.e., they are shared only by a subset of the population) may also be used in specific cases. Two issues then arise. First, what is the relationship between binary relations B_i and B_i^E capturing respectively personal and extended goodness judgments? The two binary relations should obviously be identical when the observer i is comparing a pair of extended alternatives (x, i) and (y, i) . In the more general case where a deliberator k is comparing a pair of extended alternatives (x, i) and (y, j) , the deliberator's personal goodness judgments captured by the binary relation B_k is partially relevant because by assumption it restricts the set of values on the basis of which she forms her extended goodness judgments. But among this restricted set, only a (possibly empty) subset of values will be shared by i 's and j 's

personal conceptions of the good. The relation B_i^k compares extended alternatives on the basis of this subset of values that are shared by all protagonists. How then will a deliberator rank social alternatives from the moral point of view? The deliberator must *justify* her ranking to everyone on the basis of a sequence of extended comparative goodness judgments captured by B_i^E . But this is obviously not sufficient because the deliberator must also find a way to aggregate these extended goodness judgments. A person's all-things-considered moral ranking is then captured by a binary relation B_i^* defined over X (and *not* the Cartesian product $X \times N$).¹⁸ B_i^* is determined both by B_i (since it restricts the range of values that i uses to form her judgment) and by B_i^E . The reason why the latter should be an input in the moral ranking is twofold. First, it is an essential part of the justificatory endeavor. They permit each person to make interpersonal comparative judgments which may be used as an input in the justificatory endeavor. Indeed, they guarantee that all B_i^* are based on a minimal common conception of the good. Second, I would suggest that, at least in what Rawls (1971) calls a 'well-ordered society' based on relationships of reciprocity and mutual respect, persons have strong normative reasons to respect others' conceptions of goodness and to engage into sympathetic identification. Indeed, it might be argued that reciprocity, respect, and tolerance are core values of a well-ordered society that tend to be shared by all the members of the population. This provides reasons to consider and even to value others' conceptions of the good, at least as far as issues requiring public justification are concerned.

That said, and this is the second issue that may be mentioned, there is no reason to expect that everyone will agree on their moral ranking, i.e., that $B_i^* = B_j^*$ for all i, j . On the one hand, the fact that the binary relations B_i^* build on a common set of values does not imply that all persons will make the same tradeoffs between these values. This is just another version of the claim that persons' extended goodness judgments B_i^E will differ across the population. However, though not identical, the B_i^E relations should be regarded as partially comparable. This is due to the fact that they build on the same set of values and the same process of sympathetic identification. If this set is sufficiently large, we may expect an agreement

¹⁸ The binary relation B_i^* corresponds to what Harsanyi (1977) called a person's *moral preferences*. In Harsanyi's account, moral preferences are formed by aggregating extended preferences through a utilitarian formula that is itself derived on the basis of particular rationality and epistemic assumptions. I do not presume here any specific aggregation rule. Indeed, as I indicate below, the fact that deliberators may use different aggregation rules is one source of moral disagreement.

over a significant number of pairs of social alternatives. This minimal agreement over goodness judgments shares conceptual similarities with Rawls' thin theory of the good. There are differences though. The most significant one is that the content of this agreement is left undetermined and so that the goodness judgments may differ from one population to another. The second difference is that this agreement does not follow from a practical or empirical necessity but is more foundational. In this sense, teleological contractualism is fully 'comprehensive' rather political in Rawls' sense.

On the other hand, even if all deliberators were to agree on their extended goodness judgments, there is no reason to expect that they would necessarily agree on the way to aggregate these judgments to form their moral rankings. The likelihood of an agreement on this issue and the principles that would serve as a basis for it is another important aspect of the problem of moral disagreement that is, however, beyond the scope of this paper. This point underlines, however, the interest of partially building contractualism on teleological foundations. Sources of moral disagreement are manifold and can result both from teleological and deontological considerations. To be able to account for this fact seems to be a valuable asset within the social contract tradition.¹⁹

V. CONCLUSION: THE MANY WAYS OF DISAGREEING OVER A MORAL CODE

In conclusion, I would like to insist on the point made just above. I have claimed in this paper that the need and the difficulties surrounding interpersonal comparisons are not an artefact of social choice theory applied to issues of justice and equity. It is true that social choice theorists *do need* to make such interpersonal comparisons to show that individuals agree on such comparisons. The framework of extended preferences, though useful to account for the origins and the meaning of interpersonal comparisons, cannot establish what the social choice theorist needs, i.e., uniformity of extended preferences across a population. Now, I have argued that social contract theorists, especially those who wanted to avoid contractarianism, must also be able to account for interpersonal comparisons. This is due to the fact that a form a teleological justification is

¹⁹ While initially emphasizing the importance of teleology, Gaus' (2012) contractualism has now given up any explicit reference to the role of competing conceptions of goodness in moral disagreement. Though I cannot defend this claim here, I think that on this aspect, Gaussian contractualism—while otherwise extremely valuable and important—is making a step backward as compared to Rawlsian political liberalism, which emphasizes and builds on the disagreement over the good.

needed, even within a contractualist approach. I have suggested an approach in terms of sympathetic identification and extended goodness that falls short on establishing that individuals should agree on their judgments. Disagreement over judgments of goodness is, however, not problematic in a social contract perspective as long as individuals are in general agreement over the rules to make collective choices.

Teleological justification is indeed only a part of the justificatory endeavor. Moral pluralism, i.e., the fact that individuals disagree over their conceptions of goodness, can coexist with an overlapping consensus over core values. But individuals should also agree over principles that, *given prevailing conceptions of the good*, rule collective decision-making. One reason Rawls gave priority to the right over the good was precisely his belief that ‘reasonable pluralism’ was possible as far as conceptions of the good are concerned, but not in the case of the right. But disagreement over the right is also an important fact of modern societies. I have assumed in this paper that the two forms of justification (teleological and deontological) can be tackled quite independently from each other. This is a simplification, however. Principles about the right also build on values that should be endorsed by individuals taking the moral point of view. Further reflections on the relationship between teleological and deontological justification are required to account for the fact of moral disagreement in modern societies.

REFERENCES

- Adler, Matthew D. 2014. “Extended Preferences and Interpersonal Comparisons: A New Account.” *Economics and Philosophy* 30 (2): 123–162.
- Arrow, Kenneth J. 1963. *Social Choice and Individual Values*. New Haven, CT: Yale University Press.
- Arrow, Kenneth J. 1978. “Extended Sympathy and the Possibility of Social Choice.” *Philosophia* 7 (2): 223–237.
- Ashford, Elizabeth, and Tim Mulgan. 2018. “Contractualism.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Article published August 30, 2007; last modified April 20, 2018. <https://plato.stanford.edu/entries/contractualism/>.
- Binmore, Kenneth G. 1998. *Game Theory and the Social Contract, Volume 2: Just Playing*. Cambridge, MA: The MIT Press.
- Broome, John. 1993. “A Cause of Preference Is Not an Object of Preference.” *Social Choice and Welfare* 10 (1): 57–68.
- Darwall, Stephen. 2002. *Welfare and Rational Care*. Princeton, NJ: Princeton University Press.
- Fleurbaey, Marc. 2003. “On the Informational Basis of Social Choice.” *Social Choice and Welfare* 21 (2): 347–384.

- Freeman, Samuel. 1994. "Utilitarianism, Deontology, and the Priority of Right." *Philosophy & Public Affairs* 23 (4): 313-349.
- Freeman, Samuel. 2007. *Justice and the Social Contract: Essays on Rawlsian Political Philosophy*. New York, NY: Oxford University Press.
- Gaus, Gerald F. 1990. *Value and Justification: The Foundations of Liberal Theory*. New York, NY: Cambridge University Press.
- Gaus, Gerald F. 2011. "Social Contract and Social Choice." *Rutgers Law Journal* 43: 243-276.
- Gaus, Gerald F. 2012. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Reprint edition. New York, NY: Cambridge University Press.
- Gaus, Gerald F., and John Thrasher. 2015. "Rational Choice and the Original Position: The (Many) Models of Rawls and Harsanyi." In *The Original Position*, edited by Timothy Hinton, 39-58. Cambridge: Cambridge University Press.
- Gauthier, David. 1977. "The Social Contract as Ideology." *Philosophy & Public Affairs* 6 (2): 130-164.
- Gauthier, David. 1987. *Morals by Agreement*. New York, NY: Oxford University Press.
- Hampton, Jean. 1980. "Contracts and Choices: Does Rawls Have a Social Contract Theory?" *The Journal of Philosophy* 77 (6): 315-338.
- Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61 (5): 434-435.
- Harsanyi, John C. 1977. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.
- Harsanyi, John C. 1996. "Utilities, Preferences, and Substantive Goods." *Social Choice and Welfare* 14 (1): 129-145.
- Hausman, Daniel M., and Michael S. McPherson. 2006. *Economic Analysis, Moral Philosophy and Public Policy*. 2nd edition. New York, NY: Cambridge University Press.
- Mongin, Philippe. 2001. "The Impartial Observer Theorem of Social Ethics." *Economics & Philosophy* 17 (2): 147-179.
- Nagel, Thomas. 1989. *The View From Nowhere*. New York, NY: Oxford University Press.
- Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.
- Pattanaik, Prasanta K. 1968. "Risk, Impersonality, and the Social Welfare Function." *Journal of Political Economy* 76 (6): 1152-1169.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Rawls, John. 1980. "Kantian Constructivism in Moral Theory." *The Journal of Philosophy* 77 (9): 515-572.
- Rawls, John. 1993. *Political Liberalism*. New York, NY: Columbia University Press.
- Rawls, John. 2001. *Justice as Fairness: A Restatement*. 2nd edition. Edited by Erin Kelly. Cambridge, MA: Harvard University Press.
- Robbins, Lionel. 1938. "Interpersonal Comparisons of Utility: A Comment." *The Economic Journal* 48 (192): 635-641.
- Sandel, Michael. 2010. *Liberalism and the Limits of Justice*. 2nd edition. New York, NY: Cambridge University Press.
- Sen, Amartya. 1970. *Collective Choice and Social Welfare*. San Francisco, CA: Holden-Day.
- Sen, Amartya. 1973. "Behaviour and the Concept of Preference." *Economica* 40 (159): 241-259.

- Sen, Amartya. 1979. "Utilitarianism and Welfarism." *The Journal of Philosophy* 76 (9): 463–489.
- Sen, Amartya. 2017. *Collective Choice and Social Welfare: Expanded Edition*. London: Penguin Books.
- Suppes, Patrick. 1966. "Some Formal Models of Grading Principles." *Synthese* 16 (3/4): 284–306.
- Weymark, John. 2016. "Social Welfare Functions." In *The Oxford Handbook of Well-Being and Public Policy*, edited by Matthew D. Adler and Marc Fleurbaey, 126–158. New York, NY: Oxford University Press.

Cyril Hédoin is Professor of Economics at the University of Reims Champagne-Ardenne (France). He works at the intersection of economics and philosophy, and more particularly on issues related to rationality, rules, and institutions. He has more recently written on topics articulating normative economics with political and moral philosophy. His work has been published in many philosophy and economics journals, such as *Economics and Philosophy*, *Erkenntnis*, and *Social Epistemology*.
Contact e-mail: <cyril.hedoin@univ-reims.fr>

Integrated Moral Agency and the Practical Phenomenon of Moral Diversity

MICHAEL MOEHLER

Virginia Tech

Abstract: The practical phenomenon of moral diversity is a central feature of many contemporary societies and poses a distinct problem to moral theory building. Because of its goal to settle the moral question fully and exclusively and/or to provide better understanding of moral disagreement, traditional first-order moral theory often does not provide sufficient guidance to address this phenomenon and moral agency in deeply morally diverse societies. In this article, I move beyond traditional first-order moral theorizing and, based on multilevel social contract theory (Moehler 2018, 2020a), develop a practically sound notion of moral agency for morally diverse societies. The interrelational and dynamic notion of *integrated moral agency* developed in this article demands that agents actively exercise their rational and affective capacities, are receptive to the capacities of others, and are aware of the type of moral interaction in which they engage with others. The notion of integrated moral agency helps agents to reconcile conflicting first-order moral directives and to maximally protect agents' autonomy in morally diverse societies.

Keywords: traditional first-order moral theory, multilevel social contract theory, autonomy, integrity, respect

JEL Classification: B15, B25, B31, B41

I. INTRODUCTION

Moral diversity is a central feature of many contemporary societies. In such societies, even after careful consideration of their well-considered moral views, agents often hold divergent moral ideals that cause moral disagreement. If such disagreement is stark and this *practical phenomenon of moral diversity* (as I call it) endures, then moral diversity does not always facilitate progress, but may lead instead to severe conflict and de-

structive action. In morally diverse societies, and especially in *deeply morally diverse* societies, the ideal of a fully just society, as judged from the perspectives of all members of society, is often unattainable.

According to Rawls, one central *practical* task of moral and political philosophy is to determine reasoned ways to resolve or, if resolution is not possible, reduce moral and political conflict to ensure mutually respectful, peaceful cooperation:

One task of political philosophy—its practical role, let’s say—is to focus on deeply disputed questions and to see whether, despite appearances, some underlying basis of philosophical and moral agreement can be uncovered. Or if such a basis of agreement cannot be found, perhaps the divergence of philosophical and moral opinion at the root of divisive political differences can at least be narrowed so that social cooperation on a footing of mutual respect among citizens can still be maintained. (Rawls 2001, 2)

According to Rawls, moral diversity matters not only practically, but is also central to moral and political philosophy.¹ One, albeit not very meaningful, way to address the practical phenomenon of moral diversity for moral theory building is to discount its legitimacy. For example, some (although surely not all) moral realists who believe in moral truth and insist that they know it may discount the moral views of others if these views conflict with theirs. Moreover, some moral skeptics who do not believe in morality as traditionally conceived may discount all moral views, including the realists’ truth, because moral skeptics do not believe in any legitimate morality. From these viewpoints, the practical phenomenon of moral diversity is either spurious or does not merit further investigation.

At the level of moral theory building, different first-order moral theories often justify conflicting moral conclusions and, in this sense, compete with each other. As Gaus notes:

We often understand our ‘theories of morality’ as competing theories describing and explaining the same phenomenon. Indeed, moral philosophers often identify themselves in terms of the adherence to one or the other theory explaining what morality is all about. (Gaus 2011, 551)

First-order moral theory often assumes moral monism, although, as Gaus clarifies, “there is precious little defense of it” (2011, 554). Moreover, if

¹ See Gaus (2011, 2016), Muldoon (2016), and Müller (2019).

first-order moral theories do allow for plurality of moral conclusions (such as relativistic moral theories), then such theories are typically inadequate for principled resolution of deep moral conflict, precisely because they are too tolerant of diversity at the level of moral theory building.

As a consequence, traditional first-order moral theory typically does not offer sufficient guidance to address the practical phenomenon of moral diversity, although it may help to provide better understanding of the reasons for moral disagreement. Traditional first-order moral theory, in its quest to settle the moral question fully and exclusively, either does not allow sufficient diversity to capture the well-considered moral views of agents, or, if different first-order moral theories together do capture the well-considered moral views of agents, then such theories typically do not offer principled guidance to resolve disagreement among conflicting first-order moral theories. Either way, traditional first-order moral theory typically fails to address the practical phenomenon of moral diversity adequately.

In this article, I do not aim to criticize or discredit the role of traditional first-order moral theorizing, which Scanlon describes as follows:

All that I have said may seem simply to confirm that, as MacIntyre has written, 'Modern academic philosophy turns out by and large to provide means for more accurate and informed definition of disagreement rather than for progress toward its resolution.' If the 'resolution' in question is a matter of finding arguments that could be deployed to compel agreement between the contending parties, then I agree that philosophy has not been able to provide it and is not very likely to do so. On the other hand, 'more accurate and informed' understanding of disagreement, and of agreement where it exists, seems to me to be an important form of progress—a form that moral theory can reasonably aim at. (Scanlon 1995, 356)

In this article, I move beyond traditional first-order moral theorizing. For a moment, I put on hold—or, depending on one's perspective, build upon the lessons learned from—the more than 2,500-year-old debate in moral philosophy that aims to determine the ultimately correct view of morality and/or to gain better understanding of moral disagreement. Whatever the conclusion of this debate, if it has one, it will not be able to address the practical phenomenon of moral diversity if agents' well-considered moral views in morally diverse societies, especially in deeply morally diverse societies, are not fully captured by one particular first-order moral theory (which the very existence of different first-order moral theories suggests),

and if the precise relationship and jurisdiction of different first-order moral theories are unclear.

I consider the practical phenomenon of moral diversity to be a legitimate concern for moral philosophy. Following Rawls, the core objective of such practical moral philosophy is to determine a moral framework that, despite the ongoing and often severe moral disagreement that is reflected by the conflicting moral conclusions of different first-order moral theories, specifies the moral demands that allow agents to live peacefully with one another on the basis of mutual respect. In the recent literature, several practical (or functionalist) moral and political theories have been developed, in particular by D'Agostino (2003), Gaus (2011, 2016), Muldoon (2016), Müller (2019), Van Schoelandt (2020), and Caton (2020); although not all of these theories explicitly address the practical phenomenon of (deep) moral diversity and develop a notion of moral agency.

Without disregarding other approaches, I build upon multilevel social contract theory (Moehler 2018, 2020a), which offers one possible framework to address the practical phenomenon of deep moral diversity and moral agency. From the perspective of moral theory building, multilevel social contract theory integrates different first-order moral theories. Specifically, in its simplified version, the theory integrates Hobbesian contractarianism, Humean conventionalism,² and Kantian contractualism into one systematic moral theory.³ According to multilevel social contract theory, morality does not consist of one single system of moral rules. Instead, it consists of a multitude of such systems that are valid simultaneously and ordered hierarchically to define the demands of a complex moral world.

In this article, I contextualize and systematically develop the *basic features* of the notion of moral agency that underlies multilevel social contract theory.⁴ I argue that multilevel social contract theory offers a sound notion of moral agency for morally diverse societies where, especially under the assumption of moral under-determination, moral agency demands the active exercise of agents' rational and affective capacities. It demands that agents are receptive to others and the specific type of moral interaction in which the agents engage with others in order to elicit the most substantial common moral ground. This novel interrelational and dynamic notion of *integrated moral agency* helps to reconcile conflicting

² For discussion of Humean moral conventionalism, see Moehler (forthcoming).

³ For the difference between 'contractarianism' and 'contractualism', see Darwall (2003).

⁴ Related but conceptually different notions of moral agency have been developed by defenders of Kantian constructivist ethics, in particular Korsgaard (2009).

first-order moral directives and to maximally protect the autonomy of agents in morally diverse societies. The new notion of integrated moral agency allows agents to make sense of their complex realities in morally diverse societies and offers a principled way to resolve or, if resolution is not possible, mediate moral conflict while treating others with maximal respect.

II. THE STRUCTURE OF MULTILEVEL MORALITY

In its simplified form, multilevel social contract theory (Moehler 2018, 2020a) integrates within one systematic moral theory three different first-order moral theories, namely, ‘moral contractarianism’ that originates with Hobbes’ ([1651] 1996) moral theory, ‘moral conventionalism’ that originates with Hume’s ([1739/1740] 2000) moral theory, and ‘moral contractualism’ that originates with Kant’s ([1785] 1998) moral theory and that has been developed further by Scanlon (1998) and Southwood (2010).⁵

Moral conventionalism and moral contractualism are ‘traditional moral theories’ (as I employ the term). Traditional moral theories assume, as a basis for the justification of moral rules, that agents value moral ideals (shared or not) at least partially for intrinsic reasons or embrace such ideals for other traditional moral reasons, such as altruistic reasons or similarly motivated other-regarding reasons. As a result, traditional moral theories cannot fully capture the practical phenomenon of deep moral diversity that includes agents who do not embrace morality on traditional, noninstrumental grounds. Moral contractarianism, by contrast—as a result of its purely instrumental approach—can accommodate deep moral diversity, and thus can complement moral conventionalism and moral contractualism from the perspective of moral theory building so long as agents share the overarching goal of ensuring peaceful cooperation.

Multilevel social contract theory represents a distinct position in moral theory that, even in its simplified version, differs from Parfit’s (2011) triple theory. Parfit’s theory holds that rule consequentialism, Kantian contractualism, and Scanlonian contractualism lead to similar moral conclusions and thus represent different ways to ‘climb the same mountain’. Multilevel social contract theory, by contrast, combines three different contractarian moral theories within one systematic moral theory that entails Humean, Hobbesian, and Kantian moral features. Multilevel

⁵ See Gauthier (1997, 134–135), Watson (1998, 173–174), Darwall (2003, 1–8), D’Agostino, Gaus, and Thrasher (2017), and Moehler (2018, 11–12; 2020a, 3–7).

social contract theory does not claim that different contractarian moral theories converge to reach similar moral conclusions, although it suggests so in a weak sense with respect to Hobbes' and Kant's moral theories (as I clarify in section III). Instead, multilevel social contract theory considers different contractarian moral theories to be valid for different domains of morality.

Multilevel social contract theory purports that, in order to climb a mountain successfully, different types of moral theory may apply the higher one climbs. If, with increasing height, the air becomes thinner and the moral terrain more demanding, then different moral theories may be needed that are valid for these different circumstances and may prescribe different behaviors (what one ought to do). The higher one climbs and the thinner the air becomes, then, morally speaking, the less specific the demands of morality may be, although the demands need not necessarily be less significant. If one comes close to the top of the mountain where, after reaching the point of no return, the air is so thin that, morally speaking, no moral basis as traditionally conceived exists any longer among agents apart from the agents' goal to reach the top of the mountain, then pure instrumental morality is the only guide left to ensure one's survival.

Multilevel social contract theory also differs from Southwood's (2010) moral theory, which holds that existing contractarian moral theories, especially Hobbesian contractarianism and Kantian contractualism, do not offer plausible moral foundations. According to Southwood, Hobbesian contractarianism:

At best appears to get morality wrong in the right way [...] on account of its reliance on an implausibly personal and partial characterization of the moral point of view. Kantian contractualism at best appears to get morality right in the wrong way [...] on account of its reliance on a substantive conception of practical reason. (Southwood 2010, 190)

Southwood's (2010, 88-96, 124-128) 'deliberative contractualism' represents an intermediate position that, according to this view, is superior to Hobbesian contractarianism and Kantian contractualism, and demands that agents actively engage in deliberation with others to reach consensus on a common code by which to live.

Multilevel social contract theory, by contrast, keeps the core features of Hobbesian moral contractarianism, Humean moral conventionalism, and Kantian moral contractualism intact, and, in doing so, maintains the

strengths of each of these theories. In order to capture the practical phenomenon of moral diversity and offer a principled way to resolve conflict in deeply morally diverse societies, multilevel social contract theory considers the moral domain to be heterogeneous. It assumes that agents may agree on different social contracts that are valid simultaneously for different types of moral interaction and that such contracts may be restricted in their universality and generality. Nevertheless, it requires that agents are fit for social cooperation and morality, and will ultimately agree on regulations for all relevant types of moral interaction to ensure mutually respectful, peaceful cooperation.

Moreover, multilevel social contract theory assumes that moral rules are ordered hierarchically, with lower-level moral rules taking priority over higher-level moral rules in the regulation of moral interactions. The theory assumes that agents will always justify moral rules to themselves and others based upon the most substantial common moral ground with others, both in the traditional and purely instrumental understandings of morality, for the most local domain. The theory assumes that, based on some common denominator as a starting point or an end point for the justification of moral rules that may vary for different types of moral interaction, agents will always agree on the least invasive system of moral rules to regulate their moral interactions. In the following, I clarify the specific nature of multilevel morality that determines the basic features of its underlying notion of moral agency.

III. THE NATURE OF MULTILEVEL MORALITY

In keeping with the plurality thesis, the practical view of morality defended by multilevel social contract theory does not aim to capture all aspects of morality, but focuses on *social* morality as opposed to *personal* morality. Specifically, the goal of multilevel morality is to harmonize moral interactions among agents while ensuring peaceful cooperation independent of the precise origin of morality, such as evolution from actual social practices or justification by rational procedures or both.⁶ Nevertheless, one core challenge for multilevel social contract theory (to the extent that the practical view of morality underlies this challenge) is to show that

⁶ Despite this feature, genuine social moral rules, as understood here, are considered to be distinct from merely 'socially constructed norms' that may or may not have moral significance. For discussion of socially constructed norms and their relationship to moral theory, see Valentini (2021). For discussion of potential limitations of this view of morality, see Morris (2020) and Moehler (2020b, 97–98).

multilevel morality, especially for the domain of pure instrumental morality, defends a genuine form of morality because, under certain conditions, the demands of pure instrumental morality are assumed to override the demands of traditional morality.⁷

The first aspect of this challenge concerns moral motivation and is expressed by Prichard's dilemma.⁸ Prichard argues that there is no good reason to act morally. If one refers to moral reasons, such as altruistic reasons, then one presupposes the persuasive force of morality, which is circular reasoning. If one refers to nonmoral reasons, such as self-interest, then one provides the wrong kind of reasons to act morally. In contractarian moral theory, Scanlon's contractualism provides a direct response to Prichard's dilemma. Scanlon (1998, 147–158) argues that there is a third type of reason to act morally that is neither a moral reason nor a nonmoral reason as traditionally conceived, but that stems from the consideration of an action's wrongness and its justifiability towards others. According to Scanlon, this account of moral motivation:

Is closely enough connected to our ideas of right and wrong to be clearly an account of *moral* reasons, but it is not so closely identified with these ideas as to amount to the trivial claim that the reason we have to avoid certain actions is just that they are wrong. (Scanlon 1998, 187)

Even if correct, Scanlon's contractualist moral theory is valid only for agents who *de facto* are "moved by the aim of finding principles that others, similarly motivated, could not reasonably reject" and who are not moved merely by "seeking some kind of advantage" (1998, 5). Although Scanlon (1998, 158–160) highlights the importance of agents caring about such justification, Scanlon's theory, as a result of its contractualist features, cannot sufficiently address the practical phenomenon of moral diversity that includes agents who, at least in certain types of moral interaction, may exclusively seek their own advantage despite sharing a common goal with others. Such agents fall outside the scope of Scanlon's moral theory. This finding does not constitute a criticism of Scanlon's theory which, as a result of its traditional moral nature, pursues a different task.

Multilevel social contract theory, within the domain of pure instrumental morality, includes such agents who may exclusively seek their own

⁷ For related discussion, see Moehler (2020a, 15–16).

⁸ See Prichard (1912).

advantage. Yet, the theory does not fall prey to what Southwood (2019) calls the ‘concessional fallacy’, namely that a moral theory may not be sufficiently demanding because it overly caters to agents’ self-interests. Multilevel social contract theory relies on the ‘principle of subsidiarity’, and thus justifies moral rules always on the most substantial traditional moral ground (shared or not). The theory defers to pure instrumental reasoning only in cases where traditional morality fails. In Hobbesian terms, multilevel social contract theory takes agents ‘as they are’ as a basis for morality, although the thinner the traditional moral ground among agents, the more carefully agents must reflect on their traditional moral views and their importance compared to the agents’ aim of reaching their overarching goal.

Because multilevel social contract theory, for the domain of pure instrumental morality, does not require that agents follow moral rules on the basis of what are traditionally conceived to be moral reasons (although it does not rule out such reasons) but allows agents to be motivated entirely by self-interest, the theory faces the ‘wrong kind of reasons’ objection. More fundamentally, the problem concerns the priority of prudential moral reasoning over traditional moral reasoning for the domain of pure instrumental morality.⁹ To be clear, within the domain of pure instrumental morality, multilevel social contract theory allows agents to value their traditional moral views highly. It generally only rules out that agents value their traditional moral views infinitely over other agents’ traditional moral views, their lives, and the general goal of ensuring peaceful cooperation.

Such agents, whom I call *homo categorical*,¹⁰ are not willing to make concessions in conflict situations, which puts them deeply at odds with others and renders mutually respectful peaceful cooperation impossible in societies where the practical phenomenon of moral diversity endures. Such agents, often as a result of ideological considerations, are dogmatic and fall outside the scope of pure instrumental morality. Also, if agents, such as extreme glory seekers or suicide bombers, do not embrace the overarching goal to ensure peaceful cooperation, including an interest in preserving their lives and the means necessary to do so, then they fall outside the scope of the theory.¹¹ Even if pure instrumental morality imposes only weak normative constraints on the reasoning of agents, the

⁹ See Gaus (2019, 110–111).

¹⁰ See Moehler (2018, 103).

¹¹ See Moehler (2020b).

theory has normative force and thus considers some agents to be potentially irrational.

For cases of conflict in which moral reasoning is reduced to instrumental reasoning and peace is at risk, multilevel social contract theory assumes that the demands of pure instrumental morality override the demands of traditional morality. In societies in which the practical phenomenon of moral diversity endures, by definition agents disagree on what is morally right as traditionally conceived and what qualifies as genuine moral reasons. Under the assumption of deep moral diversity, traditional moral reasons are just one type of reason that may or may not help to harmonize the behavior of agents. Under such conditions, traditional moral reasons are not privileged reasons and may often be the main cause for conflict.

In cases where pure instrumental morality applies, multilevel social contract theory assumes that life, and the human cooperation that is necessary to sustain it, are more important than traditional morality. Or, traditional morality is misconceived if it pits agents against each other and endangers their existence.¹² In such cases, agents must distance themselves from their traditional moral views to evaluate their overall interests and mediate conflicts according to the demands of pure instrumental rationality to ensure that, in Rawls' terms, "social cooperation on a footing of mutual respect among citizens can still be maintained" (2001, 2). According to multilevel morality, agents who are not able to do so are not fit for cooperation and social morality. Multilevel social contract theory does not claim that there is no space outside of morality. Instead, it merely extends the limits of morality as far as methodologically possible within the bounds of social morality and moral agency.

The second aspect of the challenge to show that multilevel morality represents a genuine form of morality concerns the role and nature of moral emotions. Arguably, pure instrumental morality does not sufficiently consider the importance of agents' affective capacities, such as anger, indignation, guilt, blame, and resentment, and more generally the moral psychology of agents. Moreover, as Southwood (2008, 185–186; 2010, 34–42) stresses, even if pure instrumental morality were to consider such capacities, it may misconstrue their nature as primarily self-directed. To be clear, the fact that multilevel social contract theory prioritizes reason over emotion within the domain of pure instrumental morality does not imply that the theory neglects the importance of such traditional

¹² See Moehler (2019, 145).

moral concepts. Instead, it merely considers them to be typically part of traditional morality. According to multilevel social contract theory, both reason and emotion have their roles in moral agency. Nevertheless, in situations where traditional morality turns agents against each other and threatens their existence, traditional moral concepts cannot serve as the ultimate moral compass because they are often the very cause for conflict.

The third aspect of the challenge concerns the content of morality. In order to be considered legitimate, the moral rules that are justified by pure instrumental morality must, in the relevant sense, resemble traditional moral rules. Although this consideration is not essential for my argument here, the moral rule that I defend for the domain of pure instrumental morality under the specific circumstances described (under different circumstances, different rules may be justifiable) fulfills this condition.¹³ I call this rule the *weak principle of universalization* and argue that it can be considered to be a weak version of Kant's categorical imperative, because it weakly expresses the moral ideals that underlie Kant's moral law. Nevertheless, the rule is not as general and universal as Kant's categorical imperative because it applies only to specific types of conflict and only to agents in this empirical world. The advantage of the weak principle of universalization is that agents do not need to embrace these moral ideals on traditional moral grounds, because these ideals can be justified instrumentally. Ultimately, multilevel social contract theory stresses the respect that agents owe each other if they want to coexist peacefully despite the endurance of the practical phenomenon of moral diversity. Based on its general assumptions, multilevel social contract theory defines the bounds of social morality and thus also the bounds of meaningful moral agency.

IV. AUTONOMY, INTEGRITY, AND AGENCY

Multilevel social contract theory, within its defined bounds, aims to provide agents with maximal autonomy in morally diverse societies. To this end, the theory allows several levels of agreement among agents on moral rules that may be restricted with respect to their universality and generality—within both the domain of traditional morality and that of pure instrumental morality, as well as within the domain of what Van Schoelandt aptly calls “intermediate moralities” that combine features of both types

¹³ See Moehler (2018, 133–139; 2020a, 48–52).

of morality (2019, 133).¹⁴ Although the discussion in this article focuses primarily on triple theory that combines Hobbesian contractarianism, Humean conventionalism, and Kantian contractualism, multilevel social contract theory allows for further levels of moral rules that define a more fine-grained *n*-level social contract theory and that determine a complex web of moral relationships in diverse societies.

In order to allow for moral diversity, multilevel social contract theory does not require that potential demands stemming from nonmoral interpersonal relationships must be consistent with the demands of morality for the entire moral domain, as, for example, Scanlon's traditional moral theory requires. According to Scanlon's theory, morality defines the most fundamental demands on human agency, and any additional nonmoral social obligations that may arise through love, friendship, and other social relationships, if adequately conceived, must respect the priority of traditional morality. Such demands must be sensitive to the demands of traditional morality. As Scanlon puts it:

Interacting with others *qua* chess players, *qua* lovers, or poets, is a special case of interacting with them *qua* rational creatures. If we didn't think of them as having the status of rational creatures, we wouldn't be able to relate to them in the way that we do. Therefore, I would say that a relationship to others that brings the moral requirements of 'what we owe to each other' in train is presupposed by the more specific forms of relationships. That is one of the reasons why moral requirements take precedence over other relationships. (Scanlon and Voorhoeve 2001, 30)

According to Scanlon's theory, a person who murders another person for a friend is not a real friend. Scanlon considers the priority of traditional morality and its presupposition by other interpersonal relationships to be a necessary condition for the creation of a kingdom of ends in the Kantian sense.

Also, in the context of his discussion of moral relativism, Scanlon stresses that, although his theory allows for variable moral practices (whereby an action that is wrong in one context may be morally unobjectionable in another context), ultimately the moral force of such practices "is explained by appeal to a single substantive moral principle", a position that Scanlon calls "parametric universalism" (1998, 340). Following Kant,

¹⁴ For further discussion of the features of traditional morality related to compliance issues, see Van Schoelandt (2019, 132).

Scanlon's moral theory assumes that the categorical imperative is valid for the whole moral domain and thus any lower-level moral rules that are expressed in the form of hypothetical imperatives, which do not qualify as genuine moral rules in Kant's moral theory, cannot override or run counter to the demands of the categorical imperative from the perspective of moral agency.¹⁵

Multilevel social contract theory, by contrast, does not require that the demands of different levels of morality are consistent with each other, because conflicts within the domain of lower-level morality may often be the very reason for appeal to higher-level morality.¹⁶ For justificatory purposes, according to multilevel social contract theory, the normative content of higher-level morality is independent of the normative content of lower-level morality. According to multilevel social contract theory, higher-level morality has moral authority precisely in situations where lower-level morality has failed to resolve conflict. In agreement with Scanlon's theory, multilevel social contract theory considers morality to be foundational. However, in contrast to Scanlon's theory, multilevel social contract theory considers morality also to be layered and hierarchical.

As such, within the domain of pure instrumental morality, a plurality of reasons—including reasons that stem from personal commitments, such as friendship, love, and other social relationships—may be part of the justificatory basis for moral rules. Furthermore, the demands of such personal reasons need not be consistent with the demands of traditional morality because, within the domain of pure instrumental morality, traditional moral reasons themselves may be controversial and the primary cause for conflict. Within the domain of pure instrumental morality, traditional moral reasons are not privileged reasons. Instead, agents' non-moral reasons as traditionally conceived may be just as important as their traditional moral reasons for ensuring mutually respectful, peaceful co-operation.

The diversity of reasons that multilevel social contract theory allows within the domain of pure instrumental morality helps to protect the integrity of agents, which, according to Bernard Williams, is an important feature of a sound moral theory (to the extent that Williams can be considered to be interested in moral theory). In his critique of utilitarianism, Williams (1973, 100–118) argues that agents' lives are structured around

¹⁵ See Kohl (2018).

¹⁶ See Moehler (2018, 21).

people and projects they care about, including their own wellbeing, families, friends, and the fulfillment of basic necessities of life, as well as cultural and aesthetic interests. Such relationships and projects, and the commitments that follow from them, give meaning to agents' lives and constitute their identity as persons. Integrity demands that agents' identities are protected. Williams argues that utilitarianism, because of its focus on the group level, does not sufficiently protect the integrity of agents. Utilitarianism demands strict impartiality and impersonality of agents with regard to their own projects and the projects of others, which may alienate agents from their own lives.¹⁷ A sound moral theory must provide room for personal relationships and the pursuit of agents' important life projects.

Scanlon's moral contractualism responds to this criticism. According to Scanlon (1998, 160), the morality of 'what we owe to each other' provides agents with sufficient room to pursue personal relationships and life projects, because moral rules that do not allow such personal space could be reasonably rejected by agents. That is, although Scanlon's moral theory assigns priority to traditional moral reasons as the foundation for social relationships, his theory leaves sufficient room for personal considerations. However, this provision holds only if all members of society are reasonable in the specific way presupposed by Scanlon's theory, because the demands of the concept of reasonableness determine the claims to which agents can object.¹⁸ In morally diverse societies, the demands of the concept of reasonableness are likely to be controversial and, in deeply morally diverse societies, not all agents can be assumed to be reasonable. As a result, Scanlon's moral theory, which is limited to the bounds of traditional moral reasoning, may compromise the integrity of agents who disagree with the prevalent traditional morality. The interests of such agents are excluded from Scanlon's moral theory.

Multilevel social contract theory, by contrast, includes the interests of such agents, in cases where traditional morality fails to resolve moral conflict. Within the domain of pure instrumental morality, not only traditional moral reasons can serve as reasons for the justification of moral rules, but also other reasons that may stem from important personal relationships and life projects. Moral theory, if it aims to address the practical phenomenon of moral diversity, must consider such reasons to

¹⁷ See also Moehler (2013).

¹⁸ For further discussion of this consideration, see Suikkanen (2019).

which agents may be strongly committed. Even if multilevel social contract theory limits the bounds of social morality and moral agency (as discussed in section III), the theory helps to maximally protect the integrity of agents in deeply morally diverse societies where traditional moral reasons are controversial.

Moreover, multilevel social contract theory suggests a complex notion of moral agency that entails that, in deeply morally diverse societies, agents may need to follow different moral rules for different types of moral interaction. To illustrate this idea in a nonmoral context, Nguyen (2019, 2020) suggests in his analysis of ‘games’ (in their broadest sense) and the ‘rules’ that govern them that games do not merely create structures for social interaction, but that they also create ‘temporary selves’. Games can be seen as describing (sometimes arbitrary) rules for different types of social interaction that shape agents’ experiences. Nguyen’s analysis of social interaction through the lens of games shows that agency is typically not only socially embedded, modular, and fluid, but also often requires that agents adopt, at least temporarily, certain ends for successful social interaction. Moreover, playing different games with different rules may allow agents to understand more fully their own selves and the importance of constraints for personal and social development.

The notion of moral agency that is implicit in multilevel social contract theory does not hinge on Nguyen’s analysis. Nevertheless, for some readers, Nguyen’s analysis may be helpful for better understanding the complex notion of moral agency that is suggested by multilevel social contract theory. Multilevel social contract theory—with its different levels of morality that define a complex web of moral rules that are valid simultaneously for different types of moral interaction—suggests a socially embedded, differentiated, and fluid notion of agency that is similar to the one found in Nguyen’s analysis. One central difference to Nguyen’s analysis is that multilevel social contract theory focuses exclusively on the domain of morality and thus underlies the (non-arbitrary) constraints of moral theory. In the following, I lay out the basic features of the notion of integrated moral agency that underlies multilevel social contract theory.

V. INTEGRATED MORAL AGENCY

As the previous discussion of the structure and nature of multilevel social contract theory implies, the notion of moral agency that is defended by multilevel social contract theory differs from other notions of moral

agency by *integrating* different contractarian moral theories within one systematic moral theory.¹⁹ Multilevel social contract theory integrates traditional moral reasoning, as captured by Humean moral conventionalism and Kantian moral contractualism, with pure instrumental or prudential moral reasoning, as captured by Hobbesian moral contractarianism.²⁰ Because most readers will be familiar with traditional moral reasoning, I will focus primarily on the aspects of moral agency that pertain to the domain of pure instrumental morality and their integration with traditional morality.

For the domain of pure instrumental morality, multilevel social contract theory relies on a particular Hobbesian model of moral agency that I call the '*homo prudens* model'.²¹ The *homo prudens* model assumes that, in moral interactions in which traditional moral reasoning fails and pure instrumental rationality is the only means to ensure peace, agents, in addition to reflecting upon their goals and empirical conditions, are forward-looking and value their lives and the expected gains from peaceful cooperation more than they value noncooperation *per se*.

The *homo prudens* model is a close cousin of the *homo economicus* model (the predominant model of human agency in economic theory) that is often thought to represent a particularly liberal notion of human agency. According to Gauthier (1986, 345), instrumental rationality, especially in the way it is employed by Gauthier in the context of his theory of 'morals by agreement', captures the reasoning and affections of economic agents and, more specifically, the affections of the 'liberal individual'. Gauthier (1986, 353–354) argues that the liberal individual is only a recent invention and that it is unclear whether or not her 'ecological niche' can be realized and sustained. As such, Gauthier's moral theory may be bound by time.

This characterization of the *homo economicus* model does not apply to the *homo prudens* model employed by multilevel social contract theory. Multilevel social contract theory assumes that instrumental rationality is an essential and permanent part of human agency because the capacity for instrumental reasoning is generally necessary for agents to survive in this empirical world. In this empirical world, instrumental reasoning can be viewed as a timeless feature of human agency. Moreover, according to multilevel social contract theory, the *homo prudens* model represents a

¹⁹ For discussion of other notions of moral agency, especially notions of moral agency developed by supporters of Kantian constructivist ethics, see Korsgaard (2009).

²⁰ For Hobbes' prudentialism, see Abizadeh (2018, 110).

²¹ See Moehler (2018, 95; 2020a, 48).

partial model of human agency that is assumed to be valid only for moral interactions in which moral reasoning is reduced to instrumental reasoning and instrumental morality is the only means to ensure peace.

In other words, agents' behavior within the domain of traditional morality is not assumed to be guided by the demands of the *homo prudens* model and, more specifically, the ideal of individual utility-maximizing behavior. Instead, as discussed in section III, such behavior is assumed to be guided by traditional morality and its core notions, including moral emotions and agents' affective capacities that, in the specific case of moral contractualism, may have a certain liberal content. The *homo prudens* model does not aim to capture the whole range of moral behavior, but only the behavior relevant to certain types of moral interaction. As such, depending on the specific type of moral interaction, one and the same agent may be guided sometimes by traditional morality and sometimes by the *homo prudens* model. According to multilevel social contract theory, different types of moral interaction may require different constructions of morality, depending on the common moral ground among agents, both in the traditional and purely instrumental understandings of morality.

This feature of multilevel social contract theory does not imply that agents are assumed to have a split personality. Instead, multilevel social contract theory merely assumes that, under the endurance of the practical phenomenon of moral diversity, agents consider all relevant reasons for the justification of moral rules in the light of their overarching goal of ensuring peaceful cooperation. Because peaceful cooperation is an interdependent variable and because, in the domain of pure instrumental morality, agents are a potential threat to each other, agents may need to abstract from certain aspects of traditional morality—at least in certain types of moral interaction—in order to reach their overarching goal of ensuring peaceful cooperation (as conceptually expressed by the restriction of the generality of moral rules). For the justification of moral rules, agents must consider their goals and moral views as well as those of other agents, even if agents are not intrinsically motivated to justify their actions towards others, as assumed by traditional moral theories.

Further, according to multilevel social contract theory, moral agency is *interrelational*. According to multilevel social contract theory, moral demands depend not only on the specific type of moral interaction, but also on the agents' cooperative partners (as conceptually expressed by the restriction of the universality of moral rules). Because multilevel social

contract theory demands that agents always determine the most substantial common moral ground with their cooperative partners, in terms of both the traditional and purely instrumental understandings of morality, sometimes agents may appeal to their full-fledged traditional morality, or certain parts of it, and sometimes they may need to rely on pure instrumental morality.

Stated differently, agents may be bound by their traditional morality with regard to some agents and/or some types of moral interaction and by pure instrumental morality with regard to other agents and/or other types of moral interaction. Agents may be under the authority of different moralities (or intermediate moralities) at the same time, depending on their cooperative partners and the precise form(s) of moral interaction. Agents may be in multiple moral relationships with each other that require different forms of moral reasoning. In this sense, agents may be considered to engage in moral role-playing, which demands that agents take on different moral perspectives and reconcile conflicting first-order moral directives by accepting different moral constraints (rules) for different agents and/or different types of moral interaction. Agents may be considered to play different 'moral games' that are governed by different moral rules.

Phenomenologically, engaging in such moral role-playing may provide agents with a better understanding of their own selves and other agents, which may lead agents to a less rigid and more constructivist understanding of morality. In addition, perceiving moral agency in such an integrated and interrelational manner allows for differentiation of agents' moral views, which facilitates the reconciliation of conflicting moral views both within and among agents, especially in morally fragmented societies that are prone to conflict.²² Moral differentiation, which is inherent in the structure of multilevel social contract theory, requires that agents consider the universality and generality of their own moral views and the moral views of others, as well as reflect upon the origin and legitimacy of such views and their underlying values.

Moral differentiation may render it more acceptable for agents to make concessions with regard to their moral views as temporary means for social interaction, because at no point is an agent's entire worldview at stake. Moral differentiation may also help to preserve agents' integrity

²² For discussion of this point, see Ostrom and Ostrom (2002, 96). For discussion of the notion of fragmentation in the context of collective moral responsibility, see Braham and van Hees (2018).

in morally fragmented societies, and thus to address Williams' (1973) integrity concern discussed in the previous section, even if, as Williams (1985) concedes, not all human projects may ultimately be fully reconcilable.²³ The developed notion of moral agency allows agents to connect their moral views systematically to form a unified, integrated moral self, even if the notion of moral agency demands that, due to the complexities of deeply morally diverse societies, agents carefully differentiate their moral views.

Moreover, the integrated and interrelational notion of moral agency that is defended by multilevel social contract theory encourages agents to perceive disagreements as opportunities for discovery, learning, and change. It invites agents to probe their views both conceptually and historically and, if they consider the moral views of others to be legitimate, to revise their views and construct the best moral world together based on the agents' combined perspectives. The notion of moral agency asks agents to exercise their moral capacities as an integral part of the determination of the constraints of morality that agents impose on themselves and others. Such active engagement with others may allow agents to become better moral agents over time because it requires that the agents carefully reflect on their views and the moral views of others, even if disagreements remain and consensus is a false ideal in morally diverse societies.

Returning to the distinction between traditional and purely instrumental morality, in practice, there will probably be a continuum with regard to agents' moral reasons and motivations that involves, at the most foundational level, traditional moral concepts and that ascends to pure instrumental morality. As a result, in practice, agents' moral reasons and motivations may not be captured fully by moral conventionalism, moral contractualism, and moral contractarianism. As indicated in section II, the division of contractarian moral theory into three types of moral theory is intended merely to clarify the core differences among different contractarian moral theories. In practice, under the phenomenon of moral diversity, moral agency is typically more complex. Multilevel social contract theory can account for such complexity that determines a complicated web of moral relationships among agents within the domains of both traditional and pure instrumental morality.

²³ For critical discussion of Williams' view in the context of virtue ethics, see Cottingham (2010).

A final feature of multilevel social contract theory stems from the fact that it regards traditional morality to be culturally-dependent and path-dependent, because the theory regards traditional morality to be grounded at least partially in evolving social moral practices, whereas the demands of pure instrumental morality are assumed to be based upon general and fairly stable assumptions about human nature, empirical facts, and conditions of rationality. Nevertheless, in the form discussed, the demands of multilevel morality are valid only for moral interactions among agents who live in this empirical world and who reason in the mode of *homo prudens*. This feature of pure instrumental rationality stands in contrast to Kant's moral theory, which assumes that moral rules carry with them absolute necessity and thus are valid for all rational beings in the universe.²⁴ As a result of this feature, the precise moral demands of multilevel social contract theory cannot be determined *a priori*, but are constructed by the exercise of practical reasoning in moral interactions.

In this sense, multilevel social contract theory suggests a *dynamic* notion of moral agency. As a result of combining evolutionary aspects of morality in the domain of traditional morality with static (although not necessarily eternally fixed) rationalistic aspects in the domain of pure instrumental morality, the notion of moral agency defended by multilevel social contract theory requires that agents (re)construct the demands of morality under evolving social conditions by exercising their rational and affective capacities. Multilevel social contract theory takes seriously the primacy of evolved morality and its relationship to reason and thus the lessons of Gaus' (2018a, 2018b, 2018c, and forthcoming) work on social morality. Nevertheless, one essential difference between multilevel social contract theory and Gaus' moral theory is that, in the face of moral diversity, multilevel social contract theory, which follows Hume's notion of evolutionary processes, combines evolution and reason *hierarchically*, whereas Gaus' theory, which follows Hayek's (1960, 1973) notion of evolutionary processes, combines them *sequentially*.

Also, whereas both theories respond to the challenge of moral diversity by defending decentralized, dynamic processes of moral explanation and justification, Gaus' (2011, 2016) moral theory makes room for moral diversity by dispensing with the ideal of justifying specific moral rules. Multilevel social contract theory, by contrast, makes room for moral di-

²⁴ See Kant ([1785] 1998, AK 4:389).

versity by integrating different contractarian moral theories and consideration of their adequate scope. Conceptually, multilevel social contract theory suspends the requirement of uniqueness for lower-level morality to allow for moral diversity. The theory defends uniqueness only when it is necessary to ensure principled conflict resolution after all other moral means have failed.

Overall, the notion of moral agency that multilevel social contract theory defends is complex, and demands the active exercise of agents' rational and affective capacities. Multilevel morality demands that agents are receptive to others and the specific type of moral interaction in which they engage so that agents always elicit the most substantial common moral ground for their moral interactions, especially under conditions of moral uncertainty and under-determination.²⁵ In practice, in order to avoid that agents are unable to act because too many considerations apply, which Scanlon (1998, 170) calls 'moral gridlock', or because too much uncertainty prevails, agents may idealize some of the features of their empirical environment and/or their counterparts, especially if further information is unavailable and inaction would cause significant harm. Also, in practice, moral heuristics may evolve that can serve as shortcuts for behavior in a morally fragmented world.

To conclude, because its goal is to settle the moral question fully and exclusively and/or to provide better understanding of the reasons for moral disagreement, traditional first-order moral theory often does not offer sufficient guidance to address the practical phenomenon of moral diversity and moral agency in contemporary societies. Multilevel social contract theory can address this phenomenon and, in doing so, it reorients moral theory. It reorients moral theory to be practical and accept that, in a morally diverse society, no single correct system of moral rules exists. Instead, many such systems work together simultaneously and hierarchically to protect the autonomy of agents maximally in morally diverse societies. As a result of its structure and nature, multilevel social contract theory suggests a practically sound notion of moral agency for morally diverse societies. In this article, I have systematically placed into context and developed the basic features of this notion of moral agency, even if space restrictions do not allow discussion of all its aspects. The developed interrelational and dynamic notion of integrated moral agency suggests one particular way for agents to make sense of their complex

²⁵ For discussion of some aspects related to ethical decision-making under risk and uncertainty, see Rowe and Voorhoeve (2018) and Rowe (2019).

realities in morally diverse societies and offers guidance with regard to treating others respectfully in a complex moral world.

REFERENCES

- Abizadeh, Arash. 2018. *Hobbes and the Two Faces of Ethics*. Cambridge: Cambridge University Press.
- Braham, Matthew, and Martin van Hees. 2018. "Voids or Fragmentation: Moral Responsibility for Collective Outcomes." *The Economic Journal* 128 (612): F95-F113.
- Caton, James. 2020. "Moral Community and Moral Order: Developing Buchanan's Multilevel Social Contract Theory." *Erasmus Journal for Philosophy and Economics* 13 (2): 1-29.
- Cottingham, John. 2010. "Integrity and Fragmentation." *Journal of Applied Philosophy* 27 (1): 2-14.
- D'Agostino, Fred. 2003. *Incommensurability and Commensuration: The Common Denominator*. Aldershot: Ashgate.
- D'Agostino, Fred, Gerald Gaus, and John Thrasher. 2017. "Contemporary Approaches to the Social Contract." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Article published March 3, 1996; last modified May 20, 2021. <https://plato.stanford.edu/entries/contractarianism-contemporary/>.
- Darwall, Stephen L., ed. 2003. *Contractarianism/Contractualism*. Oxford: Blackwell Publishing.
- Gaus, Gerald. 2011. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press.
- Gaus, Gerald. 2016. *The Tyranny of the Ideal: Justice in a Diverse Society*. Princeton, NJ: Princeton University Press.
- Gaus, Gerald. 2018a. "It Can't Be Rational Choice All the Way Down: Comprehensive Hobbesianism and the Origins of the Moral Order." In *Buchanan's Tensions: Reexamining the Political Economy and Philosophy of James M. Buchanan*, edited by Peter J. Boettke and Solomon Stein, 117-147. Arlington, VA: Mercatus Center.
- Gaus, Gerald. 2018b. "The Priority of Social Morality." In *Morality, Governance, and Social Institutions: Reflections on Russell Hardin*, edited by Thomas Christiano, Ingrid Creppell, and Jack Knight, 23-56. Cham: Palgrave MacMillan.
- Gaus, Gerald. 2018c. "Self-Organizing Moral Systems: Beyond Social Contract Theory." *Politics, Philosophy & Economics* 17 (2): 119-147.
- Gaus, Gerald. 2019. "Moral Conflict and Prudential Agreement: Michael Moehler's *Minimal Morality*." *Analysis* 79 (1): 106-115.
- Gaus, Gerald. Forthcoming. *The Open Society and Its Complexities*. New York, NY: Oxford University Press.
- Gauthier, David. 1986. *Morals by Agreement*. Oxford: Clarendon Press.
- Gauthier, David. 1997. "Political Contractarianism." *The Journal of Political Philosophy* 5 (2): 132-148.
- von Hayek, Friedrich A. 1960. *The Constitution of Liberty*. Chicago, IL: University of Chicago Press.
- von Hayek, Friedrich A. 1973. *Law, Legislation and Liberty. Volume 1: Rules and Order*. Chicago, IL: University of Chicago Press.

- Hobbes, Thomas. (1651) 1996. *Leviathan*. Edited by Richard Tuck. Cambridge: Cambridge University Press.
- Hume, David. (1739/1740) 2000. *A Treatise of Human Nature*. Edited by David Fate Norton and Mary J. Norton. Oxford: Oxford University Press.
- Kant, Immanuel. (1785) 1998. *Groundwork of the Metaphysics of Morals*. Translated and edited by Mary Gregor. Cambridge: Cambridge University Press.
- Kohl, Markus. 2018. "Kant's Critique of Instrumental Reason." *Pacific Philosophical Quarterly* 99 (3): 489–516.
- Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.
- Moehler, Michael. 2013. "Contractarian Ethics and Harsanyi's Two Justifications of Utilitarianism." *Politics, Philosophy & Economics* 12 (1): 24–47.
- Moehler, Michael. 2018. *Minimal Morality: A Multilevel Social Contract Theory*. Oxford: Oxford University Press.
- Moehler, Michael. 2019. "Replies to Gaus, Van Schoelandt and Cooper: Prudence, Morality and the Social Contract." *Analysis* 79 (1): 140–153.
- Moehler, Michael. 2020a. *Contractarianism*. Cambridge: Cambridge University Press.
- Moehler, Michael. 2020b. "Minimal Morality, Bargaining Power, and Moral Constraint: Replies to D'Agostino, Thrasher, Morris, and Vanderschraaf." *Analytic Philosophy* 61 (1): 87–100.
- Moehler, Michael. Forthcoming. "Strategic Justice, Conventionalism, and Bargaining Theory." *Synthese*.
- Morris, Christopher W. 2020. "Morality's Many Parts: Symposium on Michael Moehler's *Minimal Morality*." *Analytic Philosophy* 61 (1): 57–69.
- Muldoon, Ryan. 2016. *Social Contract Theory for a Diverse World: Beyond Tolerance*. New York, NY: Routledge.
- Müller, Julian F. 2019. *Political Pluralism, Disagreement and Justice: The Case for Polycentric Democracy*. New York, NY: Routledge.
- Nguyen, C. Thi. 2019. "Games and the Art of Agency." *The Philosophical Review* 128 (4): 423–462.
- Nguyen, C. Thi. 2020. *Games: Agency As Art*. Oxford: Oxford University Press.
- Ostrom, Vincent, and Elinor Ostrom. 2002. "Public Goods and Public Choices." In *Polycentricity and Local Public Economies: Readings from the Workshop in Political Theory and Policy Analysis*, edited by Michael D. McGinnis, 75–103. Ann Arbor, MI: University of Michigan Press.
- Parfit, Derek. 2011. *On What Matters. Volumes 1 and 2*. Oxford: Oxford University Press.
- Prichard, Harold A. 1912. "Does Moral Philosophy Rest on a Mistake?" *Mind* 21 (81): 21–37.
- Rawls, John. 2001. *Justice as Fairness: A Restatement*. Edited by Erin Kelly. Cambridge, MA: The Belknap Press of Harvard University Press.
- Rowe, Thomas. 2019. "Risk and the Unfairness of Some Being Better Off at the Expense of Others." *Journal of Ethics and Social Philosophy* 16 (1): 44–60.
- Rowe, Thomas, and Alex Voorhoeve. 2018. "Egalitarianism under Severe Uncertainty." *Philosophy & Public Affairs* 46 (3): 239–268.
- Scanlon, Thomas M. 1995. "Moral Theory: Understanding and Disagreement." *Philosophy and Phenomenological Research* 55 (2): 343–356.

- Scanlon, Thomas M. 1998. *What We Owe to Each Other*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Scanlon, Thomas M., and Alex Voorhoeve. 2001. "Kant on the Cheap." *The Philosophers' Magazine* 16: 29-30.
- Southwood, Nicholas. 2008. "A Deliberative Model of Contractualism." *Politics, Philosophy & Economics* 7 (2): 183-208.
- Southwood, Nicholas. 2010. *Contractualism and the Foundations of Morality*. Oxford: Oxford University Press.
- Southwood, Nicholas. 2019. "Contractualism for Us As We Are." *Philosophy and Phenomenological Research* 99 (3): 529-547.
- Suikkanen, Jussi. 2019. "Ex Ante and Ex Post Contractualism: A Synthesis." *The Journal of Ethics* 23 (1): 77-98.
- Valentini, Laura. 2021. "Respect for Persons and the Moral Force of Socially Constructed Norms." *Noûs* 55 (2): 385-408.
- Van Schoelandt, Chad. 2019. "Between Traditional and Minimal Moralities." *Analysis* 79 (1): 128-140.
- Van Schoelandt, Chad. 2020. "Functionalist Justice and Coordination." *Social Theory and Practice* 46 (2): 417-440.
- Watson, Gary. 1998. "Some Considerations in Favor of Contractualism." In *Rational Commitment and Social Justice: Essays for Gregory Kavka*, edited by Jules L. Coleman and Christopher W. Morris, 168-185. Cambridge: Cambridge University Press.
- Williams, Bernard. 1973. "A Critique of Utilitarianism." In *Utilitarianism: For and Against*, edited by John J. C. Smart and Bernard Williams, 77-150. Cambridge: Cambridge University Press.
- Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. London: Fontana Press.

Michael Moehler is Founding Director and Core Faculty member of the Kellogg Center for Philosophy, Politics, and Economics at Virginia Tech. He received his Ph.D. in philosophy from the London School of Economics. His research and teaching expertise lie in the history of moral and political philosophy, rational choice theory, public reason theory, distributive justice, the welfare state, and political economy. He is author of *Minimal Morality: A Multilevel Social Contract Theory* (Oxford University Press) and *Contractarianism* (Cambridge University Press).
Contact e-mail: <moehler@vt.edu>

The Paths to Narrow Identities

JEAN-PAUL CARVALHO

University of Oxford

I. MULTIDIMENSIONAL IDENTITIES

Up until the first agricultural revolution 12,000 years ago, the population of *homo sapiens* lived in small bands of nomadic families consisting of 25–50 members (Henrich 2015). They were hunter-gatherers, highly egalitarian, did not engage in economic exchange, and divided labor based on age and sex. Marriage was exogamous and patrilocal. When a band grew too large, it split. Identity was based on rules tied to ancestral lineage, and the scope for identity choice was limited. Since humans began to permanently settle and to domesticate plants and animals, there has been a dramatic increase in social complexity driven by the invention of writing, money, cities, democracy, world religions, scientific and industrial revolutions, nuclear power, space travel, and the internet, among many other developments. These products of cultural evolution have provided people with immense scope within which to develop and project ideas of who they are and what is their place in the world.

Our identities, both personal and social, are structured in a particular way, with multiple dimensions such as ethnicity, gender, class, religion, nationality, sexual orientation, political preferences, and cultural tastes. We might interact with a neighbor based on our geographic proximity and mutual interest in cooperation without any attention paid to, though possibly with full knowledge of, the other aspects of our identities which do not exactly match. This is not a surprise to anyone. It is a universal human experience and could be described as a natural state. What is puzzling is that at certain points in history, a dimensionality reduction takes place, and society becomes suddenly and radically simplified. Social interactions and conflict become narrowly organized around one salient dimension of identity, with all other dimensions switched off. A personal example may help to illustrate.

As a child in Sri Lanka, my family were caught up in the 1983 riots, which came to be known as Black July. When the riots broke out in our

area, my grandfather was at the family businesses. Before my father attempted to reach my grandfather in person, he phoned and told my grandfather that if he did not arrive by a specific time, my grandfather should leave and seek shelter in the neighboring Sinhalese-owned business. My grandfather knew the owner well and had rescued him financially on several occasions, so this arrangement was assumed to be safe. When my grandfather entered the premises, the owner, who was surrounded by several others, addressed him in the Tamil language, telling him curtly to 'sit down there'. In this way, my grandfather was marked out as Tamil, the ethnic minority being targeted by the communal violence. The peculiar thing is that, though my grandfather was indeed of Tamil ethnicity, he did not identify as such. As a member of the liberal cosmopolitan class, he almost never spoke the Tamil language (certainly never to his Sinhalese friend). In fact, all languages except for English were prohibited at home. The complex identity that he had built over the course of his life—businessman, film critic, husband, father, and so forth—had in a few hours of rioting been collapsed into one aspect, which he did not choose and which was neither central to his self-conception nor, up to that point, to his social identity. We can surmise that this sudden and imposed narrowing of his identity had as profound an impact on him as the looting and burning of his businesses. After meticulously organizing his life around work for over forty-five years, he never worked another day.

Unfortunately, such narrow identification—an extreme phenomenon—dominates theoretical and empirical work in economics to the point at which readers may lose sight of the broad identities they encounter everywhere in their daily lives. Understanding identity is an important step toward broadening the scope of economics to encompass the social and political environment in which market behavior is embedded. Inspired by the seminal work of Akerlof and Kranton (2000, 2010), economists are now examining the effect of identity on many forms of economic behavior. However, the concept of identity that has been mostly employed is unidimensional, with notable exceptions such as Sen (2006), Akerlof (2017), Sambanis and Shayo (2013), and Carvalho and Pradelski (2021). This is an important omission, and not just in the study of conflict. For example, Carvalho and Pradelski (2021) show that standard approaches to reducing structural inequality that treat identity dimensions as independent can be counterproductive. Due to spillovers across identity dimensions, such as race and gender, interventions that aim to reduce inequality along one identity dimension can increase inequality along another. What is re-

quired are more holistic approaches that account for the multidimensionality of identity and the connections between identity dimensions.

Nobody has done more than Amartya Sen to call such issues to the attention of economists and bring them within the scope of economic analysis. Sen shows how historical episodes of conflict are triggered by identity-based concerns and in particular “the odd presumption that the people of the world can be uniquely categorized according to some *singular and overarching* system of partitioning” (2006, xii; emphasis in the original). Sen’s proposed solution is simple. Identity-based conflict, Sen proposes, can be solved through epistemic means by escaping (mentally) from a singular conception of identity. Narrow identification is seductive and can be exploited by “artisans of terror” (2006, 2). The solution is to switch to the right, more complex conception of identity through individual will. Identity should not be viewed purely as inherited, but also as shaped by individual choice. I wholeheartedly agree with Sen that narrow identification is the exception, not the rule. It is not a natural state from which we must plot an escape route, but rather an extreme and unnatural one from which we must wonder how we got there. But then the question arises: if we have plural identities, why do narrow identities emerge at all? A clearer map of the paths to narrow identities will tell us how to better avoid identity-based conflict. Dasgupta and Goyal take up this question in their paper titled “Narrow Identities” (2019; henceforth DG).

II. THE DASGUPTA-GOYAL MODEL OF NARROW IDENTITIES

DG recognize that narrow identities form through social interactions and that groups play an important role in this process. This is a major advance. Because narrow identities are an equilibrium phenomenon they cannot be undone through individual will and right thinking alone. Specifically, DG’s model consists of a finite population of individuals, N , and two groups, A and B . Individuals and groups are *ex ante* identical. Individuals can choose to join one or both of the groups. Only the extensive margin (membership) is considered, not the time or effort devoted to each group. The payoff from joining group $k \in \{A, B\}$ is increasing in the size of group k and is also a function of the size of the other group $k' \neq k$. An individual i is said to have a narrow identity if i joins only one of the groups. Society can be said to have narrow identities when each individual $i \in N$ joins exactly one group. Note that narrow identities can exist in a monomorphic equilibrium (where all individuals join one group) or a polymorphic equilibrium (where different individuals join different groups). Disregarding membership costs, narrow identities do not occur without the interven-

tion of group leaders, who maximize aggregate group payoffs. The main conclusion of the paper is that narrow identities emerge as an equilibrium when (i) groups impose negative externalities on each other that are increasing in group size (for example, competition for scarce resources) and (ii) group leaders respond to this by imposing restrictions on dual membership to limit the size of the other group.

The DG model is a club model in the tradition of Iannaccone (1992) and the subsequent literature on religious clubs (Iannaccone 1998; Iyer 2016; Carvalho 2019). There are, however, notable differences. In the DG model, as in the religious club model, rules governing outside activity by group members play a critical role. There are two types of rules explored in the religious clubs literature. The first is stigmatizing behavioral practices and proscriptions imposed by religious groups, which act as a tax on outside activity (Iannaccone 1992; Aimone et al. 2013; Carvalho 2013; Carvalho and Koyama 2016). These rules contribute to reducing outside activity even when inputs to the club (for example, religious effort) are difficult to monitor. Second, religious clubs can impose a minimum participation constraint on group activity or, equivalently, cap the amount of time or money group members spend on outside activity (Carvalho 2016; Carvalho and Sacks, forthcoming). The restriction on group membership that arises in equilibrium in the DG model is closer to the second type of rule. The difference is that the DG model focuses on the extensive margin and does not consider the intensive margin, that is, the amount of time, effort, or money devoted to the group. The focus on membership choice, however, allows the authors to study participation in multiple groups, something not explored by religious club models. The purpose of membership rules in the DG model also bears some resemblance to the purpose of membership rules in the literature on religious clubs. In the religious club model, restrictions on outside activity play a strategic role in screening out uncommitted types and inducing club members to divert resources to the club. In this way, restrictions on outside activity limit the standard free-rider problem in collective production. In the DG model, they limit negative externalities generated by other groups. For example, in a conflict setting, restrictions on membership in the DG model mean that individuals need to ‘pick a side’, and cannot benefit regardless of which side wins. The difference is that the religious club model focuses on *intragroup* externalities while the DG model focuses on *intergroup* externalities, an important distinction to which I will return below.

II.I. Extensions

There are three immediate ways one can build on the work of DG in studying narrow identities and identity-based conflict. First, identities can be made explicitly multidimensional, as suggested by the context I have provided above. Of course, it is possible to interpret DG's model in this way, with each group $k \in \{A, B\}$ representing one dimension of an individual's identity. By choosing both groups, an individual identifies with both aspects of their identity. Otherwise, they have a narrow identity. Since each individual has the same choice set, $\{A, B\}$, this interpretation applies only to homogeneous populations. Second, the very notion of identity suggests *ex ante* heterogeneity, which could be built into the model. Third, multidimensionality and heterogeneity point to different means of defining narrow identities.

I will suggest one possible definition. Let each individual i 's identity be denoted by a vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ik}, \dots, x_{iK}) \in \mathbb{R}^K$, where x_{ik} is a coding of i 's identity in dimension k , for example, gender (male = 0, female = 1) or race (white = 0, black = 1). The distance between two identities \mathbf{x}_i and \mathbf{x}_j is not the standard Euclidean distance. Rather, each individual has a (weighted) perceived distance from every other individual.¹ Let i 's perceived distance from j be given by

$$d_i(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^K \theta_{ik} (x_{ik} - x_{jk})^2}, \quad (1)$$

where $\theta_{ik} \in [0, 1]$ is the weight i assigns to identity dimension k and $\sum_{k=1}^K \theta_{ik} = 1$. The broadness of i 's identity is captured by the distribution of weights over the identity dimensions. An identity is narrow if an individual assigns almost all weight to one identity dimension (for example, ethnicity).² Note that in some circumstances the weights are best thought of as homogeneous: $\theta_{ik} = \theta_k$ for all $i \in N$. In other circumstances, they are likely to vary across individuals.

Now consider a repeated population game in which players are matched in pairs and the payoff to a player i from interaction with j is decreasing in the perceived distance $d_i(\mathbf{x}_i, \mathbf{x}_j)$. In addition, for each player i , we can let

¹ In addition to requiring comparability across dimensions, this requires that identity be measured in each dimension according to an interval scale.

² One specific measure of the broadness of i 's identity is entropy (Shannon 1948):

$$H(\theta_i) = - \sum_{k=1}^K \theta_{ik} \log(\theta_{ik}).$$

the weights $\{\theta_{ik}\}_{k=1}^K$ evolve as a function of the history of play, including past choices in pairwise interactions, individual investments in identity, and political interventions. Thus, individuals can reduce their perceived social distance to some members of the population and increase their perceived social distance to others. In this way, the ‘singular and overarching system of partitioning’ described by Sen (2006) can come into being.

Note that DG allude at the end of their paper to a variant of their model in which individuals choose between exclusive (that is, narrow) identities, which increase social distance, and shared identities, which reduce social distance. This variant can be interpreted as a reduced-form version of the model suggested here. A version of it has already been studied by Carvalho (2017, section 3.1), who examines the conditions under which exclusive equilibria are stochastically stable. By analyzing a richer model of multidimensional identity, we could make further progress in understanding how narrow identities emerge.

III. PATHS TO NARROW IDENTITIES

While a formal model is beyond the scope of this comment, one can still distinguish between two different paths to narrow identities. Narrow identification is an extreme outcome, and extreme outcomes typically arise from positive feedbacks. The two paths correspond to positive feedback processes operating between and within groups. Similarly, Carvalho and Sacks (2021) analyze a dynamic model in which identity-based organizations can, under certain conditions, strengthen identification within an identity group over time. This can occur via different paths, most notably through biased cultural transmission (within-group) and endogenous discrimination (between-group). However, identity in their model is unidimensional, so the narrowing of identity is not examined.

III.I. Intergroup Dynamics

To illustrate, suppose each individual i can take an action e that is helpful or an action h that is harmful to their partner in an interaction. We know that individuals care more about ingroup than about outgroup members, even when there are minimal differences between groups (Chen and Li 2009). With multidimensional identities, we can suppose that action e yields a larger payoff to i than h if and only if the perceived social distance between i and j , $d_i(\mathbf{x}_i, \mathbf{x}_j)$, is sufficiently small. For each individual $i \in N$ and identity dimension $k = 1, \dots, K$, the identity weight θ_{ik}^t increases at time t if i is matched with a player who has a different identity

in dimension k and who chooses h . If e is chosen, the identity weights remain unchanged.

Consider a state in which identification is 'broad', that is, θ_{ik} is equally distributed across identity dimensions. Then e could be chosen by all players, and identification would continue to be broad. Let us now shock the system. Suppose that individuals either have identity 0 or 1 in dimension k , and let there be, for some reason, a sequence of plays of h whenever 0 and 1 types (in dimension k) are matched. Then θ_{ik}^t rises for all players in such matches. Eventually, h could become a best response. In this way, persistent conflict emerges between the 0-s and 1-s in identity dimension k , regardless of all they might have in common along other dimensions. The negative shock setting society down this path can come from decentralized forces. However, it can also be, and often is, engineered by political entrepreneurs. For example, in the case of Black July in Sri Lanka, electoral rolls and lists of Tamil-owned businesses were released to rioters so they could efficiently target Sri Lankan Tamil homes and businesses. This event itself was preceded by a gradual escalation of ethnoreligious conflict and led to a full-blown civil war (Tambiah 1992; Powell and Amarasingam 2017). This path to narrow identification through intergroup dynamics is the closest to the DG model of narrow identities.

III.II. Intragroup Dynamics

A second path, not considered by DG, is created by interactions within groups, as exemplified by the formation of cults. A cult is a strict sect whose doctrine is at variance with the mainstream culture in which it is located (Stark and Bainbridge 1985). Absorption into a cult can be thought of as a process of narrowing identity:

The same story makes the headlines again and again. An anguished family is trying to 'rescue' its child, who has, the parents charge, been 'stolen' by a cult, sometimes after only a single weekend of involvement. The parents describe the child as a humorless 'zombie' — where formerly he or she was self-possessed, intelligent and completely 'normal.' (Collins 1982, B5)

The three main theories of cult formation described by Bainbridge and Stark (1979)—the psychopathology model, the entrepreneur model, and the subculture-evolution model—have two common elements. Firstly, the cults provide *compensators*, that is, relationships, experiences, or material goods that members find missing in their regular lives. Perhaps

the most important compensator produced by such groups is a sense of meaning and belonging (Carvalho, forthcoming). Secondly, these group-specific goods are produced through social interaction within the group. In particular, the subculture-evolution model views cults as “the expression of novel social systems, composed of intimately interacting individuals who achieve radical cultural developments through a series of many small steps” (Bainbridge and Stark 1979, 283).

This can be modeled within the framework suggested here as follows. Each individual chooses whether to join one of a number of groups or to interact in an unrestricted manner in mainstream society. Suppose there is a set of individuals $G \subset N$ whose members have a rare trait in dimension k , for example, a rare cosmological belief. (They could well have mainstream traits in all other dimensions.) Suppose also that interacting in a group increases the weight on identity dimensions in which there is low within-group variation but high between-group variation. Then, if members of the set G were to find each other and form a group, θ_{ik} would rise for $i \in G$. This shift in identification makes the group more valuable to members. As such, a group leader could elicit larger contributions to club goods, making group membership even more valuable (see Carvalho 2016, 2020). Through such a process, the group can gradually become more cohesive and group members more narrowly identified with their rare dimension- k identity. This is one possible model of cult formation. Note that such groups are mostly nonviolent. However, under certain conditions, strict clubs can transition to violent activity (Berman 2009; Berman and Laitin 2008) or be infiltrated by militants (Carvalho, forthcoming).

IV. CONCLUDING REMARKS

DG's paper is an important advance in understanding identity formation and conflict. As narrow identities are formed through social interactions regulated by groups, narrow identification cannot be undone through individual will and right thinking alone. In the DG model, narrow identities emerge as an equilibrium when groups impose negative externalities on each other that are increasing in group size and group leaders respond to this by imposing restrictions on dual membership to limit the size of the other group. By examining membership in multiple groups, DG perform an extension of the standard club model in the economics of religion. The DG model paves the way for a fuller, dynamic analysis of the formation of narrow identities. In this comment, I have made several suggestions for extensions, including (1) explicitly multidimensional identities,

(2) heterogeneity, and (3) a new definition of narrow identities. In terms of dynamics, there is an important distinction to be made between positive feedback effects within and between groups. Of course, much more needs to be done, both theoretically and empirically, to flesh out these models. Such work could enrich our understanding of identity formation and thereby help to mitigate identity-based conflict.

REFERENCES

- Aimone, Jason A., Laurence R. Iannaccone, Michael D. Makowsky, and Jared Rubin. 2013. "Endogenous Group Formation via Unproductive Costs." *The Review of Economic Studies* 80 (4): 1215–1236.
- Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *The Quarterly Journal of Economics* 115 (3): 715–753.
- Akerlof, George A., and Rachel E. Kranton. 2010. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton, NJ: Princeton University Press.
- Akerlof, Robert. 2017. "Value Formation: The Role of Esteem." *Games and Economic Behavior* 102: 1–19.
- Bainbridge, William Sims, and Rodney Stark. 1979. "Cult Formation: Three Compatible Models." *Sociology of Religion* 40 (4): 283–295.
- Berman, Eli. 2009. *Radical, Religious, and Violent: The New Economics of Terrorism*. Cambridge, MA: The MIT Press.
- Berman, Eli, and David D. Laitin. 2008. "Religion, Terrorism and Public Goods: Testing the Club Model." *Journal of Public Economics* 92 (10–11): 1942–1967.
- Carvalho, Jean-Paul. 2013. "Veiling." *The Quarterly Journal of Economics* 128 (1): 337–370.
- Carvalho, Jean-Paul. 2016. "Identity-Based Organizations." *American Economic Review* 106 (5): 410–414.
- Carvalho, Jean-Paul. 2017. "Coordination and Culture." *Economic Theory* 64 (3): 449–475.
- Carvalho, Jean-Paul. 2019. "Religious Clubs: The Strategic Role of Religious Identity." In *Advances in the Economics of Religion*, edited by Jean-Paul Carvalho, Sriya Iyer, and Jared Rubin, 25–41. Cham: Palgrave.
- Carvalho, Jean-Paul. 2020. "Sacrifice and Sorting in Clubs." *Forum for Social Economics* 49 (4): 357–369.
- Carvalho, Jean-Paul. Forthcoming. "Religion and Terrorism: The Religious Utility Hypothesis." In *Handbook of the Economics of Terrorism*, edited by Atin Basuchoudhary and Gunther Schulze. Cambridge: Cambridge University Press.
- Carvalho, Jean-Paul, and Mark Koyama. 2016. "Jewish Emancipation and Schism: Economic Development and Religious Change." *Journal of Comparative Economics* 44 (3): 562–584.
- Carvalho, Jean-Paul, and Bary Pradeliski. 2021. "Identity and Underrepresentation: Interactions between Race and Gender." Working Paper; first version January 2, 2019; revised May 3, 2021. Available at SSRN. <https://ssrn.com/abstract=3299477>.
- Carvalho, Jean-Paul, and Michael Sacks. 2021. "Radicalization." Working Paper; first version December 26, 2018; revised March 6, 2021. Available at SSRN. <https://ssrn.com/abstract=3297267>.

- Carvalho, Jean-Paul, and Michael Sacks. Forthcoming. "The Economics of Religious Communities." *Journal of Public Economics*.
- Chen, Yan, and Sherry Xin Li. 2009. "Group Identity and Social Preferences." *American Economic Review* 99 (1): 431–457.
- Collins, Glenn. 1982. "The Psychology of the Cult Experience." *The New York Times*, March 15, 1982, B5. <https://www.nytimes.com/1982/03/15/style/the-psychology-of-the-cult-experience.html>.
- Dasgupta, Partha, and Sanjeev Goyal. 2019. "Narrow Identities." *Journal of Institutional and Theoretical Economics* 175 (3): 395–419.
- Henrich, Joseph P. 2015. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.
- Iannaccone, Laurence R. 1992. "Sacrifice and Stigma: Reducing Free-Riding in Cults, Communes, and Other Collectives." *Journal of Political Economy* 100 (2): 271–291.
- Iannaccone, Laurence R. 1998. "Introduction to the Economics of Religion." *Journal of Economic Literature* 36 (3): 1465–1495.
- Iyer, Sriya. 2016. "The New Economics of Religion." *Journal of Economic Literature* 54 (2): 395–441.
- Powell, Christopher, and Amarnath Amarasingam. 2017. "Atrocity and Proto-Genocide in Sri Lanka." In *Understanding Atrocities: Remembering, Representing, and Teaching Genocide*, edited by Scott W. Murray, 19–47. Calgary: University of Calgary Press.
- Sambanis, Nicholas, and Moses Shayo. 2013. "Social Identification and Ethnic Conflict." *American Political Science Review* 107 (2): 294–325.
- Sen, Amartya K. 2006. *Identity and Violence: The Illusion of Destiny*. New York, NY: W. W. Norton.
- Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423.
- Stark, Rodney, and William Sims Bainbridge. 1985. *The Future of Religion: Secularization, Revival and Cult Formation*. Berkeley, LA: University of California Press.
- Tambiah, Stanley Jeyaraja. 1992. *Buddhism Betrayed? Religion, Politics, and Violence in Sri Lanka*. Chicago, IL: University of Chicago Press.

Jean-Paul Carvalho is Associate Professor of Economics at the University of Oxford and fellow of New College. He is a specialist in the fields of political economy and social dynamics.

Contact e-mail: <jean-paul.carvalho@economics.ox.ac.uk>

Deepening and Widening Social Identity Analysis in Economics

JOHN B. DAVIS

Marquette University and University of Amsterdam

I. DASGUPTA AND GOYAL'S CONTRIBUTION

Standard rationality theory explains individual behavior as utility maximizing in every possible circumstance and situation individuals may encounter. No matter how the world is organized or institutionally structured, individuals always behave in one single way. As a universal, top-down characterization of behavior, it rules out there being any evidence that might disconfirm it. As Paul Samuelson observed long ago, the utility maximization hypothesis “is in the nature of an axiom, or definition, not subject to proof in any empirical sense, since any and all types of observable behaviour might conceivably result from such an assumption” (Samuelson 1937, 156). Employed, then, without fear of refutation, by contemporary standards of science where theories are accepted or rejected according to how well they stand up to the evidence the world provides, the utility maximization hypothesis falls short.

One way to correct this failing is to incorporate the concept of identity into how economics characterizes individuals. Asking how we identify individuals ties their behavior to ‘who’ they are (Kirman and Teschl 2004), who they are reflects how they occupy the world, and in principle this allows us to determine whether our characterizations of their behavior are refutable—that is, whether these characterizations fit what we observe about the world. One might say attention to identity dampens the universalism most standard rationality accounts of behavior assume. So Dasgupta and Goyal (2019; henceforth DG) are to be commended for incorporating the concept of identity into their analysis of individual behavior.

They are also to be commended for making the concept of social identity central. There are many ways to investigate how social factors influence individual behavior, but using the concept of social identity has a distinct advantage—it organizes those factors in a framework that has

been extensively investigated, both conceptually and empirically, for many years in social psychology. Needless to say, the amorphous and ambiguous concept ‘social’ has the potential to create endless disputes over its scope and meaning. The social identity concept limits and determines what counts as social by elaborating a specific set of behavioral hypotheses associated with well observed social circumstances where people are said to have social identities.

Yet, DG neglect that social psychologists distinguish *two* main types of social identity referred to as *relational* social identities and *categorical* social identities. Relational social identities are seen as “identifications of the self *as* a certain kind of person”—a role-based social identity—while categorical social identities involve “identifications of the self *with* a group [or category] as a whole”—a collective identity (Thoits and Virshup 1997, 106; emphasis added). For example, in the case of relational social identities, people identify themselves *as* employers when they employ others or *as* employees relative to their employers; similarly with students and teachers, people sharing households, and many other social settings that involve role relationships. In the case of categorical social identities, people identify *with* others who, for example, share the same ethnicity or gender, and they do so independently of interacting with these others in role relationships or indeed of ever even coming into contact with them.

These two types of social identities, then, are sometimes treated in social psychology as two dimensions or

two levels of [people’s] social selves—[(i)] those that derive from interpersonal relationships and interdependence with specific others and [(ii)] those that derive from membership in larger, more impersonal collectives or social categories. (Brewer and Gardner 1996, 83)¹

The reason for saying that they involve two dimensions or ‘levels’ of a person is that the two types of social identity involve two quite different kinds of behavioral motivations.

DG, however, build their analysis exclusively around social-group categorical social identities as if relational social identities either do not exist or can be ignored for the purposes of their analysis. In one variation of DG’s (two-group) model, “a group’s payoff [to members] is an increasing function of its own size but a decreasing function of the size of the other group” (2019, 397). Their theory, they point out, is reminiscent of club

¹ See also Brewer (2001), Reynolds, Turner, and Haslam (2003), and Davis (2011, 201ff).

theory but it interprets clubs in social-group identity terms and is then modeled in such a way that social-group (or club) membership can be understood in terms of how social groups compete in attracting individuals and maintaining membership levels. Over time, DG argue, the rules that groups employ become endogenous to the competition for members between groups.² Recall that, in the standard taxonomy of goods, club goods (also termed local public goods) are defined as goods that are excludable and non-rivalrous for particular groups of people.

One reason for focusing exclusively on people's social-group categorical social identities is that this makes social identity analysis primarily about the importance of social groups—a relatively unexplored subject in economics. This seems to be what DG are doing. The particular challenge this strategy then faces is how to limit intersectionality, or people having multiple social-group identities and belonging to different groups—what DG call 'multiple identities'. If people move back and forth between different social-group categorical social identities, because they believe they belong to different groups, cross-group mobility becomes important. In that case, group membership explains less of their behavior than in the case where people maintain strong loyalties to one or only a few groups. DG's payoff analysis, then, rules out this sort of outcome by showing how it can be in individuals' interest to maintain strong loyalties to one or only a few groups and to narrow their social-group profiles.

What, then, does setting aside people's relational social identities miss? First, it obviously misses incorporating social roles into an economic social identity analysis. Second, it misses how relational social identities are also a source of categorical social identities and not just one of two types of social identity. Thus, not only does one have a relation to others in social roles *as* a kind of person, but one also shares this identity *with* many others outside these role relationships. For example, students see themselves in relational social identity terms *as* students in virtue of their interaction with their teachers, but they also share the identity of simply being students in categorical social identity terms because they identify *with* other students.

Consider the difference, then, between a role-based type of categorical social identity and a categorical social identity simply involving identification with others outside of role relationships. Categorical social identi-

² A similar argument about group dynamics was previously made by Horst, Kirman, and Teschl (2007).

ties specifically attached to relational social identities mix personal acquaintance with others in those role relationships with impersonal association with others outside of them. Thus, such identities are behaviorally more complicated than categorical social identities *not* connected to social roles in that the former somehow combine two kinds of motivation: one based in regular personal contact with people in role settings and the other in impersonal recognition of shared social categories. That is, students are motivated both by how they see themselves relationally *as* students vis-à-vis their teachers and also by how they identify categorically *with* other students.

In this more complicated behavioral situation, the issue is how the motivations associated with personal acquaintance, interaction in a specific relational setting, and the role responsibilities this involves connect up and interact with the more impersonal sort of motivation associated with identifying with others one doesn't know and will likely never meet. Does one kind of motivation dominate or take precedence in people's decision-making? Do role relationships persist or break down in light of people's perceived categorical identities? Or, apropos DG, do people's associated categorical social identities matter more or less in role settings?

Note an important difference, then, in where people's understanding of the two different types of categorical social identities comes from. In the case of categorical social identities per se, social scientists and other third-party individuals in public media, government, and elsewhere have an important say in ascribing social categories to individuals. In contrast, in the case of categorical social identities that arise specifically out of relational settings, it is individuals themselves, who, based on their own experience—for example, students studying with other students in relational settings—determine whether it is important to them to act upon a given social category ascribed to them by others. That is, in the first case, categorical social identities are largely ascribed to people independently of their own experience while, in the second case, role-based categorical social identities also depend on self-ascription. I return to this difference in connection with the issue of individual agency in the next section.

The vision, then, that DG pursue is one in which categorical social identities independent of relational ones are our only concern. In this framework, people have increasingly *narrow* social group identities associated with how the payoffs to individuals of maintaining primary loyalties to at most a few groups tend to rise. Intersectionality is potentially

an issue for this framework, but DG's payoff analysis makes it a non-issue.

This makes the world today a space of more and more intense social group rivalries—a world that is arguably increasingly polarized into largely non-communicating social groups locked into a competition with one another; a competition brought about by utility-maximizing individuals' impersonal, rational calculation of the payoffs to increasingly exclusive social-group loyalties. A possible corollary here is that democracy may matter less and less in this world because people ultimately vote according to their perceived group interests, and groups are more likely to promote partial interests rather than group-independent conceptions of the common good. In a time of increasing political partisanship, DG can hardly be blamed for seeking to model the social world in this way.

I will argue, however, that there is a more serious problem involved in focusing exclusively on people's categorical social identities, namely, that it risks eliminating individuals as distinct and independent agents which is fundamental to explaining what individuals' personal identities are. DG recognize that people have both personal and social identities, but do not explain what the former involves, thus leaving unaddressed what the relationship between the two is. This comes out in the inherent ambiguity associated with saying that people 'identify with' others.

II. THE PROBLEM OF INDIVIDUAL AGENCY

The categorical social identity sense of 'identify with' that DG employ is where individual agents cognitively associate themselves with others in social groups and yet still maintain their independence as individual agents while doing so—a weak sense of 'with' that preserves the person's individual autonomy. Here, the identification with others is done under the individual's control. Yet, the 'identify with' idea can also mean that individuals give up their status as independent individual agents when they associate themselves with others in social groups—a strong sense of 'with' in which the person's status as an individual is unclear. In effect, people may become representative agents of those social groups³ and substitute group utility functions for their own under pressure from others—in that case, the social group, not the individual, effectively becomes the agent. Indeed, this was the meaning of 'identify with' that was investigated in the famous Robbers Cave experiment (Sherif et al. [1961] 1988),

³ Note that this is not the macro-level representative agent concept of new classical economics but instead a micro-level one with a different purpose.

cited by DG (2019, 401). The experiment influenced many social psychologists and social identity theorists to argue that at least in some circumstances—particularly when individuals feel they are at significant risk—people cease to act as independent individuals and form strong allegiances to social groups. Indeed, this is one way to explain despotism in mass movements in history.

While DG clearly accord considerable weight to the influence social groups have on people, they clearly hold the weak sense of ‘identify with’ that preserves individual agency since, in their model, people are individual utility maximizers. DG also do not say that groups are also agents, even if this might be an interpretation or possible implication of their model. DG are, like many economists, ontological individualists who hold that only individuals exist and that only individuals are agents.

Yet, given the two possible meanings of ‘identify with’, it is incumbent upon DG, and those who hold the weak sense, to say how individuals retain a status as independent individual agents when group loyalties may count as important factors in explanations of their behavior. In this regard, restricting their social identity analysis to categorical social identities (and setting aside relational ones) creates a problem for DG, since the social distance individuals have from those with whom they identify in the case of categorical social identities—as manifest in the impersonality of this relation—diminishes their individuality and thus what might distinguish them as distinct and independent individuals. Essentially, if individuals see themselves as being simply like others, they become indistinguishable from others. In contrast, as I will argue more fully below, relational social identities, because they depend on personal role relationships, potentially differentiate people from one another according to how those role relationships distinguish them *as* certain kinds of people.

I’ve previously argued (for example, in Davis 2011) that to be able to say people are distinct, independent individuals, one needs to provide an adequate account of how they are individuated or shown to be distinct and different from one another. Individuation, that is, is an identity criterion (or test) for picking out individual agents (of any kind); and such a criterion needs to be satisfied when explanations and models are populated by agents, if these explanations and models are to be assumed to refer to real agents in the world.⁴

⁴ Of course, one could argue that the theory is normative, not descriptive of real-world individuals, and that it states what people ought to do in order to be rational. But DG presumably see their theory as descriptive.

I've also argued, however, that the individuation criterion implicitly assumed for utility-maximizing individuals is circular and inadequate because it cannot show that people are distinct and independent individuals. It relies on saying that a person is a distinct and independent individual in virtue of having their *own* preferences or utility function. However, it's question-begging to say that something is a distinct and independent entity because it has its *own* characteristics. So, despite the fact that DG assume people are distinguishable, independent individuals, their utility function analysis does not give us grounds for supposing that it is the weak sense of 'identify with' that applies in the world they envision.⁵

To be clear, I am also an ontological individualist. I believe individuals exist and can act as independent individual agents, though there are also social circumstances when they do not do so—I think human history has repeatedly, sadly demonstrated this when we examine the uptake on totalitarianism. From this I infer that we need to develop explanations of behavior that show us where the grounds lie for individuals to be able to behave and be explained as independent individual agents; then, we need to normatively and institutionally try to make the social circumstances involved objects of social policy. From this perspective, let us return to how relational social identities work.

III. RELATIONAL SOCIAL IDENTITIES AS A BASIS FOR EXPLAINING INDIVIDUAL AGENCY AND PERSONAL IDENTITY

Again, relational social identities concern how people in role settings identify themselves *as* certain kinds of persons, but because roles often refer to broad social categories, they also involve identifying people *with* others via their categorical social identities. In order, then, to distinguish these kinds of categorical social identities from those not connected to social roles, we might say that the former are nested or embedded within those relational social identities. This, I will argue, offers a way of explaining how the different kinds of motivations associated with the two kinds of social identities fit together, and how people can then be individuated as distinct and independent economic agents.

Generally speaking, individual agency is understood as a person's ability to bring certain things about. In social roles, then, people are seen as being able to bring things about because having a role implies a person

⁵ I believe the source of this failure, as Samuelson saw, is that the utility maximization hypothesis is taken as an axiom or a definition.

is expected to do certain things, and this assumes that the person is able to do such things. In contrast, having categorical social identities *per se* tells us nothing about individual agency since simply possessing a categorical social identity does not tell us that there are certain things people are expected to do. They simply concern a property assigned to people by third parties without any reference to agency. However, in the case of categorical social identities that specifically arise out of relational social identities, there do exist expectations, framed by people's associated social roles, regarding what people are expected to do.

For example, when a student says that a teacher is not acting as a teacher should, and when the student uses how she compares herself categorically to other students, the standard for this cross-student comparison and identification is what teacher-student roles generally involve. That is, the student's interpretation of her shared student categorical social identity is nested or embedded in the relational teacher-student social identity. Thus, to answer the question above regarding how relational social identity motivations and categorical social identity motivations connect up or are related, the former dominate and take precedence in determining how people understand the latter.

This, then, bears on the issue of individuation and what is needed to show that people are distinct and independent individuals. They are distinct and independent individuals if they can be seen to be individual agents, and they can be seen to be individual agents in social roles where they have responsibilities they are expected to be able to individually fulfill. So, fulfilling a specific role distinguishes a person from others in relation to others. Roles are, of course, largely given, but fulfilling them is up to the person who accepts them.⁶

In contrast to the standard utility-function conception of the individual, where individuating people in terms of only their own characteristics is circular, in this case, people are individuated not solely in terms of their own characteristics, but also in terms of how their characteristics are understood in relational role settings. Thus, as a conception of personal identity, people can then generally be individuated as distinct and independent individuals to the extent that across their many different role relationships in workplaces, households, and so on, people are expected to act on the various responsibilities these relationships involve. Personal

⁶ Previously (Davis 2003, 145–147), I used collective intentionality theory—how shared intentions relate individuals behaviorally—to explain how roles and responsibilities individuate people, but the simple idea of role fulfillment also explains this.

identity is thus relational; we are distinct and independent individuals relative to others with whom we interact.⁷

This view accordingly implies a different vision of society than the one DG offer. People's categorical social identities are indeed important, but their impact on society is mediated by people's social roles and relational social identities—at least if we wish to argue that people are distinct and independent individuals (and contrary to the hypothesis that social groups are agents rather than individuals). On this view, the world need not be seen as becoming increasingly polarized across rival, competing social groups. Rather, social conflict lies in whether role relationships operate as they are functionally expected to do or are prey to manipulation and unfairness undermining individual autonomy, particularly where some people have power over others. I touch on this issue below.

IV. RELATIVELY CLOSED AND RELATIVELY OPEN BEHAVIORAL DOMAINS

The social identity analysis above raises a difficult methodological issue regarding how one explains the interaction of economic agents' different behavioral motivations. This issue does not arise when utility-maximizing behavior is taken to be the only type of economic behavior. But people's two kinds of social identities produce two different kinds of behaviors.

Suppose, then, that we regard the different types of behaviors associated with role settings and group affiliations as occupying two different behavioral domains. People's relational social identities and social roles refer us to a domain where people are motivated by the responsibilities they believe they have—a domain of deontological behavior. Following DG, categorical social identities refer us to a domain where people act according to payoffs they expect from ties to social groups—a domain of instrumental behavior. An economic approach to social identity analysis, distinguishing people's relational and categorical social identities, could then be developed around explanations of the relationship between these different domains of life.

The analysis above made one key assumption in this regard. It argued that if we take individual agency seriously and focus on what makes people distinct and independent individuals, their role-relationship behaviors should have prior importance and their social-group behaviors should be

⁷ There is, of course, much more that needs to be said for a complete relational account of personal identity. Any relational account, however, shares the idea that without relationships to others, it is impossible to say what one is—thus, a human being deprived of such relationships throughout their life would, on this view, not count as an individual agent.

seen as modifying how people address these behavioral priorities. Again, when a student says that a teacher is not acting properly when the teacher judges the student by her ethnicity or gender rather than by how she behaves as a student, the standard for this complaint is what the teacher's role requires, and remedying the situation involves emphasizing the responsibilities of that role.

Note, then, that social roles, because they are usually structured around interconnected sets of rules, constitute a 'relatively closed' type of behavioral domain. The personal character of role settings—people being in contact and regular interaction with others—arguably reinforces their 'relatively closed' nature. In contrast, whereas social roles involve relatively well-structured domains of activity, people's categorical social identities motivate a more diffuse set of behaviors and thus constitute a 'relatively open' type of behavioral domain. That one possesses a social-group identity does not point toward specific kinds of behaviors other than acting in the interest of the group, which can mean many things. The 'relatively open' character of this domain is also reinforced by the social distance between people who share categorical social identities and the more impersonal connection this involves. Moreover, that people have multiple categorical social identities competing for their attention moderates the weights they place on any single social group identity.

Thus, the behavioral complexity associated with people having two kinds of social identities can be explained in terms of how the 'relatively closed' relational social identity domain interacts with the 'relatively open' categorical social identity domain. What, then, the priority given to social roles and the modifying effects upon them that group identities have tells us is that the effects of the latter domain work within and not simply from outside the social roles domain. The social roles domain is only 'relatively closed', not fully closed, because categorical social identities matter when we attend to the performance of responsibilities attached to social roles. At the same time, the domain constituted from people's categorical social identities is only 'relatively open', not fully so, because it is still structured, if loosely so, by intersectionality and by the fact that people have multiple social identities.

One implication of this sort of conception lies in how we understand normative economics and order different ethical goals in economics. Standard normative economics makes efficiency its principle, if not exclusive prescription, on the grounds that all behavior is utility maximiz-

ing. Yet, if social roles are of primary importance, because only they explain how people are distinct and independent economic agents, then economics should make the fulfillment of role responsibilities an important basis for its prescriptions. That is, in a world of ever-increasing specialization and continued expansion of social roles, economic gains are not the product of continual extension of the scope of utility maximization in the world, but of assuring that people's behavior efficiently fulfills what their roles entail.

However, there is a problem with this understanding since social roles in many instances embed and sustain discrimination against individuals according to their social-group memberships. Role relationships often place one party in a position of authority, and this allows them to abuse those relationships when it serves their interests.⁸ Thus, re-orienting normative economics also needs to give emphasis to values associated with combatting social discrimination—a concept of justice tied to the idea that people should be valued as individuals, not according to their social characteristics.

DG, then, are to be commended for incorporating the concept of identity into their analysis of individual behavior. Their framework allows us to raise new issues regarding economic behavior and to potentially expand the normative thinking underlying economics. This comment seeks to further extend DG's framework by deepening and widening social identity analysis by emphasizing the differences between people's relational and categorical social identities.

REFERENCES

- Brewer, Marilynn B. 2001. "The Many Faces of Social Identity: Implications for Political Psychology." *Political Psychology* 22 (1): 115–125.
- Brewer, Marilynn B., and Wendi Gardner. 1996. "Who Is This 'We'? Levels of Collective Identity and Self-Representations." *Journal of Personal and Social Psychology* 71 (1): 83–93.
- Dasgupta, Partha, and Sanjeev Goyal. 2019. "Narrow Identities." *Journal of Institutional and Theoretical Economics* 175 (3): 395–419.
- Davis, John B. 2003. *The Theory of the Individual in Economics: Identity and Value*. London: Routledge.
- Davis, John B. 2011. *Individuals and Identity in Economics*. Cambridge, NY: Cambridge University Press.
- Davis, John B. 2021. "A General Theory of Social Economic Stratification: Stigmatization, Exclusion, and Capability Shortfalls." Unpublished Manuscript.

⁸ In Davis (2021), I discuss how people in positions of power in role relationships can discriminate against others by manipulating intersectionality.

- Horst, Ulrich, Alan Kirman, and Miriam Teschl. 2007. "Changing Identity: The Emergence of Social Groups." Economics Working Paper No. 0078. Institute for Advanced Study, School of Social Sciences, Princeton, NJ.
- Kirman, Alan, and Miriam Teschl. 2004. "On the Emergence of Economic Identity." *Revue de Philosophie Économique* 9: 59–86.
- Reynolds, Katherine J., John C. Turner, and S. Alexander Haslam. 2003. "Social Identity and Self-Categorization Theories' Contribution to Understanding Identification, Salience and Diversity in Teams and Organizations." In *Research on Managing Groups and Teams. Volume 5: Identity Issues in Groups*, edited by Jeffrey Polzer, 279–304. Bingley: Emerald Publishing.
- Samuelson, Paul A. 1937. "A Note on Measurement of Utility." *The Review of Economic Studies* 4 (2): 155–161.
- Sherif, Muzafer, O. J. Harvey, B. Jack White, William R. Hood, and Carolyn W. Sherif. (1961) 1988. *The Robbers Cave Experiment: Intergroup Conflict and Cooperation*. Middletown, CT: Wesleyan University Press.
- Thoits, Peggy A., and Lauren K. Virshup. 1997. "Me's and We's: Forms and Functions of Social Identities." In *Self and Identity: Fundamental Issues*, edited by Richard D. Ashmore and Lee Jussim, 106–133. Oxford: Oxford University Press.

John B. Davis is Professor Emeritus of Economics at Marquette University and Professor Emeritus of Economics at the University of Amsterdam. He is the author of *Keynes's Philosophical Development* (Cambridge University Press, 1994), *The Theory of the Individual in Economics: Identity and Value* (Routledge, 2003), *Individuals and Identity in Economics* (Cambridge University Press, 2011), and co-author with Marcel Boumans of *Economic Methodology: Understanding Economics as a Science* (Palgrave, 2010). He has been a visiting professor at the Sorbonne, Cambridge University, Erasmus University Rotterdam, and Duke University. He was editor of the *Review of Social Economy* (1987–2005), co-editor with Wade Hands of *The Journal of Economic Methodology* (2005–2019), and is the editor of the book series "Routledge Advances in Social Economics". He is a past president or chair of the History of Economics Society, the International Network for Economic Method, and the Association for Social Economics, and a past vice-president of the European Society for the History of Economic Thought.

Contact e-mail: <john.davis@marquette.edu>

Social Identities: Narrow and Broad, Exclusive and Inclusive, Firm and Fuzzy

PETER FINKE

University of Zurich

Disciplinary boundaries certainly have their reasons and advantages, for example, regarding the possibility to specialise and the common ground they create for on-going debates. But this comes at a price, as we all know. One is that I would probably never come across a piece of work like Dasgupta and Goyal's "Narrow Identities" (2019; henceforth DG) were it not for the kind invitation to participate in this (virtual) mini-symposium. I would therefore like to thank the organisers and editors of this issue, as well as the authors of the original paper, for the opportunity to raise an anthropological word in this context.

The paper sets itself the aim of explaining why people around the world tend to stress narrow or exclusive identities, such as ethnicity, when each one's personhood is composed of so many more distinctions. This is, indeed, a remarkable fact, and one with great—and often fatal—consequences. It is highly appreciated to see this approached in such a meticulous and consistent way as only economists tend to do. I write this as an economic anthropologist who has been working on identity issues for many years, and who is, in general, amicable to the rigour and logical derivation of mathematical models and causal explanations—more so, at least, than many of my colleagues.

Having said that, I was presumably not invited to simply back the arguments and conclusions made here but to take a critical look from a different theoretical perspective and methodological background. As indicated, some of these differences are disciplinary or terminological. To begin with, I fully agree with the authors that individual and social identities may be analytically distinct but always closely interrelated, and thus, should be looked at jointly. I also strongly agree with the point made in the very beginning that a person's identity has indeed many facets, such as the mentioned "language, personal interests, customs, religion,

and ethnicity” (DG 2019, 395; missing here are issues of gender and class, for example). Yet, when in later passages the word is about the multiplicity of identity, I wonder whether complexity would not fit better as a description relating to facets. All these various aspects do not stand parallel or in opposition to each other but, at least for most of us, add up to make us the personalities we are (Finke 2014). It took me also a while to grasp what the authors actually mean by narrow identities. But if it is the opposite of multiple, shouldn't it be single identities, or if narrow, shouldn't its counterpart be broad identities? Or inclusive and exclusive? And isn't much of what the authors write about more the—often opportunistic—affiliation with particular groups and categories rather than an identity that tells us who we are or believe to be? Membership is not the same as identification, although the two may often go together. But, as said, these are terminological issues, not a question of disagreement on an argument.

Most of the critical issues I shall raise here have to do with assumptions the authors make; others point to neglected aspects that, for an anthropologist or social scientist, are essential to think about. Let me take them in turn.

I. GROUP SIZE

The problem of the paper starts with a couple of assumptions, not all of which are totally convincing. Yes, as the authors claim, the main function of identity seems to be for people to find their place in the larger social world, and to include guidelines on how to act accordingly as well as a commitment to do so. This, in turn, enables mutual trust and solidarity. I also follow the idea that people evaluate the pros and cons of joining or leaving groups depending on the respective benefits these have to offer. I am not so sure about the assumption that, all else equal, the larger one's *own* group is, the better off one is. First of all, while this undoubtedly increases the chance to achieve common goods, in many situations, it also means sharing the booty with more members than one would want to, depending on what is at stake. If you want to save the world from climate change, the more collaborators the better. The same is not the case if you plan to rob a bank or win a Nobel prize. Therefore, in alliance theory and related bodies of literature, the trick of the game is rather to end up at the optimal size to just succeed rather than to maximise one's followers (Schlee 2004). Building coalitions in many European parliaments is a telling example of that. The optimal size here is 51 per cent. This allows you

to assert your aims without having too many parties on board with their own interests and stakes.

In other cases, exclusiveness, or the strict limitation of membership in a given group, may be a much sought-after advantage, for example, in the case of aristocratic estates or religious factions built on ideas of purity and esoteric wisdom. Here, the smaller the own group the better, as long as this distinction towards others manifests in some sort of benefits in itself. In turn, narratives of the true faith being narrowly defined and only accessible for a small minority are of greater relevance here.

II. METHODS

A greater concern I have, and one which I presumably share with more or less all of my disciplinary colleagues, is about the methodology. Mathematical models are, of course, a perfectly legitimate way to look at the world and try to explain it. But they are obviously only one, and not the one that allow the most accurate and reliable predictions, in my view. I think as anthropologists we are not so much puzzled by the fact that economists develop models rather than doing fieldwork to observe what people actually do and how they interrelate (although there are, of course, economists who do just that). What keeps us perplexed is that economists seem *not* to think that it might be fruitful to also incorporate studies which investigate ‘identities in the wild’ into their model-building. Admittedly, me and others are guilty of the same. I do not recall having cited a single economist when writing about identity theories—although I do when talking about markets, transaction costs, or social cooperation—rather, I stick with fellow anthropologists, a few sociologists, and eventually one or two psychologists.

III. GROUP BOUNDARIES

Some analytical flaws that I identify in the paper’s argument may derive out of that. One is the taken-for-granted nature of ‘ethnic groups and boundaries’ (Barth 1969). As has been firmly established since the days the so-titled volume was released, this is not an easy relationship. Ethnic groups may lead long lives in spite of a constant flow of people across their boundaries. And none of these necessarily are given by birth or are primordial, as indicated by the authors of “Narrow Identities” (DG 2019, 395). As all social entities, identity-bearing groups or categories are necessarily culturally constructed—where else would they come from—and manipulable, although once there, they unfold an often strong force to

bind people together or put them at each other's throat under given circumstances. But they are equally never unquestioned and are exposed to permanent changes in membership, shape, and boundary delimitation. And membership in them is hardly inherited in the strict sense but has to be actively promoted by way of socialisation for each new generation. It is only such social and political pressure that makes them appear natural in the eyes of the people concerned—a phenomenon that has also been called “*emic*” primordialism (Gil-White 1999, 792). Some authors have even argued that the division of humankind into neatly-bounded ethnic groups is an invention of colonialism, at least for the African case, and that traditional societies were based more on perceptions of gradual change. Social identities were then located along a continuum. Villagers *A* see themselves as closest to villagers *B*, who in turn think of being equally similar to *C*, while the latter drop *A* as compatriots in favour for villagers *D*, and so on (Elwert 1989).

Even then, not all ethnic or national identity categories strictly ask for single membership. I have made the case of the Uzbeks in Central Asia who allow local Tajiks, Arabs, or Turkmens to switch affiliation but also to add their new affiliation to their original belonging, and to continue to speak a different language, as long as the overall repertoire of cultural expressions and associated political loyalty is adhered to. People may then be Tajiks and Uzbeks, the latter both as an overarching national category and as a local ethnic belonging (Finke and Sancak 2012; Finke 2014). Leach (1954) has described similar patterns for the Kachin in Myanmar and Astuti (1995) for the Vezo in Madagascar. Such perceptions also allow identities to change and not necessarily consider them to be acquired by birth. Other constellations do not allow such fluidity, as is the case for most societies based on unilineal descent models, which define the belonging of individuals as being determined by—socially defined—genealogies (Gil-White 1999; Finke and Sökefeld 2018).

IV. POWER, FREE-RIDING, AND SWITCHING RULES

Another flaw, in my understanding, is the homogeneous nature that these groups and categories seem to have for the authors. As I mentioned, rightly, the authors see identity as a way to find one's place in the world, but these places and positions are not equal. And neither are the chances to decide on one's identity or belonging. The terms power and inequality are conspicuously missing in the text, as if identifying oneself and others takes place in a simple equilibrium game. But in the real world, people

have very different stakes and different means at hand to pursue their interests. Manipulating or limiting the options for ethnic belongings for others are as part of this as is inciting someone to act in certain ways to one's own benefit. The history of humankind as one of conflicts and wars, both between groups as well as within, is a telling story of that. People may have little choice of which group they belong to, as exemplified by the (in)famous 'one-drop rule', which defined racial categorisations in North America.

Ethnic or other social groups often do not fulfil the expected outcomes for yet another reason. Their main function may be the provision of public goods but as in all spheres of life, there is always an opportunity and temptation for free riding. This may spoil all the efforts one invests into joining or leaving groups as well as the overall aims—if there is such a thing—of the group in question. Individuals, as every economist is well aware, differ in their preferences, and this also applies to the meaning that identities and attachments to others have for them. Such a view is by no means trivial or of marginal relevance for the sustainability of larger categories of humans but may well jeopardise any benefits sought with such affiliations. It has been argued, and the authors take a similar point towards the end of the paper, that most ethnic groups remain at the level of a category and hardly ever lead to much collective action so that conflicts between them usually have to be initiated by entrepreneurial actors seeking an individual benefit in this (Eriksen 1993). And the latter is certainly not per se “an increasing function of the aggregate group payoff” (DG 2019, 410). Elites may have an interest in increasing collective benefits for the greater amount of revenues this implies. But this can hardly be taken for granted. One may look at all the Napoleons world history has experienced. Or what to make out of group norms such as bride-wealth and polygyny that have been described as means of elites—elderly men in this case—to monopolise power, livestock, and human resources. Those disadvantaged by the types of institutions prevalent in a given society not only lack bargaining power to change this but also face social pressure that disenables them to vote for membership in a different group (Ensminger 1992).

V. THE COGNITION OF IDENTITY

Finally, and somewhat in contrast to the last point, identities—narrow or broad, exclusive or inclusive—also develop a life of their own, and people begin to strongly believe in their content. Thus, a change of affiliation is

not always easy and transcends simple calculations of costs and benefits. Being English, Swiss, Peruvian, or Mongol includes not only a feeling of cosiness to one's place and people but is also attached with ideas of right and wrong, of how things should be done properly, and what constitutes punishable behaviour. The world of social norms and ideologies comes in here and at times may unfold a strong impact on people's feelings and loyalties, making the latter difficult to switch (Finke and Sökefeld 2018). Having grown up with a particular image of oneself and one's origins, it may be almost unthinkable to change this by, for example, turning from being German into being French. But, of course, historically this is exactly what happened in places like Alsace and many other parts of the world again and again (and back and forth).

Summing up: While identities of any kind entail certain cultural models or schemes that influence people's worldviews, they also constitute—much in line with the arguments of the paper to be now discussed—institutional guidelines for concrete behaviour and interactional patterns. As I have argued elsewhere (Finke 2014), belonging to specific groups or categories may reduce transaction costs of various kinds, economic as well as social ones, and enable the provision of public goods. A logical consequence of this is that individuals may want to adapt their affiliations to those categories better suited to reducing transaction costs. But there may be significant exit and entry costs—including social ostracising and moral discomfort on the part of the individual concerned—to prevent such a shift. This may partly explain the longevity of ethnic groups that seem to be on the losing side of history.

Whether there is a strict positive correlation with the size of the groups in question is a different matter and one which I am more sceptical about. I also hesitate to define group affiliations and individual identifications as necessarily positive, providing communities with public goods. Sometimes the benefits narrowly stay with individuals. And a quick look at the world of today paints a picture of meanness and horrors conducted in the name of ethnic groups or other bearers of 'narrow identities', not only to outsiders but also to everyone perceived as a potential traitor within (those who do not support the collective aim, as defined by its leaders). This is certainly not only an issue of weak governance, as the authors claim. To a certain degree, it is rather exactly the consequence of the human trend to go for mutually exclusive categories when defining themselves. I am not sure if the models the authors develop and weigh

against each other will provide us with an answer to that. But it is certainly worth a try and welcomed food for thought.

REFERENCES

- Astuti, Rita. 1995. *People of the Sea: Identity and Descent among the Vezo of Madagascar*. Cambridge: Cambridge University Press.
- Barth, Fredrik. 1969. "Introduction." In *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*, edited by Fredrik Barth, 9–38. Bergen: Universitetsforlaget.
- Dasgupta, Partha, and Sanjeev Goyal. 2019. "Narrow Identities." *Journal of Institutional and Theoretical Economics* 175 (3): 395–419.
- Elwert, Georg. 1989. "Nationalismus und Ethnizität: Über die Bildung von Wir-Gruppen." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 41 (3): 440–464.
- Ensminger, Jean. 1992. *Making a Market: The Institutional Transformation of an African Society*. New York, NY: Cambridge University Press.
- Eriksen, Thomas H. 1993. *Ethnicity and Nationalism: Anthropological Perspectives*. London: Pluto Press.
- Finke, Peter. 2014. *Variations on Uzbek Identity: Strategic Choices, Cognitive Schemas and Political Constraints in Identification Processes*. Oxford: Berghahn Books.
- Finke, Peter, and Meltem Sancak. 2012. "To Be an Uzbek or Not to Be a Tajik? Ethnicity and Locality in the Bukhara Oasis." *Zeitschrift für Ethnologie* 137 (1): 47–70.
- Finke, Peter, and Martin Sökefeld. 2018. "Identity in Anthropology." In *The International Encyclopedia of Anthropology*, edited by Hilary Callan. Hoboken, NJ: John Wiley & Sons. First published September 5, 2018. <http://doi.org/10.1002/9781118924396.wbiea2142>.
- Gil-White, Francisco J. 1999. "How Thick Is Blood? The Plot Thickens...: If Ethnic Actors Are Primordialists, What Remains of the Circumstantialist/Primordialist Controversy?" *Ethnic and Racial Studies* 22 (5): 789–820.
- Leach, Edmund R. 1954. *Political Systems of Highland Burma: A Study of Kachin Social Structure*. Cambridge, MA: Harvard University Press.
- Schlee, Günther. 2004. "Taking Sides and Constructing Identities: Reflections on Conflict Theory." *Journal of the Royal Anthropological Institute* 10 (1): 135–156.

Peter Finke is Professor for Social Anthropology at the University of Zurich. An expert in economic anthropology, he has conducted field research in Mongolia, Qazaqstan, and Uzbekistan on issues of economic transformation, institutional change, and social identity since the early 1990s. Contact e-mail: <peter.finke@uzh.ch>

Group Membership or Identity?

MIRIAM TESCHL

EHESS, Aix-Marseille Université, CNRS, and AMSE

Liberal cosmopolitans, Dasgupta and Goyal say, argue that identity should primarily be recognised as “unfettered choices by individuals regarding where they belong” (2019, 396). If identity thus consists of some inherited dimensions, complemented and/or completed by deliberative choice, then identity is fundamentally multiple. However, Dasgupta and Goyal observe that “all over the world we see people defining themselves in narrow, exclusive terms and being so regarded by others” (396). The key point in their paper is to give an economic explanation of the emergence of narrow identities and, by doing so, to show that narrow identities may be in certain cases individually and socially desired. This, of course, is quite a surprising result—one that stands in opposition to the liberal claim and belief that choosing and recognising multiple identities is necessarily always better. How do Dasgupta and Goyal (2019; henceforth DG) achieve this result?

I. DASGUPTA AND GOYAL’S MODEL

A key aspect of their model is the idea that groups impose positive and negative externalities on other groups. A group’s payoff always increases in own-group size, and this increase is, by assumption, equally distributed among the members of the group.¹ Hence, joining a group is always advantageous, and this might lead us to believe that joining all groups may be best for all. This is indeed the case as long as there are no (or at least only positive) spill-overs between groups, and as long as joining any group comes at no (or a sufficiently small) cost. The situation becomes more complicated, as expected, with negative spill-overs. As is common with externalities, they are not considered by individual agents who are concerned for their own benefit, and who do not notice the costs their

¹ Later in the text, DG (2019) also discuss the case where payoff division is such that group size is a public good to which all individuals have equal access.

actions impose on other agents. If an agent belongs to a group and then joins another group, this may lead to costs for the original group due to negative spill-overs. That is, payoffs in the first group are then decreasing with the membership size of the second group. This is why, in some cases, it would be socially more efficient to have a single group. At the same time, however, an individual has an incentive to deviate from single group membership as the gain from an additional group membership accrues fully to the individual, while the costs of that additional group membership are shared among all members of the first group. Clearly, if the negative spill-overs are high, relative to the additional gain from an extra group membership, single membership will automatically emerge. In any other case, multiple memberships emerge—these, however, are not socially efficient due to costly negative externalities. Imposing a rule that requires exclusive memberships may remedy this inefficiency.

Hence, given this model, we can distil three key messages. (1) Contrary to claims by liberal cosmopolitans, there are cases in which ‘unfettered choices by individuals regarding where they belong’ lead to single membership or, in DG’s words, ‘narrow identity’, and not to multiple identities. (2) A narrow identity may be individually and socially more efficient than multiple identities. (3) ‘Fettered choices’—choices constrained by rules imposing exclusive group memberships—may lead to a socially and individually efficient outcome.

As is often the case with economic models, we judge them to be good models when they produce novel and/or counterintuitive results. Given the tendency, I may say, among liberal academics to adhere to the claims of liberal cosmopolitanism (possibly because we often have multiple identities ourselves—coming from one country but working in another, for example), DG’s results are indeed counterintuitive and quite surprising. Do these results give us enough reason to doubt and revise our own liberal intuitions, which are to acknowledge our multiple identities?

In what follows, I will argue for the idea that we should *not* revise our liberal intuitions. First, I will show that there is a difference between reasons for group or identity choice and reasoning about those reasons, which is a liberal and, so far, primarily *normative* concern. I will then question whether playing a strategic game, as DG propose, responds to such liberal concern. I will next distinguish between different contexts in which group or identity choices can take place and argue that had such contexts been more explicit in DG’s model, then DG’s results on narrowness may have been less convincing. That is, we would have arguably

found a narrow identity only in very extreme cases. However, I will also argue that DG's definition of a narrow identity may not fully coincide with what we commonly understand under 'narrow identity'. Finally, I will question the idea that identity can be represented by one's group choice. While this is a convenient representation of identity for the purpose of a model, it is doubtful that DG's claim—that "the very richness of the notion [of identity] suggests it may pay to distil it into an almost 'presocietal' form" (2019, 396)—holds. At least, there should have been a more prominent effort to show what exactly we can learn from the results of the model about real-world issues.

II. NORMATIVE VERSUS POSITIVE

DG's explicit aim is to provide an explanation for why "all over the world we see individuals defining themselves in narrow and exclusive terms" (2019, 414). They contrast this explanation with the liberal view that people have multiple identities, and that the "sanctity of narrow social identities by those having them are unwarranted, even delusional" (396). However, the liberal cosmopolitan view that unfettered choices will lead to the recognition of us having multiple identities is a normative 'call', rather than an empirical claim. It is an empirical fact for liberals that we *have* multiple identities as we belong, even without any prior choice, to different 'groups' simply by being born at a particular time and place into a particular family. But it is a normative issue that—in Amartya Sen's words, an author singled out as a 'liberal cosmopolitan' by DG—" [w]e do have the opportunity to determine the weights we have reason to place on our different associations and distinct identities" (Sen 2004, 86). Sen insists that:

The reasoning in the choice of relevant identities must [...] go well beyond the purely intellectual into contingent social significance. Not only is reason involved in the choice of identity, but it may require some collateral social analysis of the grounds of relevance. (Sen 2004, 86)

Hence, unfettered choices do not simply amount to the decision to join one or more groups. Sen's key point on identity is actually twofold. First, there is always an element of choice involved in determining one's identity. We do not simply 'discover' the identity we are born into, and we do not have to uncritically adhere to it; we are always left with some choice. Second, choosing means evaluating *which* group memberships we give

priority and also *why* we do so. This implies that there is always a self-reflective and critical element in determining 'who we are'. That is, even if we are born into particular groups and have not, strictly speaking, chosen to belong to them, it is our prerogative to evaluate how much importance we give to those group memberships. If we engage in such evaluation—which, according to Sen, we *should*—then we unmistakably recognise the empirical fact that we always have multiple attachments and affiliations—that is, multiple identities—to which we give different priorities.

III. REASONS FOR CHOICE

Sen has been quite silent on what, exactly, is involved in reasoning about the weighting of identities and how this reasoning should work. The question we may thus ask is: what does this *reasoning* process involve? Does it mean playing some strategic game, as in DG's model? If yes, then reasoning may lead to the recognition of some narrow identity only. In that case, liberal cosmopolitans such as Sen would be mistaken about their view that people would necessarily recognise multiple identities.

As is typical with economic choices, agents understand payoffs in terms of costs and benefits, and the highest payoff is usually the most attractive one. What we know about the payoffs in the given model is that they depend primarily on the interplay between one's own and another group's size. Group size is thus the primary operative *reason* in DG's model and it is indeed a plausible reason in certain contexts. For example, a bigger fan club might be better at promoting and financing a club, a greater Amnesty International or Greenpeace presence might give more force to their respective initiatives, a higher number of Fridays-for-Future demonstrators might amplify their message for action. DG mention the promotion of fundamental research or the chance of obtaining public funds, which might improve the higher the number of group members is. In their model, payoffs literally increase with own-group size, and either increase or decrease—depending on the valence of the respective spill-over—with the other group's size.

This means that, in this model, no difference is made between the benefits of size and the benefits of consequences of such size, although there may be no necessary link between the two. Certain outcomes may also be achieved by a small group of, for example, well-connected and influential people, especially in our digital world. Hence, another reason for joining a group—independently of group size—may be the

competence of the group's members, or their connections with certain decision makers. Obviously, we may be able to think of many more reasons for why individuals join groups, and such reasons are arguably context-dependent and influenced by different 'grounds of relevance'. DG's model, however, does not capture these different reasons, nor does it present a more detailed analysis of the context in which their chosen reason for group choice—namely, group size—applies. It is a specific model focused on a particular reason for choice, but it is presented as a general model of group choice. It does not constitute a general model of the processes of *reasoning* and *scrutiny* about that particular reason, which is the concern of liberal cosmopolitans. In that sense, playing a strategic game does not capture *reasoning in the choice of relevant identities*. It is thus somehow curious that DG hold the specific results of their single-reason-based model against the general, normative call of liberal cosmopolitans for reasoning and scrutiny about reasons, as these two endeavours are quite different.

IV. CONTEXT OF CHOICE

In DG's model, people can choose to participate in either one or two groups. But what if we assume that people make group choices across different contexts? We could define a context as the different sets of groups from which one can choose. These sets of groups can be constituted on the basis of different interests, such as professional interests, hobbies, or family activities. Arguably, given a choice among two groups, there may be a context, where I may choose to belong to one group, and another, where I may choose to belong to both groups. For example, if I'm a professional football player, and thereby an important member for my club, but I would also like to play tennis, then, of course, my starting to play tennis imposes costs on my football team. Such costs include having less time to train for my football games or not getting enough sleep to stay in top form for both training programs. As long as I like and enjoy playing both sports, I may indeed have multiple memberships, even if it would be better for my football mates that I dedicated my effort exclusively to football. If I (and my teammates) start suffering from the increase in lost games, I may reconsider and stop playing tennis. Of course, I may have also signed a contract with my football team which does not allow me to play any sport other than football; if this arrangement were good for me—and my team, as well as our fan club—it would be socially efficient. This means that, *in the context of my professional opportunities*,

having a narrow identity and focussing on one particular professional activity makes absolute sense. However, when choosing among leisure activities, I may well choose to belong to a reading club *and* a chess club (and forego playing tennis, even if this choice were also available to me). Thus, by repeating the choice process, as presented by DG, across different contexts with different groups, I may end up with a set of different identities, that is, multiple identities after all. But this is exactly what liberal cosmopolitans claim: choosing one's identity will lead to the recognition of multiple affiliations and attachments. It would be a very special case, and an extreme life-style choice indeed, if someone *continuously* chose one and the same identity across different contexts—it would be interesting to study the conditions under which this kind of 'narrow and exclusive' identity emerges.

V. WHAT IS A NARROW IDENTITY?

Notice that, by definition, a narrow identity in DG's sense is a situation in which everyone joins one group, whereas, in the case of multiple identities, some people may join more than one group (2019, 402). Everyone joining *one and the same* group thus counts as a narrow identity. One may think, however, that this is not quite the definition of narrow identity that liberal cosmopolitans have in mind. If everyone joins one and the same group, and this is socially efficient, then surely liberal cosmopolitans would also support such narrowness. The more compelling case of narrowness in DG's model is, of course, when people choose to belong to different groups, which then implies that agents lack any community. In DG's model, individuals can also choose to belong to any subset of groups, which is a strong assumption that rules out a number of situations where the choice set is restricted, at least for certain people. Indeed, one may think that it is these restrictions that may motivate narrowness because they create or reinforce differences between 'us' (insiders) and 'them' (outsiders).

In summary, it is not quite clear what DG are quarrelling with when referring to 'liberal cosmopolitans'. Of course, their results on the efficiency of narrow identities seem to contradict liberal claims, according to which multiple identities emerge from reasoning about identity. But, at a closer look, there may be no opposition: first, because the meaning of narrow identity (in DG's model and in liberal claims) does not seem to be exactly the same, and second, because by repeating the choice process

across different contexts, it is reasonable to think that multiple affiliations will emerge after all.

VI. IS IDENTITY GROUP MEMBERSHIP?

This leads to a further point. Identity takes part in DG's analysis only as a definition, but it is not explained in the formal choice model as such. We talk about identity here because we *interpret* group choice as identity choice. That is, choosing a particular group makes me that person the group is characteristic of. Joining a football club makes me a footballer. Joining a reading club makes me a reader and book lover. This is a rather simple view of identity, which may apply in some cases (for example, when I am indeed a professional football player, and I like the idea of being seen as a footballer and am presenting myself as such); but it does not in others (for example, I am a member of the local school's parents association, but I don't take that membership to be part of my identity; it's an interest, not something that defines me and that I care to mention when I present myself).

Following Kwame Appiah (2006a, 2006b), DG explicitly use a "stripped-down formulation of social identities" (2019, 398). In brief, Appiah proposes to label groups as a way of distinguishing people who have a group's label from those who don't. This is a helpful way to 'exploit' identity (thus defined) in a formal model and a common way to *describe* people. But is this identity as we understand it when we talk about ourselves? Do we care about identity defined in this way? Would we present ourselves by listing our group memberships? It's a convenient way to describe 'what' the person is, socially speaking, but such a description does not necessarily tell us 'who' the person is—this involves, as liberal cosmopolitans think, a reflection on one's own involvement in groups (Kirman and Teschl 2004).

Hence, if we strip the model down to what it does—namely, explain when it is better to have single membership versus dual or multiple memberships—and leave out any wordy interpretations of narrow or multiple identities, then the paper's claim that it explains why people seek to define "themselves in narrow and exclusive terms", as DG (2019, 414) put it, does not hold. Rather, the model explains why people play only football and not tennis, and when they may do both. It does not explain, for example, the Israeli-Palestinian conflict, the situation of the Uighurs or Rohingyas, or even the French yellow vest movement, all of which can be seen as social identity conflicts.

The question, therefore, is: what do we really learn by abstracting and simplifying certain realities to a point where the situation becomes mathematically tractable, but the nature of the question (for example, ‘why do people define themselves in narrow terms?’) is reduced to something different (in the given case, to ‘why do people join only one group?’)? Clearly, the model appears refreshingly counterintuitive when framed as a model of identity choice, and it provides interesting food for thought. Yet, I cannot help but wonder two things. First, abstracting is the very essence of mathematical modelling, no quarrels with that. The question is how to interpret the results and, in the given case, whether they are not ‘hyper-interpreted’—that is, whether it is not the case that more is read into them than what is really there. Put yet differently, how do we return from an explicitly “minimalist” (2019, 397) and “lean notion of identity” (396) to real-world identities? Second, if DG think that the model provides any particular tools or insights for analysing the world in a novel and interesting way, it would have been good to make those explicit. So far, we are left with excessively simple examples, such as joining religious congregations and research institutions, which do not really improve our understanding of real-world narrow identities.

Economic decisions are a driving factor of a number of choices; many of them also happen outside any typical market setting. It is always an interesting exercise to look for economic reasoning in non-market decisions. This does not (necessarily) amount to economic imperialism but is constitutive of a real intellectual endeavour, as DG clearly show. But, from this model, we learn when single group memberships are efficient, not why people seek to define themselves through narrow identities. Clearly, the latter topic attracts more attention than the explanation of single group memberships. But that topic is also a substantially more complex issue. Economists are used to simplifying the world. There is nothing wrong in admitting that one is simplifying complexity and in dealing with it in an economic way; but then, one also needs to clearly single out the benefits and limitations of such an economic analysis.

REFERENCES

- Appiah, Kwame Anthony. 2006a. *Cosmopolitanism: Ethics in a World of Strangers*. New York, NY: W. W. Norton.
- Appiah, Kwame Anthony. 2006b. “The Politics of Identity.” *Dædalus* 135 (4): 15–22.
- Dasgupta, Partha, and Sanjeev Goyal. 2019. “Narrow Identities.” *Journal of Institutional and Theoretical Economics* 175 (3): 395–419.

Kirman, Alan, and Miriam Teschl. 2004. "On the Emergence of Economic Identity." *Revue de Philosophie Économique* 9: 59–86.

Sen, Amartya. 2004. "Social Identity." *Revue de Philosophie Économique* 9: 81–101.

Miriam Teschl is associate professor in economic philosophy at the École des Hautes Études en Sciences Sociales (EHESS) and is based at the Aix-Marseille School of Economics (AMSE). She is interested in interdisciplinary questions of wellbeing, social justice, and, in particular, decision-making under internal conflict.

Contact e-mail: <miriam.teschl@ehess.fr>

Narrow Identities Revisited

PARTHA DASGUPTA

University of Cambridge and St John's College, Cambridge

SANJEEV GOYAL

University of Cambridge and Christ's College, Cambridge

We are most grateful to the editors for inviting a discussion round our paper, Dasgupta and Goyal (2019; henceforth DG), and to Professors Carvalho, Davis, Finke, and Teschl for their most enquiring and incisive comments. They have made us look afresh at how the approach we took in our work on social identity fits into what is a substantial literature on the subject.

I. MODELLING SOCIAL IDENTITY

In DG, we constructed a *minimalist* model to explore the concept of social identity, because one's 'identity'—be it personal or social—is overly close to home; so close, that we are all tempted to make of it what we want to see in ourselves. Too much structure runs the risk of tilting the analysis toward our predispositions.

There is an obvious sense in which we all have *multiple* identities, spanning across our professional, social, and personal lives. That unquestionable fact has led influential commentators to view *narrow* social identities as a blight, founded on delusions (Sen 2006). Carvalho, in his comments on our paper, goes further and claims that “narrow identification is the exception, not the rule. It is not a natural state from which we must plot an escape route, but rather an extreme and unnatural one from which we must wonder how we got there” (79).

The thought is beguiling, but the evidence points in a different direction from what Carvalho thinks it does. The contours of our emotions were etched in the palaeolithic era. In times long ago, social groups were tiny, perhaps twenty members in a band and twenty-five bands in a tribe. We humans evolved as small group animals. If multiple identities were the norm, people would have identified themselves with

members of all tribes they ever encountered. Evidence from archaeological and genetic records of intertribal warfare over scarce resources tell us that they didn't (see, for example, Henrich 2015, for an account of the violence that frequently characterised encounters between tribes). That is why, in their study of the diversity of human natures, evolutionary biologists have examined why we are *even today* disposed toward narrow identities, why when called upon, we take sides. They have explained why our social horizons are often restricted even when we have recognized that the benefits of collective enterprises could be most effectively realized if we admitted others into our group (Ehrlich 2000). In our paper, we constructed a model of personal incentives and group interests to offer an explanation that complements the one provided by evolutionary biologists.

The dilemma people face in having to choose sides is exemplified in an early book in the *Iliad*, where Agamemnon raises the question before the Greek armies whether they should sail home. They had been there for nine years, and Troy still stood intact. Of course, he raises the question knowing what the answer would be, and it is Odysseus who supplies it: the Greeks and their allies had made a commitment, and to leave then would be to break it. Homer tells us that the Greeks saw themselves as Greeks (the poet says they called themselves Achaeans) and were joined by, among other bonds, marriage alliances. But they and their allies would appear to have been a loose coalition, otherwise it would be hard to understand Achilles' wrath in the opening book, and his bitter complaint that he had brought his army to fight a people who had done him no wrong. Achilles' re-entry into the coalition had to do with seeking vengeance for Patroclus' death at the hands of Hector, a very personal loss. The tragedy that followed was inevitable once the Achaeans stood firm on the commitment made nine years previously. The analysis in our paper was not constructed to identify contingencies where people's loyalties would be tested, it was designed to uncover reasons people form loyalties in the first place. Contingencies oblige us to take sides, and even when the occasions are resolved peacefully, we know which side we would be on when the next push comes to shove. Carvalho's moving description of his family's experience speaks to that.

An insistence that we all have multiple identities can be read as no more than that we are all humans. The point is of course that peaceful co-existence among groups doesn't mean people parade multiple identities, it only points to accommodation having been reached.

Showing tolerance toward others' beliefs is different from showing respect for them, for it may be that you judge their beliefs to be false and their customs so at odds with your own that you don't want to have anything to do with them socially, least of all engage in interminable 'conversations' with them. A recent publication from the Pew Research Center (Sahgal et al. 2021), based on interviews in seventeen languages with 30,000 adults from the various faiths in India, has reported that an overwhelming majority declared themselves to be deeply religious, but regardless of their religion, expressed their allegiance to religious tolerance and peaceful co-existence even while insisting on religious exclusivity and social segregation. Hindus especially declared their national identity, religion, and language to be closely connected. The model in our paper studied how exclusivity can arise even when we are not born into it.

Teschl opens her remarks by noting that "there is a difference between reasons for group or identity choice and reasoning about those reasons, which is a liberal and, so far, primarily *normative* concern" (107). She is right in drawing the distinction, but in our paper, we were not *advocating* narrow identities; rather, we were locating *reasons* for the prevalence of narrow identities. The 'is-ought' distinction is all important here. Teschl also suggests that the liberal cosmopolitan perspective is built on the fact that we have multiple attachments and affiliations, to which we give different priorities, and she invokes Sen (2006) in support of her view. But the fact that the multiple attachments and affiliations we have may be different from the ones we were born with, on its own proves nothing. People reasonably want to recognise the benefits and costs involved—including psychological transaction costs—before attaching priorities to them. Our model was designed to show that. We wanted to examine circumstances under which adopting a narrow identity may be in the interest of those involved. To do that, we constructed a model that showed that individuals and groups may find it in their interest to adopt a narrow notion of identity *even* when they have multiple attachments and are clear sighted.

Liberal cosmopolitanism is a natural position to adopt when you have the facilities to travel back and forth, and those around you are like minded (for example, in holding liberal cosmopolitan thoughts) no matter where you happen to be; it is altogether harder when you don't have that luxury. Perhaps the warm glow that accompanies liberal cosmopolitanism is inevitable when its advocacy is read by like-minded

liberal cosmopolitans. Explanatory models, such as ours in DG, force us to imagine what it could be like to be others far removed in social space from us; and, given the constraints they face, what their concerns could be. Rootedness cannot be airily dismissed, for it serves as an anchor. That alone tells us that liberal cosmopolitanism isn't the only emergent world view, and why it is a good idea to *not* introduce morality in every social science exercise. Social thinkers are inevitably among society's intellectual elite, and that may be a reason they so often slide from analysis to advocacy. In our paper, we consciously avoided doing that.

II. CONCEPTUAL DISTINCTIONS

Davis draws attention to a literature in social psychology that distinguishes notions of 'role-based' or 'relational' social identity from 'category-based' or 'categorical' social identity (88). He suggests that in that literature category-based identity affords little room for individual autonomy or agency, whereas role-based identity comes with expectations and norms of behaviour and allows for agency. Because he thinks our model does not consider 'relational social identities', he provides as an example of such an identity someone who identifies himself *as* an employer. In contrast, someone who identifies herself *with* others would display a 'categorical social identity'. In our model, the former is a sense of the self, influencing, for example, the benefits the employer would enjoy from joining an employers' association. Joining such an association would in our model be seen as assuming a particular social identity. Of course, becoming an employer is itself an outcome of past choices and circumstances, so even the sense of the self has a history (see section III below). For an employer, the cost of joining, say, a teachers' association would typically be huge, and usually that option would not even be open to the person. Ultimately, it matters little that we did not include relational social identities explicitly in the category of social identities, because relational and categorical social identities enter the model through different routes. In any case, our intention was not to provide a classification of the various ways social identity has been identified, it was a lot more modest. We wanted to show that individuals and groups may find it in their interest to *adopt* a narrow notion of identity.

The distinctions between various notions of identity and their relationship with individual agency is mostly a matter of terminology. Akerlof and Kranton (2000), for example, proposed a notion of identity

in which individuals belong to categories which come with associated norms of behaviour. As Davis remarks, certain notions of identity come with limitations on individual agency, and category-based identity would be an instance of that (91). That is one reason we constructed a model that uncovers the incentives groups have in restricting autonomy and insisting on exclusive identity. Individuals in our model, however, do have choice over whether to assume the exclusive identity demanded of them. In equilibrium, they choose to do so.

Finke asks why 'narrow' should be contrasted with 'multiple' and suggests that in our model 'narrow' might be better contrasted with 'broad'. Similarly, he wonders if it would have been better had we used the terms 'inclusive' and 'exclusive', or alternatively 'single' and 'multiple' (100). While all this is a matter of terminology, Finke is right to suggest alternative nomenclatures. Indeed, in section 6 of our paper, we present a variation on our model in which we discuss the notions of inclusive and exclusive identity.

Finke rightly says that size is not necessarily an advantage, that people are known to identify with small, exclusive groups (100). Our model was designed to specify a set of conditions that are *sufficient* for the emergence of narrow identities. To explore conditions that are *necessary* would have taken us in a different direction.

III. TIMELESS VS DYNAMIC MODELS

Finke believes our model has analytical flaws. He says it regards group boundaries to be immutable, whereas in practice they are fluid and evolve over time. He cites the examples of the Uzbeks, the Kachin in Myanmar, and the Vezo in Madagascar as illustrations of groups that extended their scope by including other ethnicities (102). By contrast, he says there are groups that retain a strict and narrow definition of membership.

Our model is timeless. But even within that restriction, it is possible to accommodate the inclusion of new ethnic groups. It could be that circumstances change, so that it benefits existing members of an ethnic group to accept another group to join them. If it is in the interest of the other group to join, the combined group would be larger. In any event, years of theoretical work on evolutionary biology (see, for example, Nowak 2006) have uncovered that timeless models, beloved of economists, contain in them signatures of past experiences in a society. Parameters of a timeless model hide in them the history of past choices

and events. Irreversibility of certain classes of decisions leave an imprint because they shape the trajectory of a society's past. What is endogenous in a model with history is reduced to exogenous parameters in its timeless variant. The parameters could, for example, represent heterogeneity of people as regard their abilities, tastes, and histories. Thus, Davis argues that heterogeneity and the power relations surrounding it are central to understanding the emergence of identity in a society (95, 97). We agree, but our timeless model, as with all timeless models, has the past frozen in it. It would have served no purpose in our exercise to postulate heterogeneity among members of society. We would still have reached our conclusions: *that the advantages size may confer on groups and the negative externalities that may prevail across groups, taken together, are a reason people could choose to assume narrow identities*, or, in our model, join one side rather than another. Equilibrium in models of the sort we constructed is not necessarily unique, but we know from models of evolutionary dynamics that history matters, that even chance events in the past could have pointed toward the equilibrium that the process eventually selected. One can, of course, unpick the parameters in a timeless model and ask how they have come to be what they are in a society, such as the personal cost of harbouring multiple identities. But that would be to elaborate, not to add.

Carvalho's formulation of distance (see equation (1) in his comment, 81) and his definition of narrow and broad identity is useful. The issue of dormant allegiances is natural in our context. There is also the issue of contingencies that interact with these dormant identities. The two often go together and will be important in practice. A possibility is to study steady states or stable configurations in a dynamic setting to illuminate these relations. The stability notions could be defined in such a way as to incorporate the 'within' and 'across' group mechanisms he elaborates (82-84). Such a model would help us understand the vector of weights on identity dimensions that are stable, as a function of different parameters of initial 'dormant' identity configurations and the character of social interactions, which would be endogenous.

Davis is entirely right that we left no space to discuss the connections between 'personal' and 'social' identities, and suggests the connections are hard to conceptualise (91). It is a gift of the concept of equilibrium that it severs the distinction between cause and effect. Our sense of the self plays a role in our ability and the choices we make regarding which social group(s) to join, but social interactions with

others in the group(s) we join in turn influence our sense of the self. In a model that works through time, such mutual influences would be sequential, subject of course to uncertain contingencies; but in equilibrium, should the social system converge to an equilibrium, the mutual influences would be in balance. They would be solutions of a system of equations representing (stochastic) equilibrium conditions.

In a complex system, and a human society may be so regarded, it may be that there is no equilibrium to which it converges. It may even be that the system is chaotic, but if that were so, little could be said regarding the evolution of personal and social identities. Historical evidence, however, does not point to chaotic time series in such significant economic variables as the share of profits in national income, let alone in national income itself. So, our timeless model probably does not mislead. In any case, there are complex systems that display chaotic behaviour among small subsystems but regular behaviour in the aggregate (unpredictability of the weather in contrast to regularity of climate is an example). We are thus at one with Finke, where he writes, “ethnic groups may lead long lives in spite of a constant flow of people across their boundaries” (101).

IV. THEORY AND FIELDWORK

Finke is nearly right when he says that “economists develop models rather than doing fieldwork to observe what people actually do and how they interrelate” (101). But only nearly. Fieldwork among communities, conducted by ecological economists in recent years (see, for example, Ghate, Jodha, and Mukhopadhyay 2008), has started unravelling intricate sets of social norms of behaviour in these communities. Such behavioural norms provide incentives to community members to manage their local resource base sustainably in circumstances where the rule of law, as conventionally understood, is not near at hand. Norms are known to differ across ecological niches, the task that theoretical economists have undertaken has been to interpret those differences in terms of differences in the character of the niches (Dasgupta 2021). Communitarian management of coastal fisheries in the tropics requires a different set of rules from what they would be in temperate zones, and both would be expected to differ from those in place in grazing lands or tropical rainforests. More generally, economic theorists use data collected by scholars conducting fieldwork to make sense of them, at least when seen through the lens of economics.

REFERENCES

- Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *The Quarterly Journal of Economics* 115 (3): 715–753.
- Dasgupta, Partha. 2021. *The Economics of Biodiversity: The Dasgupta Review*. London: HM Treasury.
- Dasgupta, Partha, and Sanjeev Goyal. 2019. "Narrow Identities." *Journal of Institutional and Theoretical Economics* 175 (3): 395–419.
- Ehrlich, Paul R. 2000. *Human Natures: Genes, Cultures, and the Human Prospect*. Washington, DC: Island Press.
- Ghate, Rucha, Narpal Jodha, and Pranab Mukhopadhyay. 2008. *Promise, Trust, and Evolution: Managing the Commons of South Asia*. Oxford: Oxford University Press.
- Henrich, Joseph P. 2015. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.
- Nowak, Martin A. 2006. *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Sahgal, Neha, Jonathan Evans, Ariana Monique Salazar, Kelsey Jo Starr, and Manolo Corichi. 2021. "Religion in India: Tolerance and Segregation." Pew Research Center, June 29, 2021. <https://www.pewforum.org/2021/06/29/religion-in-india-tolerance-and-segregation/>.
- Sen, Amartya K. 2006. *Identity and Violence: The Illusion of Destiny*. New York, NY: W. W. Norton.

Partha Dasgupta, FBA, FRS, is the Frank Ramsey Professor Emeritus of Economics, University of Cambridge and Fellow of St John's College, Cambridge. His publications include *Human Well-Being and the Natural Environment* (2nd ed., 2004), *Economics: A Very Short Introduction* (2007), and his Kenneth Arrow Lectures, *Time and the Generations: Population Ethics for a Diminishing Planet* (2019).
Contact e-mail: <pd10000@cam.ac.uk>

Sanjeev Goyal, FBA, is Professor of Economics, University of Cambridge and Fellow of Christ's College, Cambridge. He published *Connections: An Introduction to the Economics of Networks* in 2007; a new book, *Principles of Networks*, is to be published in 2022.
Contact e-mail: <sg472@cam.ac.uk>

Grounding Equal Freedom: An Interview with Ian Carter

IAN CARTER (Emsworth UK, 1964) is professor of Political Philosophy at Pavia University in Italy. He has spent most of his career in Pavia, interrupted by brief periods visiting Oxford and Cambridge in the UK, his country of origin. He studied at the University of Newcastle (BA), the University of Manchester (MA), and the European University Institute in Florence (PhD), and was then a lecturer at Manchester University before moving to Pavia in 1993.

Carter's philosophical work has focused primarily on the analysis of fundamental concepts in normative political theory. His ground-breaking monograph *A Measure of Freedom* (1999) was the first book-length treatment of the problems that arise if one assumes political and social freedom to be a matter of degree and therefore to be measurable in an overall sense. It also examined the place of the measurement of freedom in a broader theory of justice. Carter has since then continued to contribute to discussions on freedom, including as a prominent liberal critic of the 'republican' conception of freedom, defending the more basic normative role of the liberal 'negative' conception. More recently, he has played an influential part in debates about the foundations of egalitarianism, developing an account of basic equality grounded in the idea of 'opacity respect'. He is currently working on a monograph that further develops the idea of opacity respect and its implications for egalitarian justice.

The *Erasmus Journal for Philosophy and Economics* (EJPE) interviewed Carter in spring and summer 2021. The interview covers Carter's intellectual biography (section I); his extensive writings on the measurement and value of freedom (section II); his reflections on the use of formal methods in philosophical work on freedom and in political philosophy more broadly (section III); his more recent work on basic equality and respect for persons (section IV); and, finally, his advice to young scholars (section V).

EJPE'S NOTE: The interview was conducted by Annalisa Costella, editor at the *EJPE*. She thanks Constanze Binder and Akshath Jitendranath for advice while preparing the interview.

Footnotes throughout the interview are editorial annotations directing readers to the relevant literature discussed in the text.

I. INTELLECTUAL BIOGRAPHY

EJPE: Professor Ian Carter, reading your CV, one gets the impression that you have always had pretty clear ideas about your intellectual interests. You started out with a Bachelor of Arts in Philosophy and Politics at the University of Newcastle and continued with a Master of Arts in Political Theory at the University of Manchester, suggesting a very firm interest in political philosophy. Is this reading appropriate? Did you know from a relatively early age what you wanted to do research on, as it seems from your CV, or has your intellectual path been less smooth than it appears on paper?

IAN CARTER: I originally enrolled to study politics at Newcastle but soon discovered that my real passion was for philosophy and political ideas rather than, say, party politics or political behaviour. So, I switched to philosophy and politics in my second year and took as many courses in political philosophy as I could. With hindsight, and especially compared to the training received by Italian students of 'political sciences' (*scienze politiche*, in the plural), both my BA and my MA do seem pretty narrow, and I had to fill in a lot of gaps in subsequent years.

I dallied with Hegel in my second year (what fun it is to learn a language so few other people understand!) but then rebelled against his obscurities when I discovered analytical political theory under the tutorship of Peter Jones, and I haven't looked back since. So, yes, in that sense I'd say my intellectual path has been quite smooth.

While your Bachelor and Master 'scream' political philosophy, the department from which you received your PhD, the one of Political and Social Sciences at the European University Institute of Florence (EUI), seems to clash slightly with your previous education as well as with the topic of your PhD thesis. Was there a specific reason that led you to EUI and to the department of Political and Social Sciences?

I can see that might look like an unusual choice. After Newcastle I went to Manchester, mostly because I was keen to study with Hillel Steiner, having read some of his work on freedom and justice which I found fascinating. A natural course would have been to continue studying in Manchester with Hillel, and I was indeed undecided about that. But I also had itchy feet and wanted to experience the culture and language of another country, in particular Italy. I had visited France and Italy during a gap year before university and in the meantime had been teaching myself Italian.

The prospect of three years at the EUI was just too enticing! What's more, the EUI did have a long-standing commitment to political philosophy—perhaps more so than now. Steven Lukes, who became my supervisor there, had been preceded by several other important theorists including Brian Barry and Maurice Cranston, so I wasn't exactly out on a limb, although it's true that most of the people in the same department were doing empirical research. I was also lucky enough to have Philippe Van Parijs as co-supervisor—Philippe visited Florence as a Jean Monnet Fellow at that time. My impression is that the Institute has since become more strictly oriented toward research relating to the European Union than 30 years ago when I joined. True, there was no philosophy department at the EUI, but there was one, of course, at Florence University, and I made the effort to follow some lectures there. And that's how I eventually wound up at Pavia, having made contact in Florence with the Italian philosopher Salvatore Veca, who was teaching in Florence but then moved to Pavia.

So, I guess the allure of Italy interrupted what would have been an even smoother path, but I don't regret the move for one moment. The EUI was a wonderful meeting place for different cultures and intellectual traditions. Although I've stayed on the straight and narrow as an analytical philosopher, and have stayed in close contact with Hillel Steiner and several other academics in Britain and elsewhere, living and working in Italy has exposed me to intellectual traditions I might not have come into contact with had I stayed in Britain. That exposure can stimulate the imagination and promote lateral thinking, something analytic philosophers are always in need of. Salvatore Veca once said to me: 'Remember, philosophy is one third rigour, two thirds imagination!'. He later told me he varies the proportions depending on who he's speaking to.

Would you say that specific events, or people, played a key role in shaping your intellectual identity before you entered university?

I think my father, who died young when I was only 17, had quite a strong influence on my political thinking. He was a lecturer in linguistics and worked on language change between East and West Germany. When I was 10, on a family trip to the Harz mountains, he took me to see the east-west border. We stood in front of a large stretch of grass, a sort of no man's land, beyond which there was a huge fence extending into the distance to the left and right, with towers placed at intervals along the fence. My Dad said: 'If you now cross over this line and start walking across the grass, that soldier up in that tower will shout "Halt!"'. And if you then

ignore him and carry on, he'll shoot you'. That made a big impression on me. Another memory that stands out is his interest and delight in a definition of 'Freiheit' he found in an East German dictionary: 'Freedom (noun): the recognition of necessity'. He was critical of the many Marxists in his university department who, he said, extolled the virtues of the Eastern bloc countries but had never visited them. 'You can't keep a people down forever', he would say, and he was right, though he didn't live to see that fence come down. My interest in negative freedom, in particular its physical dimension, and my distrust of paternalism and what I later learnt to be theories of 'positive' freedom (in Isaiah Berlin's sense of 'positive'),¹ certainly chime with those memories.

At school my favourite subject was history. I had an excellent teacher at high school who sparked an interest in radical politics and politicians of the nineteenth and twentieth centuries. My hero was Lloyd George, whose 'people's budget' introduced sweeping taxes on the rich.² In my later teens I developed a visceral dislike of the British Royal Family and the undeserved authority and privilege of the upper classes who 'toil not, neither do they spin'.

You've hinted at your interest in negative freedom, which has played a major role in your work. How did you form that interest? Was it the result of an overarching interest in political philosophy or, instead, did you come to political philosophy because you were interested in freedom?

I find it difficult to separate my interest in freedom from my interest in political philosophy, and also difficult to say that one springs from the other or vice versa. I guess the two developed in tandem and are related constitutively rather than causally.

My interest in negative freedom developed over my undergraduate years. I entered university considering myself basically left-wing but became increasingly frustrated by the collectivist and paternalist tendencies of the Left, and with their mistaken knee-jerk association of individualism with egoism and of freedom of choice with the defence of inequalities. This was during the Thatcher years in Britain, when political thinking was quite polarized, though not as much as it is today. Parallel to this, during my studies, it occurred to me that one could think of the history of

¹ See Berlin (1969).

² In 1909, Lloyd George introduced his 'people's budget' in the UK parliament. The budget sought to fund social welfare through income and land tax increases on the wealthy and was passed into law in 1910.

modern political thought as basically a history of rival interpretations of the ideal of human freedom. I became fascinated by this thought, and the more I studied, the more firmly I found myself siding with thinkers in the liberal tradition.

Given your interest in freedom, how strongly would you say that you sympathize with libertarianism, and have you always done so?

I wouldn't call myself a card-carrying libertarian, nor even a card-carrying left-libertarian, because I'm aware of how many empirical premises libertarians combine with their moral ones and I'm too wary about the empirical premises to identify wholeheartedly with a single package of political prescriptions. I'm also wary more generally of 'isms' in political thought, except as very rough categorizations. That said, I do indeed sympathize with some of the basic moral premises of libertarianism. This wasn't always the case. I first really got interested in libertarian thinking when studying Nozick with Hillel Steiner, which was an eye-opening experience for me even though I wasn't at all happy with Nozick's anti-egalitarian conclusions.

The distinction between left- and right-libertarianism is important here, as I see right-libertarians as defending indefensible inequalities, often on the basis of shaky analyses of freedom, even though freedom is supposed to be their basic value. Left-libertarians, more convincingly, take from classical liberalism the beliefs in self-ownership, private property, and markets, but combine those beliefs with the view that if each individual is really respected as a person, then each has no greater right than any other to a decent start in life. They agree with socialists about the injustice of most of today's economic and social inequalities, but they hold that the culprit isn't markets as such but the inequalities on the basis of which people enter markets and the exploitation that results. Some people think left-libertarianism amounts to squaring the circle, but that's too quick—another knee-jerk reaction. There's an intellectual challenge here that has been taken up in interesting and original ways by thinkers like Hillel Steiner, Michael Otsuka, and Peter Vallentyne.³

Charging left-libertarianism with squaring the circle seems, however, to be a strongly entrenched belief among laymen. Many would subscribe to the claim that right-wing positions usually entail a concern for freedom and left-wing ones a concern for equality. And few, among the

³ See Steiner (1994), Vallentyne (2002), Otsuka (2003).

general public, would deny that the two values are in opposition. If this is the case, what do you believe could explain it? Are political philosophers partly responsible for this?

That's an interesting question and one I often ponder. There are ways of interpreting the ideals of freedom and equality that place them in opposition, so people aren't wholly wrong when they assume that the two ideals conflict. For example, if realizing the ideal of equality involves achieving a situation where everyone pursues a similar kind of life with a similar degree of success, and this, in turn, leads to forms of collective control over people's lives, then equality conflicts with individual freedom as most liberals understand the term. That's why left-libertarians, and indeed most liberal egalitarians, are careful to clarify and circumscribe the kinds of equality their theories are prescribing. The error lies in generalizing from the specific interpretations where the ideals conflict, to a blanket claim about the incompatibility of 'liberty and equality', and that seems to be what's going on when people pigeonhole the ideals as 'right-wing' or 'left-wing'.

How responsible are political philosophers for this tendency? I guess they're responsible to the extent that they reinforce these generalizations through simplistic journalistic writings or through their university teaching. And, more specifically, to the extent that they confuse the promotion of freedom with the enforcement of property rights under the current distribution of resources, as many right-libertarian thinkers have.

You have mentioned repeatedly the distinction between left- and right-libertarians. You have also defended the left-libertarian stance as being truer to the most fundamental premise of libertarianism.⁴ A somewhat similar characterization of libertarianism can be found in the manifesto of 'bleeding heart libertarianism',⁵ a movement started in 2011 with the aim of reconciling free-market ideals and social justice ideas. What do you think about the movement, and its impact both on the general public and the scholarly environment? Have you ever considered adhering to it, or would you be sceptical of subscribing to another 'ism'?

I doubt the 'bleeding heart libertarians' have had a great impact on the general public so far, but ideas do take a long time to filter through from academia. They've surely increased awareness among right-libertarian

⁴ See Carter (2012, 2019).

⁵ See Zwolinski (2011).

thinkers, and perhaps also among some social democratic thinkers, about the theoretical possibilities of combining aspects of libertarianism with social justice or a concern for the poor or marginalized. As for me, the time I spend looking at blogs, let alone writing on them, is pretty limited (which isn't to say I'm negative about them: I'm just a slow-working hedgehog and life is short!). Regarding that blog in particular, I have strong misgivings about its name, which gives the impression that a concern for the disadvantaged is or should be motivated by compassion. Most of today's disadvantaged are due compensation as a matter of justice, rather than being the fortunate beneficiaries of duties of charity or of the consequences of adopting certain kinds of free-market policies starting from the current distribution of resources. Relatedly, my impression is that much of what has been presented as 'bleeding heart libertarianism' is less strongly egalitarian than left-libertarianism is, on my interpretation of the latter.

Given what you have said so far and the significant amount of time and intellectual energy you have devoted to 'politically charged' topics, such as those of freedom, human dignity, and respect, one might wonder whether you have ever been active in the political realm, or whether you have tried to keep your political and philosophical selves separate.

The last time I was at all active politically was as a member of the Labour Club in Newcastle. As I came to realize then, if you're campaigning as a member of a political party you have to toe the line on points you disagree with or are uncertain about, and for me this created an uncomfortable tension. So, while I admire many of the people who go into politics for the right reasons, I long ago decided that wasn't the route for me. In 2005 I published a book in Italian called *La libertà eguale* which was picked up on by a left-leaning reformist movement going by the same name—'Libertà eguale' (equal freedom). I enjoyed talking with them and exploring affinities, but the contact was somewhat limited given the abstract nature of my arguments. I realize my attitude to political activity can seem over-detached and even self-centred. As a philosopher, I feel only mildly guilty about that as I think there are others, many of them working at the more 'applied' end of the spectrum, who are better than I would be at propagating ideas.

II. FREEDOM, METHODOLOGY, AND RELEVANCE

You started your academic career with the analysis of freedom, its measurement, and its value—as the title of your PhD thesis suggests.⁶ The concern with freedom has permeated your entire body of work. Can you tell us why this theme is so important and explain the key aspects that you focus on?

The importance of the theme is something I've believed in since my student days, as I mentioned earlier. Individual freedom seems to play a central role in normative political theory, but does that role stand up to rigorous conceptual analysis? What initially exercised me was the fact that freedom is assumed to exist in degrees in so much of our political discourse and theorizing. People argue about whether freedom should be 'increased' and about 'how free' different individuals or societies are, relatively or absolutely, and normative political theorists have argued for the 'most extensive liberty', or 'sufficient liberty', 'equal liberty', or even for 'maximin freedom' (I'm using the terms 'liberty' and 'freedom' interchangeably here). And when they do so, they are usually assuming freedom to be valuable in some sense. But I found that what had been said so far about the value of freedom had rarely been connected up to the assumptions made about its measurability, and that those assumptions had themselves rarely been examined. So, I set out to fill these two gaps: to ask what sense could be made of the idea of 'more' or 'less' freedom, and to work out what our interest in that idea presupposes about the value of freedom. I was interested in particular in the role freedom plays in liberal normative theories.

And what is the role that freedom plays in liberal normative theories, in your view?

The key point is that liberals generally assume freedom to be an independent standard of evaluation—not just something that is defined in terms of other valuable things that we can already measure, such as utility (assuming we *can* measure that), or wellbeing, or violations of property

⁶ The title of Ian Carter's PhD dissertation is *The Measurement of Freedom* (Carter 1993). It was defended at the European University Institute of Florence (EUI) under the supervision of Steven Lukes, who was a professor of Political and Social Theory in the Department of Social and Political Sciences at EUI at the time (currently a professor of sociology at NYU) and Philippe Van Parijs, full professor at the faculty of economic, social and political sciences of the University of Louvain (UCL), and Robert Schuman Fellow at the EUI.

rights, or conformity to distributive patterns. That independent standard of evaluation can't be understood or operationalized unless we can provide an independent account not only of what freedom is, in this context, but also of how and why it counts as a fundamental value and what it means for it to exist in different degrees. I adopted a coherentist approach that aimed for a reflective equilibrium between all these elements. The investigation involved asking, among other things, what it means to have available a greater or lesser quantity of action—that took me into the problem of act-individuation in the philosophy of action—and whether different kinds of constraints on action can be commensurated, and how far they need to be.

Mentioning constraints makes one immediately think about a conventional distinction when it comes to freedom: the one between negative and positive freedom. According to many, this distinction has been largely surpassed by MacCallum's definition of freedom as a triadic relation.⁷ Is there still any use for the positive-negative dichotomy?

In *A Measure of Freedom* I avoided the terms 'negative freedom' and 'positive freedom' for the reason you mention. I thought they were largely redundant in light of MacCallum's analysis, at least for the purposes of fundamental theorizing. For MacCallum, as you know, any claim about freedom expresses a relation between an agent (first element), certain constraints or preventing conditions (second element), and certain actions or 'becomings' of the agent (third element), so freedom is always both negative and positive—freedom from something to do or become something. This allows for a spectrum of definitions, not just a dichotomy. But the classic distinction between negative and positive freedom has survived, and I accept that it has some utility as a rough categorization of two families of theories: freedom as the absence of obstacles imposed by other agents on actions of any kind, versus freedom as the absence of conditions that somehow impede self-realization or self-mastery—conditions that might be self-imposed, or imposed by nature, as well as other-imposed.

There is, however, still a debate about whether there is a third rough characterisation of freedom, which is republican freedom, as theorized by Pettit and Skinner.⁸

⁷ See MacCallum (1967).

⁸ See Pettit (1997) and Skinner (2002).

Yes. The republicans' contribution has captured the imagination of a great many political theorists, and so the rest of us have found ourselves adopting their categories when engaging with their work. Over the last two decades, a lot of political philosophizing about freedom has been about 'negative freedom' versus 'republican freedom'—whether they're really different and in what ways. That has been the main reason for my speaking of 'negative freedom' in more recent work.

My own view is that republicans have failed to carve out their promised third way, despite repeated attempts to do so. Either their view reduces to a liberal position focusing on the ways actions are rendered impossible by other agents, as I originally argued in my book and in a couple of subsequent articles,⁹ or else it produces very counterintuitive results that republicans themselves would not accept—for example, the result that virtually everyone everywhere is completely unfree.¹⁰ Or else republicans, in their continued efforts to distinguish their concept from the 'negative' one, have ended up changing the subject and talking about normative freedom or normative status—what sorts of things other people ought not to be allowed to do to you according to the law, or what your legal standing ought to be with respect to other people.¹¹ Republicans oscillate between these different positions, and this suggests that their concept of freedom is inherently wobbly.

But let's imagine that the republican characterization of freedom were a convincing third way. Would this then imply that 'their' freedom cannot be cashed out fully in terms of MacCallum's triadic relation?

Personally, I can't see any reason to deny that any of the specific positions between which republicans have oscillated, whether consistent or confused, appealing or counterintuitive, can be cashed out in terms of MacCallum's triadic relation. For example, the so-called 'robustness requirement' appealed to by republicans is just another way of characterizing constraints on freedom, the second element in MacCallum's triadic relation: Must a constraint have some non-trivial degree of probability in order to count as a limitation on your freedom, or is its sheer possibility sufficient for it to count? Does your freedom to do something depend simply on others not preventing you from doing it, or must those others be prevented from preventing you?

⁹ See Carter (2008, 2013b).

¹⁰ On this point, see Carter and Shnayderman (2019).

¹¹ See, again, Carter and Shnayderman (2019).

Let us look closer at your work on freedom. You have forcefully argued that the possibility of measuring freedom is a necessary condition for freedom to be valuable in the non-specific sense (that is, 'valuable as such'), that a person's overall freedom should be captured in a value-neutral way, and that the specific freedoms open to her should also be captured in a value-neutral way. These are three of your core arguments about freedom. Could you expand on how they relate to each other?

I began by asking why it is that we're interested in measuring freedom in the first place, and I came to realize that the reason has to lie in freedom being valuable as such. If freedom is valuable as such, I came to realize, it has to have a kind of value that I called 'non-specific', or what Matthew Kramer later called 'content-independent'. The basic idea is that freedom doesn't just gain value from its content as the freedom specifically to do *x* or *y* or *z*, which would depend on the other values that *x* or *y* or *z*, in particular, might help us realize or might be partly constitutive of; freedom is also valuable independently of those other values. If a particular set of freedoms has value not just in terms of those other values, but simply as freedom, then the value of that set must be a function not just of the value of the freedom specifically to do one thing rather than another, but also of 'how much freedom' that set contains. I call this idea of 'how much freedom' a person has their 'overall freedom'—where each person's overall freedom is some kind of aggregation over their specific freedoms. So, we're interested in how much freedom we have overall, because we value freedom as such. My conclusion was that overall freedom has to be understood, and measured, independently of any considerations about the values of the specific things we're free to do.

I also believe that for any one specific freedom—the freedom to do *x*—the question of whether or not you have that freedom is independent of whether or how far it's valuable for you to do *x*. In other words, the existence conditions for the freedom to do *x* are independent of the value of doing *x*. This claim could be defended as following from the account I've just given of overall freedom, given that overall freedom is just an aggregation of specific freedoms. But it's also a sensible stance to take even if you think there's no such thing as overall freedom.

Would the conclusion that freedom is to be measured independently of the values of the specific things we're free to do entail that freedom has to be measured in a value-neutral way?

Basically yes, but I should make a terminological point here: I would prefer to say, 'in a value-free way'. In my earlier work I did use the term 'value-neutral' to describe my proposed metric, but in a later article, published in 2015,¹² I tried to distinguish between value-neutrality and two other notions: 'value-freeness' and 'value-independence'. I think what you mean here by measuring freedom in a value-neutral way is what I would now call measuring freedom in a 'value-free' way. I do think a value-free metric for freedom is entailed by freedom being measured independently of the values of the specific things we're free to do.

That distinction you've just made, between 'value-neutral' and 'value-free', isn't immediately obvious. Could you expand on it?

A value-free concept is a concept that contains no ethically evaluative terms in its definition. A value-neutral concept, as I now understand it, is one that doesn't imply the superiority of any one member of a given set of substantive ethical positions. And that's not quite the same thing as being 'value-free' in the sense I've just mentioned. Value-neutrality is always relative to a given set of ethical positions, and you can think of it as matter of degree depending on the size of that set, whereas a concept is either value-free in my sense, or it's not. Value-freeness and value-neutrality are also both different from a third possible feature of political concepts, which I call 'value-independence'. I think of the value-independence of a concept as implying that you can justify its definition without any reference to ethical considerations, so only by reference to explanatory considerations. Many people think it's impossible for a concept like freedom to be defined in a way that I'm calling value-independent. I haven't taken a stance on the possibility or desirability of using political concepts that are value-independent, but I have clarified where I stand on value-freeness and value-neutrality, at least regarding the concept of freedom. If we distinguish in the way I've suggested between value-freeness and value-neutrality, then strictly speaking my claim is that our conception of overall freedom ought to be value-free. It might, in addition, be more or less value-neutral, but that's another issue from the one we're discussing here.

That's still quite abstract. Could you give an example of how these distinctions can be helpful?

¹² See Carter (2015).

One example that springs to mind is Robert Sugden's 2003 *Ethics* article,¹³ which contains an internal critique of my position, and where the distinctions we've just been discussing can help me provide an answer. Sugden thinks individual freedom (or opportunity, to use his terminology) has something like non-specific value, a point on which we agree; and he suggests that this implies freedom should be measured in a 'value-neutral' way. But he concludes, somewhat paradoxically, that freedom (or opportunity) can't be measured, because there's no such thing as a value-neutral metric for action. Freedom is valuable as such, so we're interested in how much of it we have, but it can't be measured. And that's a very puzzling conclusion. I think that what Sugden's argument really implies is that there's no such thing as a value-independent metric for action. He thinks any choice of metric will depend on adopting some evaluative perspective. This might well be true. But I also think that the non-specific value of freedom implies a value-free metric. Moreover, the value-neutrality of any metric, in my sense of value-neutrality, is a matter of degree. Once we make these distinctions, Sugden's criticism is much less worrying. What matters, if we're to capture freedom's value as such, is that we work with a value-free notion of overall freedom. That notion doesn't have to be value-independent, and it needn't even have a very high degree of value-neutrality in my sense: there can be ethical reasons for focusing on the physical dimension of action in explicating the notion of overall freedom, and that's unobjectionable inasmuch as we're not aiming for value-independence. And all of this can be true even if the particular metric I proposed is found wanting in other ways—for example, in terms of isomorphism with our common-sense comparisons—and so needs to be revised.

So, your argument that the value of overall freedom should be captured in a 'value-free' way relies on the idea that value-based approaches to freedom are unable to capture freedom's non-specific value—they reduce this latter value to the specific values of the things a person is free to do.

Yes. A value-based metric implies that you're freer the more valuable your specific options are. My argument is that value-based metrics don't really capture degrees of freedom. Rather, they capture the values that the freedom to do x or y or z help to promote causally, or of which those specific freedoms are partly constitutive. I think Dworkin and Kymlicka are right to point out that value-based metrics make the language of overall

¹³ See Sugden (2003).

freedom normatively redundant: everything can be captured by speaking of the instrumental or constitutive values of specific freedoms.¹⁴ But I disagree with the conclusions of Dworkin and Kymlicka: they only consider value-based metrics, and so conclude that the language of overall freedom must be normatively redundant.

Kramer, however, has questioned this point, arguing that, while freedom's non-specific value is not dependent on the values of the things one is free to do, it does depend on the values of the specific freedoms to do x or y or z.¹⁵ You have an earlier paper on this issue.¹⁶ What is your take on this point? Have you changed your mind since?

Kramer proposes what I call a hybrid account (in the earlier paper you've just mentioned, I called it a dualist account): he thinks degrees of freedom are primarily a matter of the physical dimensions of available action, but he introduces evaluative multipliers, so his metric reflects both non-specific value and specific value. He qualifies his value-based metric by saying that the multipliers should be formulated only in terms of the values with which freedom is non-specifically connected—the values to which freedom is also a means, or of which it is partly constitutive, in a non-specific way. I confess I've never grasped why one should think that this last move answers the accusation that value-based metrics confuse the specific and non-specific value of freedom. So, no, Kramer hasn't changed my mind on this issue. Kramer's evaluative multipliers are still based on the value of being free specifically to do x or y or z. That fact isn't changed by restricting the set of ultimate values in terms of which the specific value is measured.

I also have some more specific criticisms of his use of evaluative multipliers, which I'd better not go into here. They're set out in the final part of a long joint paper with Hillel Steiner that's forthcoming in a festschrift for Matt Kramer.¹⁷ Matt recently told me that some of the replies he's written are even longer than the papers they're replying to(!), which is somewhat daunting, though I'm looking forward to reading them.

So, are there no cases in which value-based conceptions of freedom fare better? Take, for instance, situations in which one wants to measure

¹⁴ See, for instance, Dworkin (1979) and Kymlicka (1990).

¹⁵ See Kramer (2003, 242)

¹⁶ See Carter (1995).

¹⁷ See Carter and Steiner (forthcoming).

the ‘value of freedom’, rather than its extent. Would a value-based understanding of freedom then be more appropriate?

Yes, if what you mean by ‘the value of freedom’ is ‘the value of freedom in terms of values other than freedom’! As I see it, the extent of freedom is just the value of freedom measured in terms of freedom itself. Freedom has value in both of these senses—value as such, and value in terms of other things it brings about or is partly constitutive of. Both kinds of value matter. So, value-based metrics of freedom can certainly be useful, but they aren’t what their authors claim them to be: by calling them metrics of freedom, they hide from view, and so fail to capture, the extent of freedom, which also matters for evaluative purposes. Or, more commonly, they capture both, but fail to make the distinction clearly.

In your book *A Measure of Freedom* you argue that there is a difference between freedom-based justice and justice-based freedom.¹⁸ With this, you mean that there are some definitions of freedom that are moralized and some that are not. You maintain that only the latter can play an appropriate role in a theory of justice. Can you expand on this more, and your reasons for it?

The answer to this question takes us back to a point I made earlier about the liberal normative assumption that the notion of freedom provides an independent standard of evaluation, referring as it does to a fundamental value. This means that, for any proposed set of rights, we want to be able to say what that set implies in terms of freedom. If we are liberal normative theorists, we ought to be able to defend a particular set of rights on the ground that it is good for freedom. For example, if liberals favour private property over communal property, one reason one would expect them to be able to give is that private property is, at least on the whole, better for freedom. You can’t say this and then go on to define freedom as the absence of constraints that violate private property rights. That sort of justice-based definition, or ‘moralized’ definition, would rob freedom of its role as an independent standard of evaluation. Freedom is then no longer a grounding value. This was all set out very nicely by G.A. Cohen in a series of articles in the 1980s and 1990s.¹⁹ I would say, in addition, that freedom ought to be seen as a grounding value by any liberal attaching non-specific value to freedom.

¹⁸ See Carter (1999).

¹⁹ See, for instance, Cohen (1988, 1995).

I find it interesting that moralized definitions straddle the left-right divide, something that Cohen didn't notice or at least didn't bother to point out. For example, both Dworkin and Nozick seem to presuppose moralized definitions of freedom.²⁰ In fact, Dworkin is more explicit than Nozick about doing so. This fact is itself useful in illustrating how moralizing the concept of freedom makes freedom itself redundant as an independent standard of evaluation. If they both moralize the concept of freedom, neither Dworkin nor Nozick can appeal to the value of freedom in order to show what's wrong with the other's conception of justice.

This point about the left-right divide, and the difference between moralised and non-moralised conceptions of freedom brings me to your argument that it is wrong to cash out the divide between laissez-faire liberals and egalitarians with regard to freedom as a divide between formal freedom and substantive freedom.²¹ Does this spell the end of the possibility of finding a 'rough and ready' way to characterize the contrasting views of laissez-faire liberals and egalitarians when it comes to definitions of freedom? Or are there other ways to draw a line between the two camps that you would favour?

I don't think there's a way to draw the line that's quite as 'rough and ready' as the simple distinction between merely formal freedom and substantive freedom—a distinction which can be theoretically useful but which I don't think captures the divide between laissez-faire liberals (more specifically, anti-redistributive liberals), and economic egalitarians. I do think the divide can still be captured, at least in part, in terms of differences between conceptions of freedom. One such way is by reference to G.A. Cohen's work, which I've just mentioned. He saw the difference as one between a moralized definition of freedom as the non-violation of private property rights, and a non-moralized definition, where the latter might be either freedom as not being prevented by others from performing actions (that is, negative freedom in Isaiah Berlin's sense), or freedom as the ability to perform actions, more along the lines favoured by Sen and Van Parijs.²²

Cohen's account of the divide of course also amounts to a critique of anti-egalitarian libertarianism: if you believe in equal freedom, then it's hard to deny that a non-moralized conception will have economically

²⁰ See, for instance, Nozick (1974) and Dworkin (2001).

²¹ See Carter (2011a).

²² See, for instance, Sen (1988, 1996) and Van Parijs (1997).

egalitarian implications. So, either the anti-egalitarians should explicitly defend the appeal to a moralized definition, rather than merely presupposing it implicitly, or they should admit to favouring unequal freedom. The first alternative doesn't seem to be very popular, inasmuch as there are very few explicit defences of moralized definitions (although Ralf Bader's recent work is an example,²³ as is the work by Ronald Dworkin that I mentioned earlier²⁴). The second route seems even less popular, as anti-egalitarian libertarians want to divorce questions about the distribution of freedom from questions about the distribution of resources. There might be other ways of showing how anti-egalitarian libertarians embrace a distinct concept of freedom. Hayekians think of freedom as the absence of arbitrary power, which looks rather different from negative liberty in Isaiah Berlin's sense. But then, Philip Pettit thinks of freedom in much the same way as Hayek at this abstract level, and I would classify Pettit as more of an egalitarian. So here, too, the tendency to moralize the definition of freedom might be the only way to explain this ideological divide by reference to rival definitions of freedom. This makes the critique of moralized definitions quite a potent tool in normative theorizing.

Speaking of justice and freedom, another topic you tackle in A Measure of Freedom is the appropriate role of freedom in a theory of justice. In chapter 9 of the book, you reject Steiner's argument for equalizing individual freedom in a society.²⁵ The reason behind your argument has to do with your rejection of the view that 'a universal quest for greater freedom' is a zero-sum game.²⁶ Could you expand on this (and your reasons for it) more? Do you believe that there is an (or a more) appropriate way of distributing freedom in a society?

I agree with Steiner that individual persons have a right to equal freedom. I base this claim on the premise that persons are basically equal in a morally relevant sense and are due respect as such. So, equality is certainly one basic distributive principle when it comes to allocating freedom. But, as your question implies, if equality were the only distributive principle for freedom, this would have to be because the allocation of freedom is a zero-sum game. Otherwise, we'd be indifferent between levelling-up and levelling-down people's degrees of freedom: 'very little freedom for everyone' would be a perfectly just distribution! Steiner is well known for

²³ See Bader (2018).

²⁴ Carter refers to Dworkin (2001).

²⁵ See Carter (1999, 258–267).

²⁶ On this, see Steiner (1983, 1994).

having defended the zero-sum thesis. He thinks one can never increase or decrease the total amount of freedom enjoyed by a given group of individuals. I've criticized the zero-sum thesis, arguing that the total amount of freedom of a group, understood in the 'value-free' sense we've already talked about, can indeed increase or decrease, depending on how property rights are understood, and also depending on degrees of scarcity and propensities of individuals to cooperate and to consume resources. So, we need to combine equality of freedom with some principle prescribing a certain level of freedom for all. The strongest combined principle would be maximal equal freedom, though there might be other normative considerations that count against that.

In an article for another festschrift—this time for Hillel Steiner—I argued on this basis that Steiner's zero-sum thesis actually plays a key role in his theory of justice, for methodological reasons.²⁷ Steiner wants to characterize justice in a way that's wholly independent of the good. It can be helpful to compare him with Rawls in this respect. Rawls thinks a conception of justice requires at least a 'thin' theory of the good (for Rawls, having more 'primary goods' is better than having less). For Steiner, that's a cop-out. He thinks it doesn't take seriously enough the priority of the right over the good. He wants his theory of justice to be wholly independent of any claim about what's good for individuals. So, in my terms, his idea is that we ought to analyse the concept of justice, no less than that of freedom, in a 'value-independent' way. But if the zero-sum thesis is false, then he needs to appeal to the claim that freedom is good—that having more of it is better than having less, at least *ceteris paribus*, in order to say what justice consists in, exactly. That would make the analysis of justice dependent on a particular ethical evaluative stance. And I think this partly explains why he's remained so strongly attached to the zero-sum thesis.

III. FORMAL APPROACHES TO THE MEASUREMENT OF FREEDOM

In your work, you engage to some extent with the philosophical literature that uses formal tools to analyse social phenomena, while not being a 'formal' philosopher yourself. How important has this approach been in influencing your ideas?

²⁷ See Carter (2009).

Around the time of the publication of my book *A Measure of Freedom*, and over the subsequent decade or so, I did engage quite a bit with formal theorists with backgrounds in social choice theory or philosophy or both. I was aware of the formal literature in welfare economics when working on the book, but, beyond developing a critique of Sen's work on freedom, I didn't engage with it actively until afterwards, when the 'freedom of choice literature', as it came to be known, had really started to take off. During that period Martin van Hees, especially, was instrumental in bringing together a number of political philosophers and rational choice theorists interested in freedom, and several of us organized research projects and workshops in this interdisciplinary spirit. Those workshops were fun meetings and always very stimulating.

In terms of what shows up in my published work, the impact of the formal approach probably looks fairly limited. No doubt this has partly been due to my own limits in following the more technical passages, but also to the fact that sometimes I found that the formal literature started from axioms that the authors took to be self-evident but which I had philosophical reasons for doubting. As a result, whichever literature I was looking at I mostly found myself digging down to the foundations. So, for example, I argued that Pattanaik and Xu's original axioms (in their seminal 1990 paper²⁸) were running together the three distinct concepts of freedom, choice, and freedom of choice, and that separating out these concepts could help dissolve some of the perplexities their analysis had generated.²⁹ This work wasn't just critical, as I had reflected very little on those distinctions and I found it very helpful to do so. The concept of choice is interesting in itself: in one sense it's broader than the concept of freedom, as the choices we have include powers as well as freedoms. I've recently discovered some interesting practical applications of that broader concept working with Stefano Moroni, a specialist in planning theory.³⁰

What, if anything, do you believe is gained (or lost) by the use of formal tools in the analysis of social phenomena, in general, and of freedom, in particular?

In general terms, the use of formal tools certainly brings clarity, rigour, and objectivity. You can't argue with a mathematical proof. That said, the

²⁸ See Pattanaik and Xu (1990).

²⁹ See Carter (2004) on this.

³⁰ See Carter and Moroni (2021).

costs and benefits of using a very technical language seem to me to vary depending on the theoretical context, what one's trying to demonstrate, and, more pragmatically, the nature of the audience. Analytical philosophers, who are of course strong believers in clarity and rigour, mostly get by without using more than the most rudimentary formalizations. But sometimes more technical tools can help in making a demonstration crystal clear and avoiding fallacies or sophisms, of which mainstream political philosophy certainly has its fair share.

Sometimes formal approaches are criticized for their unrealistic, idealizing assumptions—for example, those made about degrees of rationality or self-interest. I have limited sympathy with that kind of criticism. Theorizing involves abstracting, and abstracting involves removing parts of reality. If we don't remove parts of reality, we simply redescribe reality in all its perceived complexity and fail to formulate an explanatory or normative theory about it. In any case, this kind of criticism doesn't seem to apply to the 'freedom of choice literature' in the same way as it might apply to standard rational choice theory.

Thinking more specifically about how things have worked out in the case of freedom, one kind of loss is the one I pointed to a moment ago: many formal theorists seem to take their lead from only a small number of well-known philosophical texts on freedom, such as those by Mill, Berlin, or Sen, devoting nearly all their energy to their formal analysis; as a result, they run the risk of producing work that proceeds with great rigour but from dubious premises or that simply demonstrates the obvious. That said, there's nothing about the use of formal tools *per se* that makes this kind of risk inevitable.

More specifically still, and as some formal theorists themselves have pointed out, the 'freedom of choice literature' has tended to neglect the problem of identifying and assessing different types and sources of constraints on freedom. In other words, in MacCallum's terms, it mostly treats freedom as a dyadic relation rather than a triadic one, because it conceives of freedom simply as the presence of a menu of options, without asking what it is to open or close off an option. The main focus has been on how to aggregate the options, whether and in what ways preferences over options count, how to gauge degrees of similarity among options on the basis of individual preferences, and so on. Of course, this is just another way of abstracting, and isn't bad in itself. Still, it's important to be conscious of the fact that you're treating only one dimension of a multidimensional phenomenon—especially if you think, as I do, that the

plausibility and usefulness of a conception of freedom depend on its overall coherence in wide reflective equilibrium, taking into account all its dimensions, the ways in which it's supposed to ground other normative concepts such as rights and justice, and various philosophical background theories.

Finally, there's the more pragmatic issue of accessibility. Obviously limited accessibility means less impact among the audience you'd like to reach, and this is a danger when the language is very technical and the potential audience is broad. If you're trying to communicate in an interdisciplinary context—as you ought to be if you're working on freedom—it's crucial to include intuitive explanations in plain English for the benefit of those who lack the training to follow your formalizations with sufficient confidence.

To the best of my knowledge, there are two strands of the philosophical literature concerned with the measurement of freedom: one interested in the cardinal measurement of freedom that you and Hillel Steiner, for instance, contributed to, and another that purports to measure freedom (of choice) through the ranking of opportunity sets. They seem to be talking past each other, despite the appearance of being closely related. Is there a reason?

I find the term 'cardinality measure' somewhat ambiguous in this context, as it often seems to be used to cover different things—most importantly, the view that the measurement of freedom is a matter of adopting a simple counting procedure, and the view that the measure should be 'value-free' in the sense mentioned earlier. Here, two qualifications are in order. First, it's important to bear in mind that these two stances are logically independent of one another, even though the denial of the first seems mostly to have been accompanied by a denial of the second, leading to the development of various preference-based rankings of opportunity sets. Second, I don't exactly favour a simple counting procedure but favour aggregation over expected sets of conjunctively unprevented options. Still, roughly speaking, cardinality and value-freeness do characterize the position adopted by Steiner and me.

There has been some important work at the interface, so I wouldn't say that the two approaches you mention have been completely talking past each other. For example, there have been points of contact where formal theorists have attempted to produce non-preference-based

rankings of opportunity sets—as in the work of Martin van Hees.³¹ But there's certainly some truth in what you say. Most of the reasons seem to be implicit in my answer to your last question. Some are more superficial, some deeper, and I find it difficult to assess their relative importance. At the superficial end of the spectrum, theorists have talked past each other simply because they are unaware of, or lack an interest in, the fundamental concepts and language being developed on the other side of a disciplinary divide. For my part, I have paid less attention than I might have to alternative metrics developed in the 'freedom of choice literature'.

At a deeper level, there may be some differences in the reasons for our interest in the concept of freedom that generate different views about which problems are important and which solutions would be adequate. For example, welfare economists are generally happier to work with weaker comparisons. And this is natural if you're coming from an area where the main currency of evaluation has been utility, understood as preference satisfaction, and where one's main concern may be explanatory as much as normative. Starting from preferences over available options as a means of explaining individuals' economic behaviour, some welfare economists arrived at the interesting proposition that often people have a preference for having more options. So, the reasoning goes, let's try and make sense of preferences for freedom by discussing possible rankings of opportunity sets in terms of the freedom of choice they imply. If, on the other hand, you're coming from mainstream normative political theory—say, in the tradition of Rawls, or of right- or left-libertarianism—where the main concern has been equality, rights, and justice, and you're asking what sorts of institutional arrangements could realize these values, either in ideal or in non-ideal circumstances, then one of your immediate concerns ought to be whether we can make sense of cardinal interpersonal comparisons of freedom, as these comparisons are necessary in order to make sense of some of the most frequently cited principles of justice, or at least in order to compare different approximations to the ideals those principles represent—for example, approximations to equal freedom or to maximal equal freedom.

If you're coming from the direction of a deontological theory of justice, there seem to be implications also for the question of preference-dependence—which, as I've said, is a separate issue from that of cardinality versus ordinality. If you adopt a deontological perspective on distributive questions, you're likely to be sceptical about attempts to measure

³¹ See, for instance, van Hees (2004).

freedom in terms of agents' preferences. You'll be more likely to think, along Kantian lines, that freedom, in this context, is an external relation between persons and that agents' preferences are neither here nor there. This is certainly Steiner's perspective, and in my own work I aimed for an account of overall freedom that would at least be compatible with it.

You have done empirical work with regard to the measurement of capabilities,³² but not with regard to freedom itself. This might come across as surprising, given that there are specific indices devoted to the measurement of freedom across countries. An example would be the Human Freedom Index. Has any of the institutions that develop these measures ever reached out to you? Would you accept a task of counseling on the conceptual basis needed for those measurements?

I haven't really done empirical work on the measurement of capabilities. I take it the publication you're referring to is the one that came out of a research project directed by Paul Anand that resulted in a joint article. My contribution to that collective effort was mostly theoretical.

Regarding empirical measures of what you call 'freedom itself', or what I'd call social freedom or negative freedom: back in 2010 I was indeed invited to one of the workshops jointly organized by the Fraser Institute, the Cato Institute, and the Friedrich Naumann Foundation that led to their Human Freedom Index. I criticized some of their assumptions about the relation between freedom and property rights, by which they meant those property rights recognized in positive law (in other words, I was posing a variant of the critique of moralized definitions of freedom). They seemed to be divided over the usefulness of engaging in discussion of such an issue, many of them urging that they should 'just get on with developing the index', so I guess my impact there was pretty limited. Around the same time, I attended a couple of interesting workshops organized by the political scientist Leonardo Morlino, who was interested in measuring freedom as one dimension of the quality of democracy. I'd certainly consider further work in this area, though sadly time is always scarce. One project I've had in mind for many years would be to work with a political scientist on comparing the existing indices and the conceptions of freedom they assume, and the different implications of specific conceptions that are often considered to be 'rivals' yet might in fact imply quite similar indices once operationalized adequately. For example, I think social or negative freedom and republican freedom are unlikely to

³² See Anand et al. (2009).

imply different indices, and that the differences between the implications of these two conceptions and those of Sen's 'freedom as capability' might be fewer than is often supposed.

IV. EQUALITY AND RESPECT

Let me now turn to a more recent theme in your work: equality and respect. In a 2011 paper,³³ you write that we should ask what the basis of equality is. More specifically, you point out that the question of what we should equalize in a society (resources, well-being, or other things) necessarily depends on the basis of equality. Can you expand a bit more on this?

I first came to think about the basis of equality—the question of what makes us equal in a morally relevant descriptive sense—through a sense of dissatisfaction with certain prescriptive claims about equality. In particular, I was focusing on claims to the effect that certain human capabilities ought to be equalized. Sen and his followers have pointed out that humans are naturally unequal in their capacities to convert resources into functionings, as a result of which their capabilities are unequal. Yet Sen is an egalitarian. And it struck me that his affirmation of descriptive inequality, though illuminating, was depriving him of a sound reason to equalize those capabilities. Unless, that is, he could point to some other sense in which people are, in fact, equal, such that those capabilities ought to be equalized. Yet he, and others, have steered clear of that further question. If we're not actually equal in any sense, then why should anything be equalized? Treat equal cases equally, unequal cases unequally. Aristotle, who is often cited by capability theorists, recognized natural inequalities in the capability to function, but he didn't prescribe equalizing any such capabilities. Was he therefore more coherent than contemporary advocates of equality of basic capabilities?

And what was your answer to these questions? Is there a sense in which we are equal?

I developed an argument that starts from Rawls's claim that we're equal inasmuch as we all have the 'range property' of moral personality, where a range property is the property of having certain scalar properties above a certain minimum threshold. I argued that we need an independent

³³ See Carter (2011b).

reason for focusing on such a range property, one that isn't itself based on equality, otherwise our justification of equal entitlements will be circular. That independent reason, I suggested, lies in a kind of respect, which I called 'opacity respect'. To show opacity respect for a person is to adopt an external perspective, refusing to 'look inside' them, and consequently refusing to take account of the level at which they possess certain agential capacities above the threshold. In other words, respecting agents means taking their capacities as given, in our practical deliberations about how to treat them, and simply ascribing the range property to them, without further investigation, because there is something disrespectful about assessing the very capacities on which an agent's moral personality supervenes. Opacity respect might not be appropriate in all contexts, but it does seem to be appropriate in those contexts where we think people should be treated as equals—such as the context of relations between the state and citizens—and in this sense it can explain our commitment to that kind of treatment.

This basis of equality seems to be contradicted by some answers to the question of what we should equalize in society—the so-called currency of egalitarian justice. For example, some versions of the capability approach, and some versions of luck egalitarianism, prescribe equalizing certain 'internal resources' of people or, more commonly, compensating for internal resource deficits by supplying those who have such deficits with more external resources. But, if those 'internal resources' are among the capacities on which moral personality supervenes, then this policy can't be carried out without violating opacity respect. If we reject opacity respect, we no longer have a reason for focusing on the range property. And if we don't have a reason for focusing on the range property, we're back to treating people as unequal. So, my conclusion was that any egalitarian prescription, any answer to the question 'Equality of what?', has to be consistent with opacity respect in order to have a logically consistent justification.

What has been the response of luck egalitarians or capability theorists to your argument (if any)?

Some luck egalitarians have responded either by attempting to deny the entailment that the equalization of internal resources is ruled out, or by rejecting my starting premises—a sort of modus tollens argument, to which I still prefer my modus ponens argument as I haven't yet seen a convincing alternative account of the basis of equality. Gabriel Wollner,

for example, rejects my account of the basis of equality because of its undesirable implications for luck egalitarianism, and as an alternative basis of equality he goes for ‘being human’,³⁴ but I think that alternative account runs into the usual problems of speciesism or of over-inclusion. Kasper Lippert-Rasmussen, who is broadly sympathetic with luck egalitarianism, has taken the line that ‘basic equality’, as it’s come to be called, doesn’t do all the grounding work that it’s often thought to do, and has defended this view in his more recent work.³⁵ If he’s right, this might deprive my argument of some of its teeth. I’m still trying to work out why, exactly, I disagree with him, although I’m pretty sure I do! This is still work in progress.

I’m not aware of any capability theorists having addressed my argument excluding the equalization of certain basic capabilities, but perhaps many of them don’t need to. Although Sen originally advocated ‘equality of basic capabilities’ in answer to the ‘equality of what’ question, and although Nussbaum’s list of capabilities includes some very basic ones that I would think of as grounding moral personality, most capability theorists today seem to be closer to sufficientarianism than to egalitarianism in the strict sense, and so might escape my critique. That said, whether they do escape it might depend, further, on how they justify their sufficientarianism. For example, if they endorse a contractualist justification, which itself assumes basic equality, they might still be subject to my criticism.

As you have just pointed out, consistency between opacity respect and egalitarian prescriptions seems to rule out many of the egalitarian theories developed so far. Which one(s) does it not rule out?

Any theory that focuses on external resources or external relations will pass through the filter, as its application won’t involve assessing, or taking into account, levels of internal resources—in the sense of capacities in virtue of which we count as moral persons. Equality of social or negative freedom will therefore pass the test. So, a theory like Steiner’s, in which the most basic principle is equality of pure negative freedom, passes the test. So too, Rawlsian egalitarianism passes the test, as it focuses only on primary social goods and not on primary natural goods.

Seeing what did and did not pass the test was quite a revelation for me. It confirmed some of my long-standing intuitions—in particular

³⁴ See Wollner (2010, 2014).

³⁵ Carter refers to the working paper that Lippert-Rasmussen presented at the 2021 Mancept workshop, entitled “What Is It for Us to Be Moral Equals? And Does It Matter (Much), If We’re Not?”. See <https://mancept.wordpress.com/basic-equality/>.

about the importance of external freedom in a theory of justice—but also led to some surprises, as I had generally thought of myself as closer to luck egalitarianism than to Rawlsianism. Some aspects of luck egalitarianism—for example, a certain version of responsibility-sensitivity—survived this journey. Nevertheless, I have come to appreciate Rawls more than I did.

I also think that more can be said about the currency of egalitarian justice, in light of my account of basic equality, than simply pointing out which currencies can be equalized without violating opacity respect. For example, if your starting point is opacity respect, then a freedom-based theory of egalitarian justice seems to be more congruent with your fundamental egalitarian beliefs than a welfarist one. But this, too, is still work in progress.

Let me pick up on the second reply from egalitarians you mentioned, that of rejecting your starting premises. If we rejected opacity respect in favour of an alternative account of basic equality, could it be that this alternative basis of equality has no implications for what has to be equalized in a theory of justice? In other words, could the question of ‘equality of what’ then be answered independently of the question ‘what are the bases of equality’?

No, even at a more general level, I don’t accept that the two questions can be largely independent. To supply a basis for equality—to provide an account of basic equality—is to say what it is about certain individuals that makes them equal, such that they ought to be treated equally in some respect. Put this way, it should be clear that there’s an entailment-relation between the two kinds of equality: people ought to be treated equally—they ought to receive equality of some particular kind of thing *x*—because they are equal in some sense that is normatively relevant in determining entitlements to *x*. The nature of our equal entitlements is grounded in the content of our basic equality. That said, there might be some leeway. The nature of basic equality might constrain rather than completely determining the currency of egalitarian justice.

For the most part, the question ‘Equality of what?’ has indeed been addressed independently of the question ‘What are the bases of equality?’, but I think this is simply because people have generally pushed the second of these questions firmly to the back of their minds, persuading themselves that they can remain agnostic on such a ‘deep’ question when engaging in normative theorizing about equality. As a result, much of the

literature on ‘equality of what’ seems to have proceeded more through a sort of ‘intuition pumping’—that is, by comparing abstract cases of equality of certain kinds of good and asking whether such distributions really capture our intuitions about what egalitarianism truly amounts to. For example: ‘If egalitarianism meant equality of welfare, then the expensive tastes of the rich might lead us to give them more resources in the name of equality. But redistribution from the poor to the rich can’t be something true egalitarians believe in. So, “welfare” can’t be the right answer to the question “Equality of what?”’. This style of reasoning can be complex and interesting, but it only takes us so far. Once we see the relevance of basic equality to the ‘equality of what’ question, we realize that we also have to dig down to the normative grounds of equality of entitlements and not merely seek isomorphism with surface intuitions about what we ought to equalize.

Since you have mentioned intuitions, I would like to make a brief detour from the topic of equality to that of the role of intuitions in political philosophy. While they are often invoked in support of someone’s argument or against the implications of an argument, it is often unclear what exactly their normative role is in philosophizing, and where they derive their normative force from. Since you have sometimes used intuitions normatively in your own work, I wonder what your thoughts are on this issue.

The term ‘intuitions’ can refer to different kinds of beliefs which can be more or less superficial and, in that sense, more or less authoritative in our theorizing. There are linguistic intuitions, which tell us ‘what we would say’ in certain circumstances. These might seem superficial, and in one sense they are. But, when we analyse them carefully, they can also tell us something about the nature of the concepts we use, and so reveal deeper normative beliefs. I tend to follow Rawls in thinking of these normative beliefs, or ‘considered judgements’ as he calls them, as the appropriate starting point in the development of any normative theory. After all, where else can we start? But the making of a good theory doesn’t just lie in mirroring our raw intuitions. A theory needs to be internally consistent—something our raw intuitions often aren’t—and to have a plausible structure linking more basic, grounding claims with the less basic ones that are grounded in them, and it needs to cohere with other theories in wide reflective equilibrium. If a theory we’ve developed turns out to be highly inconsistent with our initial intuitions, then we have grounds for

rethinking it. Hence, for example, my own interest, when working out a theory of overall freedom, in the consistency of that theory with certain ‘common-sense comparisons’ of freedom that we ordinarily make. On the other hand, I do find myself getting a bit frustrated when I read arguments that seem to move too quickly in rejecting some theoretical claim on the ground that it conflicts with some unanalysed raw intuition. I mean arguments of the form: ‘Claim *x* can be shown to entail *y*; but *y* is “highly implausible”; therefore, we must reject claim *x*’. Raw intuitions are the first word, but they’re not the last word. We often have to revise them in our theoretical efforts to achieve overall coherence, and sometimes these revisions can be surprising and interesting.

Let me now focus more closely on respect. You also have also one article on respect and toleration.³⁶ You argue that it is true both that respect and toleration are compatible and, in another sense, that they are not. Can you expand on this?

Having developed the notion of opacity respect, I came to see that it had implications for principles other than those of equality. Toleration is one example. It’s generally assumed that tolerating something—say, a certain kind of person or practice—involves evaluating it negatively. Toleration is more complex than mere indifference or approval of something. It involves holding back from acting on beliefs or tastes that would otherwise lead us to curtail other people’s freedom. This is sometimes called the ‘objection component’ of toleration—the disapproval or dislike of some person or belief or practice—which is overridden by an ‘acceptance component’—the reason for not interfering after all. Although toleration has traditionally been seen as an important part of the theory and practice of liberalism, some have objected that it is ‘disrespectful’, exactly because it involves a negative judgement. I came to see that there’s something right and something wrong in this claim about the incompatibility of toleration and respect. First, the claim can be based on a simple confusion of ‘recognition respect’ with ‘appraisal respect’. It remains the case that toleration is compatible with recognition respect—that is, with recognizing people’s status as agents with rights to freedom. It’s only incompatible with appraisal respect—that is, respecting in the sense of holding someone or something in high esteem. But there’s also a more surprising sense in which toleration can be incompatible even with recognition respect. This is where recognition respect is interpreted as opacity respect, and where

³⁶ See Carter (2013a).

toleration involves taking into account some of the assessments that opacity respect rules out. Opacity respect involves refusing to make, or at least to take into account, certain evaluations of people's basic agential capacities. So, where the objection component of toleration consists in this specific kind of basic negative evaluation, then toleration is disrespectful in the sense of violating opacity respect. Moreover, given the line of reasoning I rehearsed earlier, failing to show opacity respect involves denying the very basic equality that grounds toleration understood as a liberal democratic virtue. So, where toleration is understood as a liberal democratic virtue, it can't be toleration of the kind that involves opacity disrespect—for example, the kind shown by the so-called 'tolerant racist'.

I would like to close this part of the interview with a question which maybe brings us full circle. This regards the interaction between freedom and equality. Would you say that your interest in equality is ultimately justified by an interest in freedom, or the other way round—that your interest in freedom was ultimately justified by an interest in equality?

I think the justification goes both ways. If you're interested in negative freedom because you observe that certain relations of oppression or confinement are unjust, then in part you're rebelling against the hierarchies that typically establish or legitimize those relations. Vice versa, it would be strange to say you're interested in equality without having at least some vague notion of the content of that ideal—of what ought to be equalized, or made less unequal, between certain people. In this sense, I find the two ideals inseparable, even though my first interest as a political philosopher, chronologically speaking, was in the concept of freedom, and for the most part, when theorizing about freedom, I abstracted from its relation to equality.

The inseparability of the two ideals, at least as I've interpreted them, has been brought home to me even more clearly by focusing on the notion of respect. Respect for persons is what grounds equal freedom: individual persons are equal, in a morally relevant way, insofar as they are due opacity respect; the object of opacity respect is people's agency; as agents, persons are due freedom; as equal persons, they are due equal freedom. And at each stage in this reasoning, the appropriate perspective on persons is an external one that doesn't involve 'looking inside' them. That perspective amounts to a kind of respect, and it grounds both equality and negative freedom, where the latter is understood as an external

relation between actions. None of this need imply that justice consists only in such relations, but it does say something about the connection between equality and freedom that I, personally, find intuitively appealing.

V. ADVICE TO YOUNG SCHOLARS

Let me close off the interview with some questions that look to the future, and, more specifically, to future generations of philosophers. What advice would you give to graduate students aiming to pursue an academic career in political philosophy?

The first thing that comes to mind is: make the most of being a full-time researcher while it lasts! Once you have an academic job, you won't have nearly as much research time. Looking back, my years as a graduate student and as a post-doc seem to have been incredibly free, although at the time of course it doesn't seem that way because of the feeling of insecurity and needing to find your way.

Regarding political philosophy in particular, it can be helpful to be aware of how vast the discipline is and how it borders, at one end, on moral philosophy, philosophy of language, metaphysics and so on, and at the other, on political science, economics, and law. When applying for jobs, if you have the luxury of being able to choose, think about where you'd feel more comfortable—in a philosophy department, if your work is more foundational or conceptual, or in a social science department, if your work is more applied or informed by empirical research. That includes thinking not only about the kinds of researchers you'd most like to interact with but also about the kinds of students you'd most like to teach.

You have hinted at the philosophy job market and the extent to which young scholars can have 'bargaining power' over their choice of where to teach and do research. What PhD students are usually told is that this largely depends on their publications. Would you have any suggestions about this more specifically?

Well, I can give a few pieces of strategic advice based on my experience as an author and as a referee. First, try to be thick skinned. Philosophy journals pride themselves on the number of papers they reject, and often they 'desk reject' pieces for fairly arbitrary reasons. Having a paper rejected after working on it for a long time feels a bit like a punch in the

stomach. But don't let it get you down: nearly all of us clock up a fair number of rejections even though many people don't like to admit it. You get over it after a day or two; and if no reason is given, or if you think the referees haven't provided strong reasons, don't hesitate in sending the piece off to another journal. Don't let it sit on your hard drive doing nothing. Of course, if you think the referees have provided convincing objections, that's another matter. Second, if you get a 'revise and resubmit' verdict, make it clear that you take the referees' points very seriously, both in the revised paper and in your cover letter. That means: if a referee makes a point that they clearly think important, don't respond to it by adding a footnote or making some similarly cosmetic adjustment. Third, don't be too surprised if you find that the part of your PhD thesis that you thought the most original actually turns out to be the most difficult part to publish. Original work provokes objections, and referees who find an idea strange seem to be more likely to reject it; whereas a diligent piece of work applying some well-established theory to some new issue in a fairly mechanical way can get nodded through unproblematically. And that's a shame, but I don't seem to be alone in having this impression, and it might be useful to bear it in mind when prioritizing the publication of one or another piece: the more original piece might take longer to find a home.

Finally, having interacted a great deal with Italian graduate students and colleagues, I have some advice for the many young researchers around the world who aren't native English speakers and are less than perfectly fluent in English: before submitting work to journals or publishers run by native English speakers, make sure that the English is not just comprehensible, but *perfect*. American and British academics do a lot of hand-wringing about their implicit biases in terms of race and gender, but much less about their implicit biases against foreigners whose first language isn't English. Avoid triggering that bias!

REFERENCES

- Anand, Paul, Graham Hunter, Ian Carter, Keith Dowding, Francesco Guala, and Martin van Hees. 2009. "The Development of Capability Indicators." *Journal of Human Development and Capabilities* 10 (1): 125-152.
- Bader, Ralf M. 2018. "Moralizing Liberty." In *Oxford Studies in Political Philosophy Volume 4*, edited by David Sobel, Peter Vallentyne, and Steven Wall, 141-166. New York, NY: Oxford University Press.
- Berlin, Isaiah. 1969. "Two Concepts of Liberty." In *Four Essays on Liberty*, by Isaiah Berlin, 121-154. Oxford: Clarendon Press.

- Carter, Ian. 1995. "Interpersonal Comparisons of Freedom." *Economics & Philosophy* 11 (1): 1-23.
- Carter, Ian. 1999. *A Measure of Freedom*. Oxford: Oxford University Press.
- Carter, Ian. 2004. "Choice, Freedom, and Freedom of Choice." *Social Choice and Welfare* 22 (1): 61-81.
- Carter, Ian. 2008. "How Are Power and Unfreedom Related?" In *Republicanism and Political Theory*, edited by Cecile Laborde, and John Maynor, 58-82. Oxford: Wiley-Blackwell.
- Carter, Ian. 2009. "Respect for Persons and the Interest in Freedom." In *Hillel Steiner and the Anatomy of Justice: Themes and Challenges*, edited by Stephen de Wijze, Matthew H. Kramer, and Ian Carter, 193-210. New York, NY: Routledge.
- Carter, Ian. 2011a. "Debate: The Myth of 'Merely Formal Freedom'." *Journal of Political Philosophy* 19 (4): 486-495.
- Carter, Ian. 2011b. "Respect and the Basis of Equality." *Ethics* 121 (3): 538-571.
- Carter, Ian. 2012. "Left-Libertarianism and the Resource Dividend." In *Alaska's Permanent Fund Dividend. Exploring the Basic Income Guarantee*, edited by Karl Widmerquist, and Michael W. Howard, 123-140. New York, NY: Palgrave Macmillan.
- Carter, Ian. 2013a. "Are Toleration and Respect Compatible?" *Journal of Applied Philosophy* 30 (3): 195-208.
- Carter, Ian. 2013b. "Social Power and Negative Freedom." In *Power, Voting and Voting Power: 30 Years After*, edited by Manfred J. Holler, and Hannu Nurmi, 27-62. Berlin: Springer.
- Carter, Ian. 2015. "Value-freeness and Value-neutrality in the Analysis of Political Concepts." In *Oxford Studies in Political Philosophy Volume 1*, edited by David Sobel, Peter Vallentyne, and Steven Wall, 279-306. New York, NY: Oxford University Press.
- Carter, Ian. 2019. "Self-ownership and the Importance of the Human Body." *Social Philosophy and Policy* 36 (2): 94-115.
- Carter, Ian, and Stefano Moroni. 2021. "Adaptive and Anti-adaptive Neighbourhoods: Investigating the Relationship between Individual Choice and Systemic Adaptability." *Environment and Planning B: Urban Analytics and City Science*.
- Carter, Ian, and Ronen Shnayderman. 2019. "The Impossibility of Freedom as Independence." *Political Studies Review* 17 (2): 136-146.
- Carter, Ian, and Hillel Steiner. Forthcoming. "Freedom Without Trimmings: The Perils of Trivalence." In *Without Trimmings. The Legal, Moral and Political Philosophy of Matthew Kramer*, edited by Mark McBride, and Visa A. J. Kurki. New York, NY: Oxford University Press.
- Cohen, Gerald A. 1988. *History, Labour, and Freedom; Themes from Marx*. Oxford: Oxford University Press.
- Cohen, Gerald A. 1995. *Self-ownership, Freedom, and Equality*. Cambridge: Cambridge University Press.
- Dworkin, Ronald. 2001. "Do Liberal Values Conflict?". In *The Legacy of Isaiah Berlin*, edited by Mark Lilla, Ronald Dworkin, and Robert Benjamin Silvers, 73-90. New York, NY: New York Review of Books.
- Dworkin, Ronald. 1979. "We Do Not Have a Right to Liberty." In *Liberty and the Rule of Law*, edited by Robert L. Cunningham, 167-181. College station, TX: Texas A&M University Press.

- Hees, Martin van. 2004. "Freedom of Choice and Diversity of Options: Some Difficulties." *Social Choice and Welfare* 22 (1): 253-266.
- Kramer, Matthew M. 2003. *The Quality of Freedom*. New York, NY: Oxford University Press.
- Kymlicka, Will. 1990. *Contemporary Political Philosophy: An Introduction*. Oxford: Oxford University Press.
- MacCallum, Gerald. C. 1967. "Negative and Positive Freedom." *The Philosophical Review* 76 (3): 312-334.
- Otsuka, Michael. 2003. *Libertarianism Without Inequality*. New York, NY: Oxford University Press.
- Pattanaik, Prasanta K., and Yongsheng Xu. 1990. "On Ranking Opportunity Sets in Terms of Freedom of Choice." *Recherches Économiques de Louvain / Louvain Economic Review* 56 (3/4): 383-390.
- Pettit, Philip. 1997. *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press.
- Sen, Amartya. 1988. "Freedom of Choice: Concept and Content". *European Economic Review* 32 (2-3): 269-294.
- Sen, Amartya. 1996. "Freedom, Capabilities and Public Action: A Response." *Notizie di Politeia* 12: 107-125.
- Skinner, Quentin. 2002. "A Third Concept of Liberty." *Proceedings of the British Academy* 117: 237-268.
- Steiner, Hillel. 1983. "How free: Computing Personal Liberty." *Royal Institute of Philosophy Supplements* 15: 73-89.
- Steiner, Hillel. 1994. *An Essay on Rights*. Oxford & Cambridge, MA: Blackwell.
- Sugden, Robert. 2003. "Opportunity as a Space for Individuality: Its Value and the Impossibility of Measuring It." *Ethics* 113 (4): 783-809.
- Vallentyne, Peter. 2002. "Equality, Brute Luck, and Initial Opportunities." *Ethics* 112: 529-57.
- Van Parijs, Philippe. 1997. *Real Freedom for All: What (if Anything) can Justify Capitalism?*. Oxford: Oxford University Press.
- Wollner, Gabriel. 2010. "Framing, Reciprocity and the Grounds of Egalitarian Justice." *Res Publica* 16: 281-298.
- Wollner, Gabriel. 2014. "Basic Equality and the Currency of Egalitarian Justice." In *Distributive Justice and Access to Advantage: G. A. Cohen's Egalitarianism*, edited by Alexander Kaufman, 180-204. Cambridge: Cambridge University Press.
- Zwolinski, Matt. 2011. "Bleeding Heart Libertarianism." *Bleeding Heart Libertarians* (blog). March 3, 2011. <http://bleedingheartlibertarians.com/2011/03/bleeding-heart-libertarianism/>.

Can One Both Contribute to and Benefit from Herd Immunity?

LUCIE WHITE

Utrecht University

Abstract: In a recent article, “Vaccine Refusal Is Not Free Riding”, Ethan Bradley and Mark Navin (2021) provide us with several reasons to doubt that vaccine refusal is a free rider problem. Here, I defend one connection between vaccine refusal and free riding and suggest that, when viewed in conjunction with their other arguments, this might constitute a reason to mandate Covid-19 vaccination.

Keywords: fairness, free riding, game theory, public good, public health, vaccination

JEL Classification: C72, I18

In their recent, illuminating contribution to the ethics and economics of Covid-19, Ethan Bradley and Mark Navin (2021) provide us with several reasons to doubt the received view that we can essentially view vaccine refusal as a free rider problem. Bradley and Navin contend that from both the subjective perspective of those who refuse vaccines, and also viewed objectively, there are several important differences between vaccine refusal and classic free riding. In making these distinctions, they draw attention to differences between the subjective views of many vaccine refusers and the views that we would expect to see among free riders, with important implications for how we should go about addressing the problem of vaccine refusal. However, their argument that vaccine refusers cannot be thought of as free riders in an objective sense—because it is not possible to both contribute to and benefit from the

AUTHOR’S NOTE: Many thanks to Philippe van Basshuysen, Donal Khosrowi, and Jan-Felix Müller for their—as always—incisive and insightful feedback. My sincere thanks also to an anonymous reviewer for their very constructive comments and, particularly, for pressing me on some essential distinctions. This research was made possible by the financial support of the Volkswagen Foundation.

public good of herd immunity¹—does not go through, particularly when it comes to Covid-19. Drawing this out can help us to better understand the various goals of a vaccination programme against Covid-19. Defending this particular parallel between vaccine refusers and free riders is also important because, in combination with the other arguments provided by Bradley and Navin, it says something about how we (in both a moral and a practical sense) should go about dealing with the problem of Covid-19 vaccine refusal.

THE ARGUMENT AGAINST VIEWING VACCINE REFUSAL AS FREE RIDING

The existence of ‘public goods’—that is, goods that are non-rivalrous and non-excludable—gives rise to the free rider problem. Because these goods are non-excludable, individuals may benefit from them whether or not they contribute to their provision. The fact that it is thus rational to benefit from a public good without contributing to it—that is, to be a ‘free rider’—coupled with the fact that if enough people refuse to contribute, the conditions for the existence of these goods are undermined, is the ‘free rider problem’. As Bradley and Navin (2021, 170) note, mass vaccination creates the public good of herd immunity. Although no country has yet passed the threshold for herd immunity, we can hope that current vaccination programmes will soon allow it to be achieved (at least in some places). Even where a fixed threshold has not been achieved, however, we might argue that vaccination still contributes to a public good, by slowing the spread (and resultant consequences of infection) to some degree—herd immunity, in other words, is not an all-or-nothing proposition (see Yates 2021).² The non-vaccinated benefit from herd immunity, because herd immunity makes outbreaks of the disease, and thus one’s chances of getting infected, more unlikely, even in the absence of individual protection.

¹ For good reason, Bradley and Navin (2021, 168n1) prefer the term ‘community protection’ to ‘herd immunity’. Although I agree with their reasons, here, I am using the more widely recognized term.

² In addition, even if we accept that there is no public good in existence until we reach a certain threshold for herd immunity, we might still posit that there is an obligation to bring this public good into existence. While this would not be an obligation to avoid free riding (which would seem to already require the existence of a public good that some individuals are unjustly benefitting from), some argue that a duty of *fairness* obligates us to contribute to the *creation* of the public good of herd immunity (see Navin 2013; Giubilini, Douglas, and Savulescu 2018). This would still suggest that individuals can be held accountable for unjustly refusing to contribute to the benefit of herd immunity, which should suffice to support this paper’s conclusion about how we should deal with the problem of vaccine refusal.

It would seem, then, to be a straightforward matter to conclude that individuals who enjoy the benefits of herd immunity without having participated in mass vaccination campaigns are free riders. Bradley and Navin, however, provide us with two strands of argument against this conclusion. First, they contend that vaccine refusers “do not have the subjective beliefs and attitudes of free riders” (2021, 171). In order to be free riders in this subjective sense, Bradley and Navin argue, vaccine refusers must acknowledge that they are indeed benefitting from the public good in question, and they must recognise that they are refusing to make some reasonable contribution towards this public good. Both of these attitudes, Bradley and Navin point out, are not characteristic of vaccine refusers. Vaccine refusers often both hold vastly overblown beliefs about the risks of vaccination (so that they do not view the costs of vaccination as a ‘reasonable’ contribution) and believe that there are no benefits to mass vaccination (thus denying that it produces a public good). It is important to point out the subjective differences between a ‘classic’ free rider and a ‘classic’ vaccine refuser because, as Bradley and Navin point out, it has implications for appropriate public policy responses. If—as would be the case with the classic free rider—an individual already believes that herd immunity is valuable and beneficial, and that the costs of contributing to this are not prohibitive, relatively minor changes to the individual’s incentive structure (in the form of either rewards or punishments) could lead them to view contribution to the public good as indeed in their best interests. Where individuals believe that the costs are extremely high, and no good will be produced as a result of their contribution, this strategy is not likely to yield the same results (we will return to the further significance of this presently).

The second strand of argument revolves around the objective criteria for free riding. Here, Bradley and Navin focus more on the moral obligation to contribute to a public good (rather than on what might motivate people to do so effectively). Where individuals are free riders, the authors note, they are refusing to contribute to something that they are benefitting from, and thus should *also* be contributing to.³ But, Bradley and Navin contend, it is not possible to both contribute to and benefit from the public good of herd immunity. One can contribute to herd immunity through possessing individual immunity (which one can gain either by being vaccinated, or by contracting and recovering from a disease). But once a person has individual immunity, they need not (indeed,

³ Bradley and Navin also provide another argument here—we will return to this below.

cannot) rely on herd immunity for protection. Because it is not possible to both contribute to and benefit from the public good of herd immunity, we cannot accuse vaccine refusers of behaving in an unjust manner by benefitting from something that they should also be contributing to.

THE BENEFITS OF CONTRIBUTING TO COVID-19 HERD IMMUNITY

This argument, however, oversimplifies the benefits that herd immunity, particularly against Covid-19, confers on each member of the community. One of the primary challenges of Covid-19, and the chief goals of public policy, has been to prevent hospital systems from becoming overwhelmed (Giubilini, Savulescu, and Wilkinson 2021; Johnson et al. 2020). The additional strain on healthcare systems (even when they are still functioning to a degree) during the pandemic has led to severe delays and disruptions in accessing needed medical care for unrelated conditions (see, for example, *The Lancet Rheumatology* 2021; Riera et al. 2021). Vaccination against Covid-19 does not confer any protection against contracting an unrelated illness and finding oneself unable to access medical care. Thus, in contributing to herd immunity by being vaccinated against Covid-19, each individual contributes to something that he also benefits from—a functioning healthcare system.

One might solve this problem by making the good at stake *excludable*—limiting healthcare access only to those who have contributed to the maintenance of the healthcare system by being vaccinated—but there are very good moral reasons not to exclude people from access to healthcare (see Feinberg 1986). If we treat access to healthcare as a non-excludable good, it generates a problem akin to the free rider problem—each individual benefits from its existence, and a widespread failure to contribute to its maintenance (by being vaccinated) will undermine the conditions for its existence. Although healthcare resources are, unlike public goods, *rivalrous* (too many individuals failing to vaccinate will lead to an overconsumption of limited healthcare resources, undermining the functioning of the system)—the essential parallel here remains. A functioning healthcare system is something that each individual can benefit from *and* contribute to, and there is thus plausibly an obligation to contribute to its maintenance through being vaccinated against Covid-19.

A second benefit that herd immunity confers on the community is an absence of the need for restrictions on the general population, which we have seen to varying degrees in many countries over the course of

the pandemic. These have included restrictions on the number of people who can meet in public or private, requirements to wear masks in certain spaces, the closure of or restrictions on the operation of businesses, the closure of schools and workplaces, and restrictions on international movement (Askitas, Tatsiramos, and Verheyden 2021). Some of these restrictions could be conditioned on vaccination status (that is, they are excludable)—being admitted to a foreign country, for example, or being able to eat at a restaurant, might be made contingent on showing proof of vaccination rather than restricted for all. But many of these restrictions—for example, most of those imposed in the UK until July 2021 (*BBC News* 2021)—were not made contingent on vaccination status despite high vaccination rates, perhaps due to the difficulty of checking the vaccination status of every unmasked person in a crowded area, or person in a group above a certain size. Where governments deem it necessary to impose general restrictions on the population in order to stop the spread of Covid-19, being vaccinated contributes to conditions that allow for the lifting of such restrictions, and this provides benefits for every member of the population.

A third way in which one can contribute to and benefit from a mass vaccination programme stems from the fact that high levels of vaccination reduce the probability of viral variants arising. If sustained transmission of Covid-19 is not contained, the likelihood of viral mutation increases. This can lead to vaccinations becoming less effective, and could even result in the emergence of a vaccine-resistant strain of the virus (Rubin 2021). In being vaccinated, therefore, you are contributing not just to herd immunity for the virus through your individual immunity; you are also contributing to the prevention of variants that you may not have individual protection against. In this way, one can both contribute to, and benefit from, the public good of herd immunity.

HOW SHOULD WE RESPOND TO VACCINE REFUSAL?

Drawing out the ways in which individuals can benefit from herd immunity to Covid-19, while contributing to this public good through being vaccinated, highlights the various and vital goals of Covid-19 mass vaccination programmes. The benefits of herd immunity through mass vaccination are not limited to the protection of the population against infection and the adverse side effects of Covid-19, but include access to a functioning healthcare system, a lack of ongoing restrictions, and protection against the emergence of future variants.

But preserving the argument that vaccine refusers are benefiting from something that they can and should also be contributing to also lends credence to the contention that vaccine refusers may be morally culpable for refusing to contribute to herd immunity through vaccination. If we think that vaccine refusers can be held responsible for their refusal to contribute despite the fact that they may hold false beliefs about vaccination (as do, for example, Brennan 2018 and Giubilini, Douglas, and Savulescu 2018), this might support incentivising or even compelling individuals to contribute to the goal of herd immunity (where the costs of doing so are reasonable—see Bradley and Navin 2021, 176). This is bolstered by another argument Bradley and Navin offer against viewing vaccination as a free rider problem in an objective sense—they claim that “free riding is individually rational, but vaccine refusal is not”. That is, they contend, because serious complications from vaccines are exceedingly rare, “it is almost always in a person’s interest to vaccinate, even when community protection makes their odds of infection very low” (2021, 173). In incentivizing or mandating vaccination, we would therefore not be imposing unreasonable burdens on the individual—in fact, each individual would be likely to benefit from this, beyond the benefits entailed by herd immunity.

Reintroducing Bradley and Navin’s arguments against viewing vaccine refusal as free riding in a subjective sense might further steer our sense of what could constitute an appropriate and effective policy response. To recap, Bradley and Navin suggest that because vaccine refusers often do not see mass vaccination as producing any benefit, and because they believe the individual costs of vaccination are very high, our typical response to classic free rider problems—introducing incentives to contribute to the agreed-upon public good—is not likely to work effectively here. This might be thought to point us, at least *prima facie*,⁴ in the direction of mandating, rather than incentivizing, vaccination where we have problems achieving or approaching the public good of herd immunity.

Scrutiny of Bradley and Navin’s arguments is thus a useful exercise in considering what constitutes effective and justifiable vaccination pol-

⁴ This is certainly not to say that this alone is sufficient to point us in this direction—several practical considerations may speak against such a policy. To take just one example, compulsory vaccination policies could lead vaccine refusers to avoid seeking medical care for themselves and their children (Flanigan 2014). For a comprehensive defense of vaccine compulsion, including sustained discussion of such practical considerations, see Flanigan (2014) and Giubilini (2020).

icy for Covid-19. I have argued, against Bradley and Navin, that the ethical argument that Covid-19 vaccine refusers are unjustly refusing to contribute to the benefit of herd immunity retains its force once we take a broader view of the resultant benefits. Vaccine refusers might thus be morally culpable for failing to contribute to the various significant benefits that herd immunity to Covid-19 provides. Coupled with Bradley and Navin's compelling arguments that incentivizing vaccination may be of limited use, and that vaccine refusal is rarely in the best interests of the individual, this could be viewed as lending support to the case for Covid-19 vaccine mandates where the public good of herd immunity cannot be achieved through other means.

REFERENCES

- Askitas, Nikolaos, Konstantinos Tatsiramos, and Bertrand Verheyden. 2021. "Estimating Worldwide Effects of Non-Pharmaceutical Interventions on COVID-19 Incidence and Population Mobility Patterns Using a Multiple-Event Study." *Scientific Reports* 11: 1972.
- BBC News. 2021. "Covid Rules: What Has Changed?" *BBC News*. Accessed 27 July, 2021. <https://www.bbc.com/news/explainers-52530518>.
- Bradley, Ethan, and Mark Navin. 2021. "Vaccine Refusal Is Not Free Riding." *Erasmus Journal of Philosophy and Economics* 14 (1): 167–181.
- Brennan, Jason. 2018. "A Libertarian Case for Mandatory Vaccination." *Journal of Medical Ethics* 44 (1): 37–43.
- Feinberg, Joel. 1986. *The Moral Limits of the Criminal Law. Volume 3: Harm To Self*. New York, NY: Oxford University Press.
- Flanigan, Jessica. 2014. "A Defense of Compulsory Vaccination." *HEC Forum* 26 (1): 5–25.
- Giubilini, Alberto. 2020. "An Argument for Compulsory Vaccination: The Taxation Analogy." *Journal of Applied Philosophy* 37 (3): 446–466.
- Giubilini, Alberto, Thomas Douglas, and Julian Savulescu. 2018. "The Moral Obligation to Be Vaccinated: Utilitarianism, Contractualism, and Collective Easy Rescue." *Medicine, Health Care and Philosophy* 21 (4): 547–560.
- Giubilini, Alberto, Julian Savulescu, and Dominic Wilkinson. 2021. "Queue Questions: Ethics of COVID-19 Vaccine Prioritization." *Bioethics* 35 (4): 348–355.
- Johnson, Helen C., Céline M. Gossner, Edoardo Colzani, John Kinsman, Leonidas Alexakis, Julien Beauté, Andrea Würz, Svetla Tsoleva, Nick Bundle, and Karl Ekdahl. 2020. "Potential Scenarios for the Progression of a COVID-19 Epidemic in the European Union and the European Economic Area, March 2020." *Eurosurveillance* 25 (9): 8–12.
- Navin, Mark. 2013. "Resisting Moral Permissiveness About Vaccine Refusal." *Public Affairs Quarterly* 27 (1): 69–85.
- Riera, Rachel, Ângela Maria Bagattini, Rafael Leite Pacheco, Daniela Vianna Pachito, Felipe Roitberg, and Andre Ilbawi. 2021. "Delays and Disruptions in Cancer Health

- Care Due to COVID-19 Pandemic: Systematic Review." *JCO Global Oncology* 7: 311–323.
- Rubin, Rita. 2021. "COVID-19 Vaccines vs Variants—Determining How Much Immunity Is Enough." *JAMA* 325 (13): 1241–1243.
- The Lancet Rheumatology*. 2021. "Editorial—Too Long to Wait: The Impact of COVID-19 on Elective Surgery." *The Lancet Rheumatology* 3 (2): E83.
- Yates, Kit. 2021. "(How) Can We Reach Herd Immunity." *The BMJ Opinion*, July 16, 2021. <https://blogs.bmj.com/bmj/2021/07/16/kit-yates-how-can-we-reach-herd-immunity/>.

Lucie White is an assistant professor at Utrecht University, specializing in bioethics. She is currently working on a couple of projects which focus, in various ways, on policy responses to Covid-19. Contact e-mail: <l.a.white@uu.nl>

Vaccine Refusal Is Still Not Free Riding: A Reply

ETHAN BRADLEY

Oakland University

MARK NAVIN

Oakland University

Abstract: In a recent article, “Can One Both Contribute to and Benefit from Herd Immunity?”, Lucie White (2021) argues that vaccine refusal is more like free riding than we have claimed that it is. Here, we critically reply to White’s arguments.

Keywords: fairness, free riding, game theory, public health, vaccination

JEL Classification: C72, I18

In a recent paper in this journal (Bradley and Navin 2021), we argued that it is generally inaccurate to claim that vaccine refusers free ride on community protection (sometimes called ‘herd immunity’).¹ We agree that vaccine refusal is often unethical—because it risks harming others and shirks a responsibility to contribute to just institutions—but we think it is not usually an instance of free riding. First, vaccine refusers are not usually *objectively* free riders. By definition, free riders rationally promote their interests by refusing to pay costs associated with supporting a public good. But vaccination is almost always good for people, such that refusal does not promote a refuser’s interests. That is, free riding is individually rational, while vaccine refusal is not. So, the problem of vaccine refusal is not about self-interest, but it is about a refuser’s false beliefs or idiosyncratic values. Also, by definition, a free rider could continue to benefit from the public good they currently enjoy if they contributed to that good. In contrast, someone who decides to get vaccinated will no longer benefit from herd immunity because they will

¹ See Bradley and Navin (2021) for references.

rely on their individual immunity for protection against disease. So, vaccine refusers could become *contributors* to herd immunity by getting vaccinated, but they would not continue to be *beneficiaries* of herd immunity. Second, vaccine refusers do not usually hold the *subjective* beliefs one might expect of a free rider. Free riders acknowledge that they are benefitting from a public good, but they refuse to incur moderate costs to contribute to it. Vaccine refusers, however, often believe that vaccination causes serious harms and that herd immunity does not exist, or at least that it is not very valuable.

Lucie White (2021) has published a response to our article. White argues, contrary to our view, that it is possible to both contribute to herd immunity and benefit from it. White's argument hinges on the identification of three secondary benefits of herd immunity: (1) a functioning healthcare system, (2) a lack of pandemic social restrictions, and (3) a reduced likelihood that new variants will arise for which vaccines may not provide good protection. In each case, someone who is vaccinated can *contribute* to these benefits—that is, they can contribute to herd immunity, which, in turn, contributes to these further goods. Vaccinated people can also *enjoy benefits* from functioning healthcare systems, from social freedom, and from the maintenance of current vaccines' protection against serious disease. In contrast, the vaccine refuser enjoys the benefits of these goods without contributing to them. Therefore, in light of the existence of these secondary goods, White concludes that there is a tighter connection between free riding and vaccine refusal than we have claimed.

We are grateful for White's generous and thoughtful engagement with our work. We agree with her that herd immunity provides both the *immediate* good of community protection from outbreaks and that it contributes to various *secondary* goods, including those White identifies. Indeed, herd immunity can also contribute to economic growth, education, and social trust and stability, among many other goods. White is surely correct to note that vaccinated people both contribute to herd immunity and can enjoy some of these kinds of secondary benefits.

While we largely agree with White's arguments, we think they have less upshot for our position than they may appear to. It may help to recall how appeals to free riding are supposed to illuminate discussions about vaccine refusal. First, if vaccine refusal were free riding, then we could *explain* the problem of vaccine refusal in terms of a conflict between collective rationality (the maintenance of herd immunity) and in-

dividual rationality (refusal to contribute to herd immunity).² And we could resolve this conflict by shifting incentives so as to make it individually rational to vaccinate. But, if, as we have argued, it is already individually rational to vaccinate, even in the absence of further incentives or sanctions, then ‘free riding’ is an inaccurate description of the causes of vaccine refusal. Second, if vaccine refusal were free riding, then we could appeal to moral arguments against free riding to *criticize* vaccine refusal. Free riding is sometimes thought to be wrongfully selfish and to demonstrate insufficient commitment to reciprocity. But, if vaccination promotes a vaccinated person’s interests, then it is not selfish to refuse vaccines. And, if, as we have again argued, someone can either contribute to herd immunity *or* benefit from it, *but not both*, then the failure to contribute to herd immunity is not a failure of reciprocity. It may instead be a failure of beneficence or of some other duty to support the creation of institutions that provide *benefits to others*. But it is not a moral failure grounded in a practice of free riding.

White’s critical commentary focuses on our point about reciprocity. If vaccination contributes to some public goods that *even vaccinated people* can benefit from, then perhaps we should characterize vaccine refusal as a kind of free riding, and perhaps we should morally condemn it for failing to demonstrate sufficient commitment to reciprocity. We are not so sure.

Many of the indirect benefits of herd immunity do not appear to be *public goods*. And, inasmuch as they are *not* public, that fact undermines the claim that vaccine refusers free ride on these goods. White notes that mass vaccination against Covid-19 reduces burdens on healthcare providers and increases access to functioning healthcare systems. But this is an *excludable* good, as unvaccinated people *can* be denied access to healthcare to preserve system capacity. (We agree with White that this kind of exclusion would be unethical.) White also notes that mass vaccination can prevent pandemic-related restrictions on movement and gatherings. This good is also often excludable, for example, via vaccine passports, though White is surely correct that some such exclusion efforts—such as checking the vaccination status of every unmasked individual in a country—would be infeasible. White further argues that mass vaccination reduces the likelihood of new SARS-CoV-2 variants. The

² We note, however, that vaccination does not guarantee individual immunity. This complicates the relationship between vaccination and herd immunity, which is why the focus should not be on vaccination per se but on the immunity that vaccination typically provides.

primary beneficiaries of a decreased likelihood of new variants are people who are already vaccinated, because new variants are more likely to result in breakthrough infections and therefore undermine the protection that vaccines grant. This good is therefore excludable because anyone who is not vaccinated will not enjoy it—unvaccinated people remain vulnerable to all variants of SARS-CoV-2. Vaccine refusers cannot free ride on a good they cannot enjoy. However, if new variants are more transmissible or more aggressive than existing variants, then unvaccinated people may also benefit from a decreased likelihood of new variants. In contrast, if new variants are less severe (as, for example, the Omicron variant appears to be), then unvaccinated people may actually benefit from the emergence of new variants, such that the prevention of new variants is not a good that they enjoy.³

The three secondary goods that White discusses are all *excludable*, which means they are not public goods, and that vaccine refusers do not *free ride* on them. However, as White notes (and as we have agreed), there are ethical and pragmatic limits to *excluding* vaccine refusers from some of these goods (for example, preventing unvaccinated persons from accessing healthcare). Therefore, some of these goods may function as public goods in societies whose institutional limitations or ethical commitments prevent them from excluding vaccine refusers. Accordingly, it may seem as if vaccine refusers may be free riders on these goods.

However, there may be other ways to *contribute* to the goods that herd immunity indirectly promotes, such that vaccine refusers may not be free riders even if those goods are *public*. Unless a non-immune person emigrates or otherwise removes themselves from society, there is only *one way* they can contribute to herd immunity: acquire individual immunity. Accordingly, someone who does not cultivate individual immunity does not contribute to herd immunity. But there are often *many ways* to contribute to the various goods that herd immunity indirectly promotes. For example, practices of social distancing, correctly using effective masks, and regular testing can reduce outbreaks and therefore reduce strains on the healthcare system, promote social freedom, and reduce the likelihood of new variants. We acknowledge that vaccination is likely to be the most efficient and effective way to contribute to these goods. But the fact that there are other ways to contribute means that someone who has refused vaccination has not thereby made use of a

³ We are grateful to an editor of this journal for suggesting this point.

public good without contributing to it. This person may be a ‘cheap rider’, if they did *less* to contribute to these goods than did vaccinated people, but they would not be a free rider if they had made some meaningful contribution.

In practice, we suspect that many vaccine refusers *also* refuse to make alternative contributions to the secondary goods that herd immunity promotes. Accordingly, it may be accurate to say that they free ride on those goods and that they have failed to do their fair share to support them. But that conclusion rests on *empirical* claims about background conditions and the behavior of vaccine refusers. When vaccine refusal is a kind of free riding, it is so only in this contingent and indirect way.

REFERENCES

- Bradley, Ethan, and Mark Navin. 2021. “Vaccine Refusal Is Not Free Riding.” *Erasmus Journal of Philosophy and Economics* 14 (1): 167–181.
- White, Lucie. 2021. “Can One Both Contribute to and Benefit from Herd Immunity?” *Erasmus Journal of Philosophy and Economics* 14 (2): 157–164.

Ethan Bradley is an undergraduate student of philosophy and political science at Oakland University (Rochester, MI, USA).
Contact e-mail: <embradley@oakland.edu>

Mark Navin is professor and chair of philosophy at Oakland University (Rochester, MI, USA).
Contact e-mail: <navin@oakland.edu>

The Different Facets of Injustice: A Critique of Nancy Folbre's 'Manifold Exploitations'

VIVEK CHIBBER

New York University

ROBERTO VENEZIANI

Queen Mary University of London

Abstract: In her recent work, Nancy Folbre (2020, 2021) undertakes an ambitious effort: constructing an intersectional political economy that aims to identify the common mechanisms and logic underpinning the many wrongs that characterise capitalism. In this paper, we focus on what we deem the three fundamental theoretical pillars of her approach. First, she challenges the oppression/exploitation distinction within Marxian political economy and proposes a broader definition of exploitation that can take manifold forms. Second, she questions the Marxian concept of class, and emphasises the variety of forms of subordination and exploitation related to social identities that cannot be reduced to Marxian classes. Finally, she advocates a more comprehensive notion of the economy beyond a focus on capitalist relations of production. It is difficult to understate the theoretical relevance of these claims, which highlight the importance of various contemporary forms of injustice—of which the exploitation of workers by capitalists is only one. As a *complement* to the Marxian theory of exploitation and class, Folbre's approach would broaden our understanding of oppressive social relations. Yet as an *alternative* to Marxian political economy, it is ultimately unconvincing: a shift of emphasis to 'manifold exploitations', social groups, and the economy does not yield a gain in analytical insight but rather an impoverishment of our conceptual toolbox. The struggle against capitalism is different from the struggle against patriarchy and racism, even if the ultimate aim should be the removal of all structures of oppression and domination.

Keywords: exploitation, oppression, injustice, intersectionality

JEL Classification: D63, B51, B54

Nancy Folbre's latest book (Folbre 2021) is the culmination of a decades-long reflection on feminist and radical themes straddling disciplinary boundaries. It is wonderfully written and passionate in its defence of a political economy firmly on the side of the oppressed. Its fundamental aim is, politically, to foster dialogue among different groups that have reason to rebel against the status quo and, theoretically, to develop a more inclusive, *intersectional* political economy to diagnose the many wrongs that characterise modern economies.

It is impossible to summarise and discuss a dense book which deals with some of the most important topics, and thinkers, in economics, political philosophy, and sociology in the brief space of a commentary article. While acknowledging the many merits of Folbre's theoretical *tour de force*, in this paper, we shall focus on one aspect of her work that we deem problematic, namely the claim that her approach builds on, but goes beyond the Marxian tradition: it generalises Marxian political economy while preserving its fundamental insights. This is a fundamental—foundational even—element of Folbre's theory and a constant theme throughout the book, which she has further elaborated upon elsewhere (Folbre 2020).

Three main theoretical pillars underpin Folbre's approach. First, she “challenges the oppression/exploitation binary within Marxian political economy, proposing a broader definition of exploitation that can take manifold forms” (Folbre 2020, 452). Standard, labour-based definitions of exploitation are, in her view, reductive as they do not capture a range of relations of subordination and oppression that characterise modern economies.¹ Second, this limitation calls into question the Marxian concept of class, as the variety of forms of oppression, subordination, and exploitation that exist in advanced economies are related to social identities that, argues Folbre, cannot be captured by standard Marxian class concepts. Building on decades of feminist work on care and unpaid labour, she argues that social conflicts cannot be reduced to class struggle, as they are

¹ A definition of exploitation is *labour-based* if labour time, or (skill-adjusted) effective labour, is used as the relevant exploitation numéraire: labour—rather than wealth, income, utility, and so on—is deemed the key normative variable of interest and the main unit of account of exploitation theory. One way to derive individual labour accounts is by means of the labour theory of value, but this is by no means the *only* way. In exploitation theory, labour accounting is simply the “way of characterizing what it is that people give one another [...] (where ‘give’ is understood very broadly to refer to any way in which some person undergoes a loss that ends up a gain to another)” (Reiman 1987, 9). More on this in section 2 below. For a comprehensive discussion, see Veneziani and Yoshihara (2018).

“more consistent with the intersectional logic of contradictory group interests” (Folbre 2020, 453). Finally, she suggests that the Marxian emphasis on class and labour-based notions of exploitation derives from a narrow view of the economy and “the assumption that capitalism is a hegemonic mode of production that constitutes the ‘economy’ or even the entire ‘world system’” (Folbre 2020, 455).

It is difficult to understate the theoretical relevance of these claims, which highlight the importance of various contemporary forms of injustice—of which the exploitation of workers by capitalists is only one. If one wants to understand the many facets of *inequality of personal* income (and other indicators of personal well-being), for example, then one has to look also at gender, race, and even citizenship. Economic and social disadvantages are the combined effect of multiple forms of subordination, including class. Further, individual motives, aims and beliefs, and therefore collective action, are influenced by various aspects of social relations, including class, race, gender, religion, and even citizenship. Finally, economic determinism suffers from major limitations, and the elimination of capitalist relations of production does not imply—as a matter of either logic or historical necessity—the end of racism, homophobia, and patriarchy.

The general appeal to a multifaceted approach to the complexity of social relations and economic structures is undoubtedly important, and Folbre’s analysis compels us to broaden our normative horizon. Nonetheless, in this paper we raise some doubts on her three key theoretical claims concerning Marxian political economy. If Folbre’s contribution was meant to highlight some social phenomena that are outside the focus of exploitation theory, then it *would* help to provide a more complex, and nuanced understanding of oppressive social relations in advanced economies. Yet Folbre’s aim is to provide an *alternative* to Marxian class and exploitation theory, which generalises it while preserving its fundamental insights. We argue that instead of building solid foundations for a more general theory of manifold oppressions and injustices in advanced economies, the proposed approach yields a loss of analytical power and conceptual clarity.

It is worth clarifying at the outset that we are not advocating analytical rigour for its own sake. As feminist and radical scholars have repeatedly pointed out, the focus on rigour and formalism tends, in practice, to select the contributions by dominant social groups as the only ones that deserve to be read or heard. Our point, rather, is that analytical precision

is required in order to clearly diagnose a form of injustice, or oppression; to explain its determinants; to identify its normative implications; and to propose appropriate remedies. A critical and emancipatory political economy—and a non-ephemeral coalition of the oppressed—can only be built around a clear and precise conceptual apparatus that allows one to identify *both* the similarities *and* the differences among the various forms of injustice that plague capitalist economies.

MARXIAN EXPLOITATION AND MANIFOLD EXPLOITATIONS

Although there are many definitions of exploitation in the Marxian tradition, they all share certain features and in this paper we ignore the differences among them. For our purposes, we illustrate the Marxian approach by focusing on the definition proposed by Erik Olin Wright (2000), which is also the focus of Folbre's critical assessment (Folbre 2020, 464–466; 2021, 67, 124–126).

According to Wright (2000), there is exploitation when three criteria are satisfied:

1. *The inverse interdependent welfare principle.*—The material welfare of exploiters causally depends upon the reductions of material welfare of the exploited.
2. *The exclusion principle.*—This inverse interdependence of the welfare of exploiters and the exploited depends upon the exclusion of the exploited from access to certain productive resources.
3. *The appropriation principle.*—Exclusion generates material advantage to exploiters because it enables them to appropriate the labor effort of the exploited. (Wright 2000, 1563)

This is a general definition that holds for various modes of production (slavery, feudalism, capitalism). Consider capitalism: workers are exploited by capitalists because they contribute more labour in productive activities than they are paid for. Part of this unpaid labour is appropriated by capitalists—who therefore improve their lot at the expense of the material welfare of workers—thanks to their ownership of scarce productive assets.

Three points should be noted about the Marxian definition of exploitation: first, labour is the main currency of exploitative relations. Second, production activities and productive relations are at the core of the definition and delineate the scope of the concept. Third, the Marxian theory

of exploitation identifies a specific mechanism which allows exploiters to appropriate labour at the expense of the exploited: property relations and the unequal ownership of certain scarce productive assets. As Wright states:

Exploitation is thus a diagnosis of 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 exploiters, by virtue of their exclusionary rights and powers over resources, are able to appropriate labor effort of the exploited. (Wright 2000, 1563)

Folbre suggests that the standard Marxian approach, as exemplified by Wright's definition, suffers from significant shortcomings and advocates the adoption of a broader definition. First, by focusing on productive activities, the Marxian approach ignores a variety of wrongs that happen outside the sphere of production narrowly conceived. Second, and related, it is 'misleading' (Folbre 2020, 458), because it emphasises mechanisms of subordination that arise *only* from relations of production and ownership of productive assets, and that are related to class positions. Third, the Marxian approach puts an excessive emphasis on labour, whose exchange distinguishes exploitation from other wrongs.

In relation to the above definition, for example, Wright argues that if the first two conditions "are present, but not the third, what might be termed nonexploitative economic oppression may exist, but not exploitation" (2000, 1564). The welfare of the oppressor (unlike that of the exploiter) depends simply on the exclusion of the oppressed from access to certain resources, but not on their effort. According to Folbre, this distinction is objectionable. For:

The mutual dependency of groups—their voluntary or coerced cooperation—need not take the form of direct control of labor; it can take more indirect forms such as contributions to public goods or enforcement of the rule of law, which also create interdependence. Many forms of implicitly coerced cooperation lead to an unfair distribution of the resulting gains—forms of exploitation that are not directly based on labor exchange. (Folbre 2020, 464)²

² Folbre argues that her definition blurs the distinction between oppression and exploitation. Yet "Wright's distinction between exploitation and oppression remains meaningful, and discrimination can affect both. Not all economic interactions can be reduced to bargaining, be it fair or unfair" (Folbre 2021, 124). So it would seem that it is more a matter of drawing the line someplace else.

Before examining Folbre's proposed generalisation, two preliminary points should be made. First, the focus on labour is much less narrow than Folbre suggests. As Wright notes:

"Appropriation of labor effort" can take many forms. Typically this involves appropriating the products of that labor effort, but it may involve a direct appropriation of labor services. The claim that labor effort is appropriated does not depend upon the thesis of the labor theory of value [...]. All that is claimed is that when capitalists appropriate products they appropriate the laboring effort of the people who make those products. (Wright 2000, 1563–1564n4)

More generally, the theory of exploitation as the unequal exchange of labour conceives of labour as the exploitation numéraire: the ethically based unit of account that measures the inequalities associated with exploitative relations (Roemer 1982; Veneziani and Yoshihara 2018). Second, it is indeed true that per se the enforcement of the rule of law, or the "willingness to help others" (Folbre 2021, 124) do not fall within the purview of Marxian exploitation theory. But it is also rather unclear that they can be properly described as instances of cooperation characterised by the *mutual* dependency of groups.

Nonetheless, it is certainly correct that the Marxian approach is limited in scope: it does not capture all forms of subordination, let alone all types of injustices; and it does focus on productive relations and on labour as the exploitation numéraire. Thus, gender relations do not fall within the purview of the above definition: women are oppressed, not exploited, because they provide services that are outside the capitalist mode of production yet are necessary for the reproduction of labour power as well as social reproduction. It is possible to appreciate the normative relevance, and the deep economic implications of gender oppression, even if it is distinguished from Marxian exploitation.

According to Folbre, exploitation should instead be defined based on "an analysis of institutional structures of collective power that shape processes of cooperation and conflict that reach beyond capitalist dynamics" (Folbre 2020, 452). Although she does not provide a precise definition, she suggests that the unequal distribution of gains from cooperation could be termed 'exploitation' "if [it] is unfair [...]. Alternatively (or in addition), the process by which the distribution was achieved may be deemed unfair" (Folbre 2020, 461).

Does this alternative approach improve on the Marxian definition? This is difficult to say, given that, lacking a precise notion of fairness, it remains underspecified. Yet, Folbre seems to emphasise the procedural aspects of exploitative relations and in particular the lack of consent, participation, and democratic deliberation. She claims that “a precise definition of ‘fairness’ therefore, may be less important than consideration of the obstacles to the development of a social environment in which truly democratic deliberations can take place” (Folbre 2020, 463).³

It is unclear whether this alternative approach could detect, and condemn, quintessentially exploitative relations: if participation and consent are required at the time certain institutions are established,⁴ then one can imagine a capitalism with a clean (democratically established) origin such that capitalist/worker relations would be deemed nonexploitative.⁵ A similar verdict would be rendered even if one focused on current institutions, since capitalism often coexists with democratic institutions. Conversely, the approach is liable to yield false positives: the relation between prison guards and prisoners, for example, may be deemed exploitative, as the mutual dependency is definitionally not supported by institutions that enjoy the active, democratic participation and support of the agents involved.⁶

At a broader level, Folbre’s discussion points to the existence of a trade-off between generality and theoretical cogency. In her attempt to extend exploitation theory to include all forms of subordination, she blurs a number of theoretically relevant distinctions—for instance, be-

³ She implicitly adopts a slightly different (and slightly more specific) approach elsewhere, when she notes that “unfair bargaining power is a form of value extraction that encompasses what Marx described as exploitation” (Folbre 2020, 464; see also Folbre 2021, 124). Yet, this approach is also problematic since it is unclear how ‘unfair bargaining power’ can be per se a form of value extraction: unequal bargaining power (whether fair or unfair) is neither necessary nor sufficient for value extraction. At best it *enables* value extraction.

⁴ This seems to be Folbre’s interpretation since she emphasises “gain-seeking behavior shaped by social institutions *established by profoundly undemocratic means*” (Folbre 2020, 464: *emphasis added*; see also Folbre 2021, 125).

⁵ Quite aside from its secondary role in exploitation theory, the normative relevance of the emphasis on the historical origin of certain institutions is rather unclear: in the context of a patterned approach to distributive justice such as the one advocated by Folbre, what matters is the *current* structure of institutions, and their *current* effect on bargaining power and distribution, not their origin.

⁶ To be sure, we are not suggesting that the correctional system in advanced capitalist societies is fundamentally just. Rather, our point is that it may be condemned as an oppressive institution without having to say, implausibly, that guards *exploit* prisoners even when no productive relations exist and no exchange of labour takes place.

tween oppression and exploitation; between different forms of exploitation; and even, to some extent, between exploitation and other forms of injustice. This may help rally the oppressed under a single banner, but it is unclear that much analytical insight is gained by saying that the relation between capitalists and workers is the same as that between husband and wife, or heterosexuals and homosexuals. It is certainly worthwhile having a conceptual framework that identifies all types of unjust social relations and the many instances of oppression in advanced economies. By bundling all these phenomena together, however, one loses sight of the fundamental differences between them, and therefore ultimately loses the ability to explain any of them.

Theft and blackmail share one important feature: they are both criminal offences. If one aims to depict the overall level of criminality in a given society, then it is perhaps harmless to bundle them together. Indeed, recognising that both are ‘crimes’ may add an important layer of explanation in that context. And surely the aim should be the elimination of *all* criminal offences. Yet in general, to insist that the distinction between the two be blurred and that they be called ‘crimes’ is not to gain analytical insight. It means losing two concepts.⁷

CLASSES AND SOCIAL GROUPS

Similar issues arise concerning the proposed generalisation of the Marxian notion of class. Folbre rightly rejects “the view that most social conflicts derive from class conflicts, or from capitalist strategies to ‘divide and conquer’” (Folbre 2020, 454). Only the crudest form of economic determinism may support the idea that property rights and the relations of production explain every social conflict in advanced economies. Similarly,

intra-class economic inequalities cannot be explained as a consequence of heterogeneous labor in capitalist wage relations, because heterogeneity itself requires explanation: why do some workers attain more advantageous skills, assets and preferences than others do? (Folbre 2020, 454)

⁷ If Folbre’s theory was conceived of as complementary to Marxian exploitation theory, focusing in particular on certain oppressive social relations that the Marxian definition ignores, then it *would* provide significant conceptual clarity. Yet Folbre’s aim is not to provide an *additional* definition of exploitation to the Marxian one, but rather an *alternative* to it, which generalises it while preserving its fundamental insights. We are thankful to the Editor of this journal for pressing us on this point.

And yet, it is unclear what is gained by generalising “the Marxian analysis of class to all socially assigned groups that share at least some common identities and interests” (Folbre 2020, 452).

A first problem is the loss of conceptual clarity at the highest level of abstraction, namely at the definitional level. Folbre argues that:

As Marxist scholars have long recognized, it is difficult to reach consensus on the operational meaning of class. It is equally, if not more difficult to reach consensus on the definition of other aspects of socially assigned (as distinct from individually chosen) group membership. (Folbre 2020, 457).

Indeed. But while the Marxian concept of class may be difficult to *operationalise*, the notion of socially assigned groups is fuzzy and vague even *at the conceptual level*. How is a socially assigned group actually defined? Is it meaningful to extend the Marxian analysis of class to *all* socially assigned groups that share *at least some* common identities and interests? Taken literally, that would imply extending Marxian class theory to football supporters, fan groups, readers’ clubs, professional associations, and so on. And yet, it is unclear, in Folbre’s theorisation, what are the common identities and interests that underpin a theoretically relevant concept of socially assigned group which generalises the Marxian notion of class.

Perhaps one objective, distinguishing feature of social groups identified by Folbre’s general approach is their position in the social structure: socially advantaged vs. socially disadvantaged groups. And yet this immediately invites the question: How does one define social (dis)advantage in a theoretically relevant way? Clearly, not all social (dis)advantages are salient, and certainly not all of them are akin to the (dis)advantages associated with Marxian classes.

Folbre seems to suggest a theoretical focus on social disadvantage that is associated with, or conducive to economic disadvantage. She notes that “all else equal, memberships in socially disadvantaged groups contribute to individual economic disadvantage” (Folbre 2021, 123). This statement is intuitively plausible, but it can meaningfully identify a causal mechanism only if the notion of *social* disadvantage is defined independently of *economic* disadvantage. This raises two main issues.

First, if socially assigned groups are to be identified based *both* on economic disadvantage *and* on a conceptually distinct social disadvantage, then it is unclear that Folbre’s approach represents, logically speaking, a generalisation of Marxian class theory in which class status is

defined entirely within the economic sphere. Second, and perhaps more important, it is not obvious how to define a theoretically salient concept of social disadvantage that is completely independent of economic disadvantage, especially if economic advantage is conceived of in a broad sense, as Folbre does. In any case the notion of social disadvantage is left largely undefined by Folbre.

More generally, one wonders how far a generic emphasis on disenfranchised agents, or socially disadvantaged groups takes us in the understanding of modern economies. “The institutional economist William Dugger has described diverse inequalities in terms of ‘top dogs’ and ‘underdogs’” (Folbre 2020, 453). But what is the analytical purchase of these categories?

In contrast, the Marxian notion of class does not explain all forms of subordination: it only partially contributes to explain patriarchal institutions and racial tensions. Yet, as complex to operationalise as it may be, it is conceptually crucial in order to understand the logic of capitalism, and its central conflict.⁸ It may be limited in scope, but this is more than compensated by the gain in analytical power.

That “capitalists, men and whites became codependent beneficiaries of the exploitation of disempowered groups” (Folbre 2021, 76) is surely true, but not particularly helpful in understanding the specific mechanisms underlying various forms of domination and oppression. To continue with our legal metaphor, it is surely true that thieves and blackmailers are both criminals. But they commit different crimes with different consequences and have different motivations. By bundling them together we do not fully understand either. Different institutions shape different group identities, and the relevant social cleavages follow different logics and are determined by different mechanisms. So, both statistically and historically, black working-class women tend to suffer from multiple oppressions. And yet, a black female owner of the means of production, who only earns profit (or dividend) income, remains a capitalist and, as a capitalist, will follow the logic of capital.

The limitations of this quest for generality are particularly evident in the conceptualisation of collective action and group conflicts. Folbre (2021, 115ff) develops a stylised model of bargaining over the gains from cooperation in which distributive outcomes are largely determined by

⁸ This does not mean that class conflict is the only or the most important conflict in advanced economies. It only means that it is conceptually the central contradiction of the capitalist mode of production.

fallback positions. According to Folbre, her “approach to collective conflict subsumes both the neoclassically-influenced concept of rent-seeking and the Marxian theory of surplus extraction under a larger rubric of ‘gain-seeking’” (2021, 115). This is not entirely correct: the emphasis on rent-seeking is characteristic of but one sub-field in neoclassical economics. Her model *is* (a version of) the general neoclassical model of bargaining: the idea of bargaining over the gains from cooperation (what she terms ‘gain seeking’) and the emphasis on the role of fallback positions is entirely consistent with neoclassical game-theoretic models of bargaining going back at least to John Nash (1953).

There is nothing wrong in adopting the neoclassical bargaining model, which provides useful insights on a number of economic conflicts. But the adoption of a theoretical framework is not neutral with respect to the kind of questions that one may ask, and the conceptual tools consistent with that framework. For example, a generic emphasis on groups and bargaining misses the specific aspects of class conflict which explain capitalist dynamics. “The notion that social institutions have intersectional effects on the bargaining power of entire groups of people builds on feminist models of bargaining between husbands and wives in married households” (Folbre 2021, 127). This may be true. But what analytical purchase does this allegedly more general framework gain us to understand different phenomena, such as race, or class?

CAPITALISM AND THE ECONOMY

In the previous sections, we have argued that there is a trade-off between an abstract appeal to generality and analytical precision. The Marxian concepts of exploitation and class may be inadequate to explain all forms of oppression, all social conflicts, and all group identities and cleavages that characterise modern economies. But this is not a shortcoming: they are not meant to subsume everything that is relevant in capitalism. They identify specific, but fundamental, phenomena that help us understand certain important aspects of social reality.

According to Folbre (2020, 455; 2021, 257), however, the narrow focus of Marxian theories on such concepts is explained by an even more fundamental theoretical problem, namely a reductionist view that equates the economy with capitalism. She proposes “an expanded definition of ‘the economy’ that extends beyond commodity production, which in turn points toward a definition of exploitation that facilitates attention to its complex interactive forms” (Folbre 2020, 452). This is the key theoretical

move, and the fundamental underpinning of Folbre's analysis: it is the shift in theoretical focus from capitalism to 'the economy' that justifies the move away from the Marxian concepts of exploitation and class.

In the previous sections, we have defended the specific analytical relevance of such concepts in the analysis of capitalism. The point is not to identify a hierarchy of wrongs based on their badness. Capitalist exploitation is not definitionally worse than slavery, and patriarchal institutions are not definitionally worse than colonialism. The point, rather, is to identify the right conceptual tools for the problem at hand. The Marxian concepts of exploitation and class do not capture all features of modern economies, but they are the right tools to understand capitalism, and more generally the relation between property rights and relations of production.

But this defence is moot if one believes, as Folbre does, that a focus on capitalism as the object of analysis is excessively narrow, if not positively misleading. For a focus on capitalist institutions and the 'logic of capital' implies, according to Folbre, the neglect or at least the downplaying of the relevance of race or gender.⁹

From this perspective, a focus on capitalism as the object of analysis makes it difficult to understand the intersectional logic of contradictory group interests, which can only be analysed if one adopts a more comprehensive view of the economy, and of production. "The concept of 'production' can be widened to include both 'reproduction' defined as the production, development and maintenance of human capabilities—and 'social reproduction'—defined as the production, development and maintenance of social groups" (Folbre 2020, 455).¹⁰

As important as the spheres of reproduction and social reproduction may be, it is again unclear that much is gained in terms of analytical power and theoretical clarity by bundling all these phenomena together under the umbrella of 'the economy'.

On the one hand, this move does not allow one to construct an approach that generalises (some aspects of) Marxian theory. In Marxian theory, production, reproduction, and social reproduction are fundamentally distinct: they are governed by different laws and agents interact in these spheres in very different ways. Contrary to Folbre's claims, to obscure the

⁹ "David Harvey notes, for instance, that capitalism is permeated with race and gender oppression, but that the 'logic of capital' is not affected by them. He makes no mention of any 'logic' of race or of gender" (Folbre 2021, 74).

¹⁰ In Folbre (2021, 20), the definition of economic activities is further expanded by including "appropriation (such as theft and war)".

distinctions between capitalism and, say, patriarchy is fundamentally extraneous to a Marxist theoretical project. For Marxists it is at best conceptually misleading to call everything a productive, or economic activity. An economy without the clergy remains capitalist, an economy without the proletariat is not.

Indeed, it is unclear what remains distinctively Marxist in the general approach outlined by Folbre. She lists three features of Marxian social theory that her approach builds on: an explanation of group-based economic inequalities focusing on how strong groups can exploit weak groups, with consequences for the individuals within them; an explanation of why people may voluntarily consent to exploitation due to structural constraints; and the emphasis on the potential for collective action to transform such structural constraints. None of these features, individually or even collectively, are distinctively Marxist.

In contrast, the approach she proposes is congenial to neoclassical economic theory, in which production, reproduction, and social reproduction can all be subsumed under the term ‘economics’. This is especially evident today, given the developments in the discipline which have led to a significant widening of the scope of the neoclassical approach to a range of phenomena outside traditional disciplinary boundaries. But it follows directly from the classic definition of economics as “a science which studies human behaviour as a relationship between ends and scarce means which have alternative uses” (Robbins 1932, 15). Observe that nothing in this definition implies per se an individualist focus, or the neglect of social conflicts. The fact that in mainstream analyses parental activities are often not explicitly and fully evaluated (Folbre 2021, 88ff) is a matter of prejudice, not the product of a theoretical barrier.

On the other hand, and perhaps more importantly, one is led again to wonder whether obliterating the fundamental differences between social phenomena obscures more than it enlightens. True, “the similarities among different forms of authoritarian hierarchy provide some clues to their coevolution” (Folbre 2021, 129). But such similarities are arguably rather superficial. One should not lose sight of the fact that production, reproduction, and social reproduction—and the authoritarian hierarchies that emerge in these spheres—are essentially distinct and no obvious relations exist between them. As Folbre herself acknowledges, “reproduction and social reproduction have particularly important implications for the evolution of patriarchal institutions that long predate capitalism” (Folbre 2020, 455). After all, “the production and maintenance of human

capabilities is a necessary—and costly—aspect of *all* economic systems” (Folbre 2021, 82–83; emphasis added). Politically, the struggle against capitalism is different from the struggle against patriarchy, even if the ultimate aim should be the removal of all structures of oppression and domination.

To return again to our metaphor, crimes against property are not the same as crimes against the person and it is not theoretically insightful to insist that they be bundled together in a more ‘general’ approach to criminality as a whole. They are different phenomena that require different analytical tools, as well as different solutions. Once this is acknowledged, then it follows that the distinction between theft and blackmail is important: it is not that one crime is more alarming, or a more severe offence, than the other. Rather, if the aim is to understand offences against property, then one should focus on theft and ignore blackmail, even if ultimately the aim should be to reduce criminality in general.

REFERENCES

- Folbre, Nancy. 2020. “Manifold Exploitations: Toward an Intersectional Political Economy.” *Review of Social Economy*, 78 (4): 451–472.
- Folbre, Nancy. 2021. *The Rise and Decline of Patriarchal Systems: An Intersectional Political Economy*. London: Verso.
- Nash, John. 1953. “Two-Person Cooperative Games.” *Econometrica* 21 (1): 128–140.
- Reiman, Jeffery. 1987. “Exploitation, Force and the Moral Assessment of Capitalism: Thoughts on Roemer and Cohen.” *Philosophy & Public Affairs* 16 (1): 3–41.
- Robbins, Lionel. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.
- Roemer, John. 1982. *A General Theory of Exploitation and Class*. Cambridge, MA: Harvard University Press.
- Veneziani, Roberto, and Naoki Yoshihara. 2018. “The Theory of Exploitation as the Unequal Exchange of Labour.” *Economics and Philosophy* 34 (3): 381–409.
- Wright, Erik Olin. 2000. “Class, Exploitation, and Economic Rents: Reflections on Sorenson's ‘Sounder Basis’.” *American Journal of Sociology* 105 (6): 1559–1571.

Vivek Chibber is Professor of Sociology at New York University and author, more recently, of *The Class Matrix: Social Theory after the Cultural Turn* (Harvard University Press, 2022) and *Confronting Capitalism: How the World Works and How to Change it* (Verso, 2022).

Contact Email: <vivek.chibber@nyu.edu>

Roberto Veneziani is Professor of Economics at the School of Economics and Finance, Queen Mary University of London. His research interests include topics of liberal principles of distributive justice, axiomatic

exploitation theory, macrodynamic models of growth and distribution, egalitarian principles, distribution of resources between generations, sustainable development, and normative principles in economics. He is also interested in the history of economic thought and in political economy from a mathematical perspective. He has published widely in economics, political science, and philosophy.

Contact Email: <r.veneziani@qmul.ac.uk>

Review of Till Düppe and Ivan Boldyrev's (eds.) *Economic Knowledge in Socialism, 1945–89*. Durham, NC: Duke University Press, 2019, 321 pp.

MARTA PODEMSKA-MIKLUCH
Gustavus Adolphus College

Economic Knowledge in Socialism, 1945–89, edited by Till Düppe and Ivan Boldyrev, is a collection of twelve essays exploring the discussions and challenges of economic scholarship produced in the communist regimes of the Eastern Bloc. In putting together this excellent collection, Düppe and Boldyrev build on their extensive expertise in the history and philosophy of science. The essays are diverse in terms of their historical and geographic context and in terms of the dynamics, and the unique challenges, they capture. Each essay can be interpreted as a study of two tightly connected themes: the influence of the socioeconomic regime on the direction of economic scholarship and the institutional challenges confronted in the creation of economic knowledge. Effectively, each chapter provides a unique perspective on the importance of the institutional context in which economic knowledge is produced and the role that the consolidation of political power plays in shaping social science. These contributions make the collection an intellectual thrill for scholars interested in the influence of politics on economic scholarship, the history of socialism, and the history of economic thought. The volume will be easily accessible to readers already familiar with the post-World War II history of Eastern and Central Europe; though, others might need to occasionally refer to additional sources, as the chapters vary in the knowledge they require for full comprehension.

The book is divided into four parts, each part consisting of three chapters. The first part, *Discourses*, includes discussions of (1) strategies for wage setting in the early years of post-war Hungary (Martha Lampland), (2) the role of political patronage in the development of reform-oriented economic scholarship in Hungary, from 1953 to mid-1970s (György Péteri), and (3) the public engagement of Czechoslovak economists in the explanation and defense of proposed economic reforms in the 1960s (Vítězslav Sommer). The second part, *Doctrines*, offers (1) a discussion of how leading Soviet economists contributed to the ideological discourse

starting at the time of the October Revolution until the early 1970s (Oleg Ananyin and Denis Melnik), (2) an overview of how the anti-statist direction in the political economy of socialism was squashed in the Soviet Union in the early 1970s (Yakov Feygin), and (3) an examination of how the argument for the over-development of the Soviet military emerged and fared (Adam E. Leeds). The third part, *Techniques*, discusses (1) the evolving role that mathematical methods and systems science—economic cybernetics—played in economic planning in the Soviet Union, from the 1960s until perestroika (Richard E. Ericson), (2) the introduction of pattern recognition and algorithmic decision-making into models of a command economy (Olessia Kirtchik), and (3) the involvement of Soviet systems analysts in international development projects as a source of political dissensus (Eglė Rindzevičiūtė). The final part, *The International*, has chapters on (1) the growing disillusionment with the prospects of a Soviet-led development in decolonizing countries (Chris Miller), (2) neoclassical economics as economics of socialism and the evolving meaning of structural adjustment, focusing on Yugoslav economists (Johanna Bockman), and (3) the influence of convergence theory—predicting eventual Eastern-Western convergence—on perestroika (Joachim Zweynert).

Here, I will focus on three chapters—those by Péteri, Ericson, and Kirtchik. These chapters yield interesting insights into some of the main themes of the volume: the specific challenges faced by economists in communist regimes and the importance of being sensitive to the institutional context at hand.

In the introduction, the editors caution readers to avoid slipping into a comparative mode, as doing so might favor questions only relevant to Western economies, while leaving out phenomena that were unique to socialism. This point is illustrated particularly well by the discussion of academic patronage in the second chapter of the volume, “By Force of Power: On the Relationship between Social Science Knowledge and Political Power in Economics in Communist Hungary”. The author, György Péteri, devotes this chapter to the analysis of the relative autonomy of economic research observed in Hungary after 1953. Péteri attributes this autonomy to the influence of István Friss, who was a member of the Central Committee of the Communist Party from 1948 to the end of his life in 1978, and the head of the economic policy department of the Central Committee between 1948 and 1954. Friss was an influential member of the high echelons of the communist regime, with carefully accumulated

connections, reputation, and prestige backing his powerful position. Péteri paints Friss as a passionate intellectual with a deep devotion to scientific economic knowledge and empiricism, passions that prompted Friss to create the Institute of Economics of the Academy of Sciences. The Institute became central to Friss' influence over the production of economic research and to the protection he granted to his protégées. In the author's account, it was Friss' political influence that, through the force of power, enabled the international success of János Kornai's scholarship, a success that would not be possible through the power of thought alone.

Péteri opens the chapter with the assertion that while patronage of academic enterprises is necessary in any economic system, the state socialist order magnifies that need. As Péteri acutely observes, the peculiarities of socialism generated the phenomenon of *the patron's dilemma*: the more the patron invests in the protection of his clientele, the more he undermines his own position and reputation that enabled him to offer protection in the first place. The patron's dilemma captures the dynamics of the monocentric, consolidated power structure of socialist regimes. This brilliant observation is made almost in passing, with the chapter far more focused on the significance of Friss' influence that, in the author's eyes, had been downplayed in the literature. Because of this focus, the author seems to miss the importance of loyalty in patronage relationships. Péteri interprets Friss' protection as motivated by the joy derived from the professional success of his protégées. But surely there was more to that motivation. The loyalty generated by the protection of controversial revisionist economists must have contributed to Friss' long-term position as the head of the Institute of Economics of the Academy of Sciences. Non-controversial economists would have had far less need to remain loyal to Friss than those whose livelihoods depended on his protection. The exchange of protection for loyalty seems a far stronger interpretation of academic patronage in a system of consolidated state power.

While the phenomenon of the patron's dilemma might be specific to autocratic regimes, other challenges encountered by economists in socialist states seem more universal, though to be sure, their frequency and intensity might vary between autocracies and democracies. One such challenge is the interaction between policy advice and political interests. To be effective, change in public policy must alter economic and political interests, hence the frequent disappointment of economists in having their policy ideas thwarted by political reality, as captured by Richard Ericson's chapter on the System for Optimal Functioning of the Economy

(SOFÉ). While other chapters in the volume attribute the resistance to reforms mainly to ideology, Ericson counts political obstacles as equally, if not more, important. SOFÉ was developed in response to the weaknesses of Bolshevik planning. It was an attempt to introduce mathematical methods and systems science (economic cybernetics) as a replacement to direct management decisions, substituting algorithms for planners. Effectively, full adoption of SOFÉ implies substantial reduction in the political influence of those in control of resource and product allocation. The ‘transitional losses trap’—the potential loss of privileges resulting from their position within the regime—explains the political resistance to its implementation. Yet, as Ericson notes, the biggest challenge to SOFÉ lied not in the ideological and political objections to its implementation but in the unresolved issues of acquiring and processing necessary information and of mitigating deep incentive compatibility issues. One is left wondering how a three-decade long intellectual effort could be committed to a program that was so inherently flawed and unworkable and why Soviet scholars continued its development despite complete awareness of the program’s weaknesses. Is it possible that they had little choice but to build this house of cards? Or were they driven by a well-hidden political realism? In either case, they were rewarded for the intricacies of their designs, and for the smokescreen they provided to the political regime.

Olessia Kirtchik’s chapter on the work of Soviet scientists in pattern recognition and economic disequilibrium amplifies some of the themes captured by Ericson; in particular, it highlights the extent of the intellectual effort that went into the development of cybernetics in the Soviet Union. However, Kirtchik pays less attention to how this scholarship fit into the institutional landscape. This leads her to a comparison between, on the one hand, the development of cybernetics in the Soviet Union and, on the other, the development of algorithms and the recent adoption of machine learning in the West. But any such similarity must surely be superficial given the dramatic difference in purposes for which these tools were designed and used. In the Soviet Union, complex adaptive systems were supposed to teach workers to respond to top-down commands. In the West, they aggregate and process information for the purpose of improving customer experience. The former attempted to improve upon top-down control processes, where the direction of change originated from a single source. The latter attempts to improve on the emergent processes of knowledge production and the direction of change is crowdsourced.

While I cannot discuss all chapters in this short review, I hope that my comments will encourage readers to explore the collection in its entirety. I share the editors' hope and desire that the volume will encourage further research on the dynamics and challenges entrenched in the scholarly pursuits under socialism. For scholars who endeavor on this quest, Düppe and Boldyrev's volume will serve as an invaluable footprint.

Given the budding nature of this field of study, and the many topics that had to be left out from the current volume, the editors might want to consider subsequent collections. Should they do that, I have three suggestions. First, I would suggest adopting a more descriptive title. The current title falls short of capturing the essence of the contribution. This is not a book about economic knowledge or the history of economic thought. Rather, this is a book about the dynamics of producing economic scholarship under a totalitarian regime, the unique (and the more universal) challenges encountered by scholars in such conditions, and the role the institutional environment plays in shaping the selection of scholarly pursuits. A more descriptive title for the subsequent volumes would be "The Creation of Economic Knowledge in Socialism". Second, given the diversity of chapters, the book would benefit from a longer introduction, providing a context for each chapter and explaining how each chapter fits with the theme of the book. A longer introduction would provide readers with alternative approaches to sampling the chapters: for example, the early versus the final decades of the Soviet Bloc, or the Soviet Union versus the Republics versus the Satellites. Third, a more careful editorial and cross-disciplinary peer-review process could have helped avoid some mistakes. For example, in the first chapter, Martha Lampland describes the economy of Hungary prior to 1948 as capitalist while simultaneously recognizing that it was characterized by state control of the means of production. These oversights do not take much away from the core value of the book. Düppe and Boldyrev, and their collaborators, must be commended for the excellent volume they have produced and should feel encouraged to continue on this path.

Marta Podemska-Mikluch is an Associate Professor of Economics at Gustavus Adolphus College. Using the analytical approach of Entangled Political Economy, Podemska-Mikluch's research analyzes the role entrepreneurship plays at the intersection of divergent institutional settings and the role regulations play in fostering and curtailing innovation. Originally from Poland, Podemska-Mikluch earned her Ph.D. in Economics from George Mason University, M.S. in Applied Economics from St. Cloud State

University, and B.A. in Political Science also from St. Cloud State. Podemska-Mikluch is the founding director of the Entangled Political Economy Research Network that brings together more than 80 scholars interested in advancing the novel approach of Entangled Political Economy.

Contact e-mail: <mpodemsk@gustavus.edu>

Website: <www.podemska.com>

Review of Michel S. Zouboulakis' *The Varieties of Economic Rationality: From Adam Smith to Contemporary Behavioural and Evolutionary Economics*. New York, NY: Routledge, 2014, 192 pp.

YAM MAAYAN

Tel Aviv University

Rationality is an essential assumption in economics as a scientific discipline. This assumption seems trivial and straightforward: people aim at pursuing their own interests and they pursue these interests rationally. However, what do economists understand by 'pursuing one's interests rationally'? This is the question Michel Zouboulakis tries to answer in *The Varieties of Economic Rationality*. Zouboulakis' central claim is that throughout the history of economic thought, different economists have understood economic rationality in different ways. In particular, although economists usually believe that rational economic behavior is identified with rational maximization, in the form of cost-benefit analysis, a closer look shows that many prominent and influential economists held different conceptions of rationality.

The book consists of twelve chapters (plus an introduction and a conclusion), the first five of which are focused on a single preeminent thinker and their corresponding conceptions of rationality: Adam Smith, John Stuart Mill, William Stanley Jevons, Vilfredo Pareto, and Lionel Robbins. In chapter 6, Zouboulakis takes up various critiques of neoclassical rationality, starting in the late 1930s, by John M. Keynes, Friedrich Hayek, Ronald Coase, Terence Hutchison, Robert L. Hall, and Charles J. Hitch, before covering—in chapter 7—responses to this critique by Fritz Machlup, Paul Samuelson, and Milton Friedman. Chapter 8 deals with a philosopher, rather than an economist: it reviews Karl Popper's influential conceptions of rationality and methodological individualism. Chapter 9 is an overview of the development of probabilistic choice theory and game theory, focusing on the work of John von Neumann and Oskar Morgenstern, Leonard Savage, and John Nash. Chapter 10 is devoted to Herbert Simon's concept of bounded rationality, while chapter 11 deals with behavioral economics: it starts with the work of Daniel Kahneman and Amos Tversky, followed by the work of experimental economist Vernon Smith,

and finishes with a discussion of the role of institutions and emotions in people's choices. Chapter 12 provides the last conception of rationality discussed in the book—the rationality of socially embedded individuals. In this chapter, Zouboulakis presents a different chronological development from that of the rest of the book, starting with Karl Marx's idea of social consciousness, followed by a review of the institutionalist tradition within economics, including the work of Thorstein Veblen, John Commons, Douglas North, and Geoffrey Hodgson. The last part of this chapter deals with economic sociology.

While reviewing numerous critical scholars that influenced the concept of rationality, the reader is taken through some of the most central issues concerning economics as a scientific discipline—for example, the question of the scope of economics, the need for grounding the meaning of theoretical concepts in empirical observations, and the proper way to interpret modeling assumptions.

In general, Zouboulakis does an impressive job defining the specific methodological approaches of the economists he deals with and categorizing the variances between the different concepts of rationality he traces. Broadly, he presents three categories related to the assumption of rationality, on the basis of which different economists can be classified. The first concerns the foundations of the rationality assumption—that is, does it need to rely on psychology as an empirical ground or can it be derived in a purely theoretical manner? The second concerns the assumption's validity—that is, is the definition of rationality universal or is it dependent on a social and historical context? The third concerns the meaning of the rationality assumption—that is, does it refer only to individual behavior, or does it also refer to the economic system as a whole? In other words, does the rationality assumption entail that rational agents have extensive computational capacities and access to all the relevant information, or do agents face some unresolved uncertainty?

Interestingly, there is no necessary relation between the answers to these questions. For example, while both John Stuart Mill and Francis Y. Edgeworth believed that the rationality assumption should rely on psychological foundations, Mill thought that rationality is socially dependent while Edgeworth saw it as universal. Similarly, while Paul Samuelson and Kenneth Arrow shared the view of perfect rationality (including complete information and unlimited computational skills), Arrow did not refer to any psychological foundations while Samuelson based his approach on behaviorist psychology.

One crucial point coming out of the historical review presented in the book is that there is no linear development of the rationality concept, starting from the classical *homo œconomicus*, through the late neoclassical *rational choice-maker*, and finishing with more complicated visions of rational agents characterizing contemporary economics.

Throughout the historical period covered in the book, not only did economists understand economic rationality differently, but also the same methodological issues came up again and again, in a cyclic manner, presenting methodological problems for attempts to define rationality—problems with no definitive solutions. For example, the concept of perfect knowledge was adopted by early neoclassical economists, criticized during the late 1930s, abandoned by game theorists during the 1940s and 1950s, readopted with the development of the general equilibrium framework and the idea of rational expectations, and then again criticized by supporters of the concept of bounded rationality in the late 1970s. A similar story can be told about the complicated relation between economics and psychology, as economists wavered between two stances, one claiming that psychology must provide the empirical ground for economic assumptions concerning human behavior, the other claiming that economists must avoid any references to psychology.

The book presents a second argument, which is more methodological than historical. Zouboulakis claims that the common lacuna in how economists understand and define economic rationality is not due to the lack of psychological foundations or the disregard of fundamental uncertainty, but rather the insufficient treatment of the social embeddedness of individuals. As Zouboulakis writes in his conclusion:

[...] it should be definitely recognized as a fact that one quintessential characteristic of humans is that they live embedded within a common system of moral values and social habits that give them a sense of social existence and identity. (138)

This statement comes up repeatedly throughout the book and guides its historical investigation. In recent years, it has been common to speak about the escape of economics from psychology—a methodological turn that peaked at the beginning of the twentieth century, inspired by Vilfredo Pareto's attempt to replace psychological assumptions with the concept of *logical choice*. Zouboulakis traces a different escape, which we can call 'the escape from sociality (or sociology)'. He claims that classical economists, among others, acknowledged the importance of the social

environment when analyzing human behavior and treated this fact carefully, even when they did not give it an explicit place within their theory. Zouboulakis describes a methodological tension in the work of classical and some of the early neoclassical economists—a tension between the way those scholars chose to define rationality and how they understood the scope of economics. For economists such as Smith, Mill, and Pareto, the less socially embedded the definition of rationality is, the narrower the scope of economics and the weaker the cogency of economic analysis is. According to Zouboulakis, this acknowledgment of human sociality is missing from neoclassical and current mainstream economics.

Following this argument, within the historical narrative in the book, the crucial ‘escape’ did not occur with Pareto’s but rather with Jevons’ work. According to Zouboulakis, Jevons was the first to define economic rationality as rational maximization, stripping it out from any particular social feature and at the same time giving it a universal status based on introspection (chapter 3). In chapter 4, Zouboulakis claims that although Pareto also abstracted rational behavior from any social context as part of his attempt to ground economics on logic alone, he made it clear that this move limits the scope of economics, as he believed economics should form only one branch in the broader investigation of human behavior and social phenomena. In contrast, Jevons used the alleged naturalness of the rationality assumption to give economics a universal validity. In this sense, Jevons is the origin of the tradition claiming rational maximization should form the fundamental point of departure for any investigation within the social sciences. This tradition includes some prominent neoclassical economists such as Garry Becker and George Stigler.

The less satisfactory part of the book contains those chapters dealing with the development of neoclassical concepts of rational choice in the post-war period (chapters 9–10). Although Zouboulakis mentions some key points in this development—for example, Paul Samuelson’s preference theory, the development of the general equilibrium framework, and the development of game theory—the book does not present a thorough methodological discussion of these issues. This is unfortunate because the post-war developments have shaped the standard conception of rationality prevalent among contemporary economists. The insufficient methodological discussion of these issues also makes it hard for Zouboulakis to relate the tensions he identifies earlier in the book to contemporary debates in economic methodology.

Zouboulakis dedicates a long (penultimate) chapter to behavioral economics (chapter 11). In this chapter, he presents the contribution of Kahneman and Tversky's work—as well as of Vernon Smith's experimentalist approach—to the way economists think about people's choices, and discusses the attempts of behavioral economists to incorporate human emotions and social institutions into models of economic behavior. However, because the development of contemporary conceptions of rationality do not get enough attention, this chapter cannot present the profound methodological issues at stake, as there is no substantial reference to the body of knowledge behavioral economics attacks. In particular, for the reader, it is hard to understand if Zouboulakis thinks that behavioral economics is a promising way to deal with the main challenge presented throughout the book—that of social embeddedness. A detailed answer to this question would have distinguished between two ways of thinking about the behavioral economic critique of the rationality assumption: (i) one holds that people do not act rationally because they lack the cognitive ability to make rational choices; (ii) the other holds that the standard definition of rationality is inadequate because it fails to deal with fundamental issues such as social norms, altruistic behavior, and the social construction of preferences. Moreover, it could have been interesting to compare behavioral economics, which, in recent years, has become a part of mainstream economics, to the more unorthodox critiques Zouboulakis presents by evolutionary institutionalists and social economists.

An interesting aspect that is not covered in the book is the connection between different conceptions of rationality and different approaches to welfare economics. Although many of the economists the book deals with have made significant contributions to the normative branch of the discipline, Zouboulakis does not discuss the interrelations between their ethical stances and their conceptions of rationality. Nevertheless, *The Varieties of Economic Rationality* presents an absorbing outline of the development of the most fundamental concept in economic theory. It reminds us that historical investigation can provoke us to rethink current methodological and philosophical questions. The book also presents some interesting points that deserve further analysis.

Yam Maayan is a PhD student at the economics department at Tel Aviv University, writing about the different ways rational choice theory has influenced normative economics.

Contact e-mail: <yamaayan@gmail.com>

PHD THESIS SUMMARY:

A Tale Between Finance and Economics: Four Essays on the History and Methodology of the Efficient Market Hypothesis

THOMAS DELCEY

PhD in Economics, January 2021

Université Paris 1 Panthéon-Sorbonne

Financial globalization is arguably the most important phenomenon in the economic history of the twentieth century. This thesis revisits the economic ideas that preceded and accompanied this transformation. Largely ignored or disparaged by economists at the beginning of the last century, finance gradually attracted more attention in the profession. This culminated, in the 1960s, in the emergence of the field of financial economics. This thesis contributes to the history of economic thought on financial markets (see, for example, Walter 1996; Mehrling 2005; and Jovanovic 2008) by discussing the evolution of such thought throughout the twentieth century. More precisely, I focus on North American economic thought, and, particularly, on the history and epistemology of one of its most central theories: the *efficient market hypothesis*. As it is broadly understood today, the efficient market hypothesis—associated with the work of Eugene Fama (1965, 1970)—claims that, in an efficient market, asset prices fully reflect all available information.

This thesis is partitioned into two broad parts: chapters 1–3 focus on the history of financial thought, and chapter 4 focuses on methodological issues in this history. The first three chapters analyse the origin of the efficient market hypothesis during the 1920s and its evolution until the early 1980s. This part of the thesis provides a new perspective on the history of financial economics by highlighting influences on the discipline by two types of economists: (1) those, such as inter-war agricultural economists like Holbrook Working, who preceded the emergence of financial economics as a subfield, and (2) those, such as Paul Samuelson, Thomas Sargent, and Robert Lucas, who remained outsiders of this subfield. These three chapters use a variety of materials—namely, the published writings of economists, private and administrative archives, and authors’ recollections of the most recent period (the 1970s and beyond). The fourth (and last) chapter offers an epistemological, rather than a historical, analysis.

Presented in comparison with Friedrich Hayek's information theory, this last chapter discusses the conceptual foundations of the efficient market hypothesis.

The first chapter traces the roots of the contemporary debate on informational efficiency to inter-war agricultural economics—at that time, an emerging field funded mostly publicly by the United States Department of Agriculture.¹ In the history of economics, information is mainly perceived as a concept that emerged from the socialist calculation debate and that was subsequently developed in a more formal framework during the post-war area. Little has been said, however, in the literature about pre-war economic thought on information. Using the archives of the Department of Agriculture and its Bureau of Agricultural Economics, I show that inter-war agricultural economists understood that market prices reflect information, far before post-war information theories, such as the efficient market hypothesis, made this point. They were at the forefront of agricultural market reforms aimed at improving the production and circulation of information in agricultural exchanges. These economists were also pioneers in the empirical analysis of derivatives markets, where they advocated for greater transparency. In this chapter, I focus on and reassess the contributions of Holbrook Working (1935, 1949), a forerunner of the efficient market hypothesis. Rather than an isolated forerunner, Working was above all an important contributor to an ongoing research program in early agricultural economics led by the Department of Agriculture. One key message of this chapter is that economic policy, rather than economic theory, came first—that is, it was economists' practice of economic policy that prompted theoretical debates about information. Through the practical reform of agricultural markets, economists at the time developed and shared knowledge on the role of information in markets—knowledge that Working would later use to build his personal version of the efficient market hypothesis.

The second chapter focuses on Paul Samuelson's contribution to the development of the efficient market hypothesis during the emergence of financial economics in the 1960s and early 1970s.² The contribution of this chapter is twofold. First, it relates the development of the modern understanding of market efficiency to Working's earlier contributions. Based on Samuelson's archives, the chapter documents Samuelson's correspondence with Working in the 1960s. I show that, in his own work on

¹ See Delcey and Noblet (2021) for the associated publication.

² For more on Samuelson's contribution, see Delcey (2019).

market efficiency, Samuelson (1965, 1973) both built on Working's ideas and translated them into the formal framework of post-war economics. Second, the chapter sheds light on the close but ambivalent relationship in the 1960s between early financial economists and traditional economics. Specifically, I compare Samuelson's evolving positions on market efficiency with those of Eugene Fama. I observe that not only did financial economics rely on the theoretical framework and methods of post-war economics for its development, but that it also developed a new type of economic expertise. While economists like Samuelson considered themselves to be traditional government advisors, the younger generation of financial economists employed their economic tools for producing evaluations useful to the practice of corporate and portfolio managers, rather than public officials.

The third chapter explores the growing role that macroeconomists played in the efficient market research program during the 1970s and early 1980s, a murky corner in the history of economic thought.³ The chapter documents the early interactions between participants in the efficient market research program and rational expectations macroeconomists. Using bibliographical analysis and the personal recollections of authors, such as Eugene Fama, the chapter documents the meeting between these two scholarly communities and explores how they influenced each other. A key place in this story is the Carnegie Institute of Technology, where Fama's student Richard Roll met Robert Lucas and Thomas Sargent. The chapter analyses the first tangible output of this meeting—a series of articles on the 'yield curve' that were published in the early 1970s and that involved Eugene Fama, Robert Shiller, Thomas Sargent, and Franco Modigliani.⁴ One of the main outcomes of this debate was the methodological influence of new classical macroeconomics on the market efficiency research programme, which led to a reformulation of the efficient market hypothesis. More precisely, market efficiency became commonly defined through rational expectations models. Furthermore, by the end of the 1970s, the debate on the yield curve had paved the way to two new research questions in finance, namely, the issue of asset valuation—how to correctly price the 'economic fundamentals' of assets—and of the dissemination of information in market exchanges, both of which relied on rational expectations models.

³ See Delcey and Sergi (2019) for more on the early interactions between macroeconomics and financial economics.

⁴ For a glimpse of this conversation, see Sargent (1972, 1973), Modigliani and Shiller (1973), and Fama (1975).

Finally, the fourth chapter is a methodological analysis that compares the notion of market efficiency, as it was developed and understood by financial economists, with Friedrich Hayek's information theory (see, for example, Hayek 1937, 1945).⁵ Hayek's rejection of the use of mathematics makes it difficult to compare his thought with mainstream (financial) economics, which relies on formal theories and econometrics. To solve this issue, the chapter introduces a distinction between two types of differences between theories, *epistemological* and *methodological differences*. I define the idea of an epistemological difference between theories as the differing core hypotheses of these theories, and I understand the methodological differences between theories as the different ways in which authors operationalize these core hypotheses. The chapter argues that there is a common epistemological core underlying Hayek's information theory and the theory of market efficiency. At their heart is the same puzzling issue, which I call *the information problem*: Is it possible to centralize locally produced knowledge for the purpose of government planning, in the case of Hayek, or for the purpose of profitably forecasting price changes, in the case of market efficiency? In each framework (Hayek's and that of market efficiency), this information problem is solved by attributing a new function to the market—that is, both theories claim that, in addition to being a traditional clearing mechanism, the market also aggregates the local knowledge of individuals into prices.

The contribution of this thesis is twofold. First, it sheds light on the analytical debates surrounding the controversial notion of market efficiency. Debated since the 1980s, the concept of market efficiency—including its meaning and implications—remains contentious in economics to this day. I analyse and discuss the diversity of formulations and interpretations of this concept throughout its history. The thesis focuses especially on the history and meaning of the concept of information, which is at the heart of the definition of market efficiency. I trace the roots of this concept back to the American economic thought of the inter-war period. While the concept of information became a guiding principle for financial research after the war, I argue that the inability to define 'information' in an objective way contributed, and still contributes, to the ambivalence of the idea of market efficiency. In particular, the ambiguity in the precise meaning of the concept of information—that is, what pieces of information are 'relevant' or 'available'—perpetuates a dual interpretation of what market efficiency refers to: (1) the absence of systematically

⁵ For more on this comparison, see Colin-Jaeger and Delcey (2020).

profitable investment strategies, or (2) the accurate valuation of asset fundamentals.

By focusing on the concept of market efficiency, this thesis also studies the origin and evolution of financial economics, for which the notion of market efficiency has guided several decades of research. Particularly, the thesis examines the roots of financial economics in the first half of the twentieth century and its emergence in the 1960s, and discusses its changing relationship with economics in the 1970s and the early 1980s. As I argue in the thesis, during the evolution of this relationship, financial economics adopted concepts and methods from post-war economics and, in the process, developed distinct new ways of thinking, such as the conceptual model of practical expertise discussed in chapter 2. This thesis thus also contributes to our understanding of the growing importance of financial economics, which is where ideas criticizing and legitimizing the financial world are mainly produced today.

REFERENCES:

- Colin-Jaeger, Nathanaël, and Thomas Delcey. 2020. "When Efficient Market Hypothesis Meets Hayek on Information: Beyond a Methodological Reading." *Journal of Economic Methodology* 27 (2): 97-116.
- Delcey, Thomas. 2019. "Samuelson vs Fama on the Efficient Market Hypothesis: The Point of View of Expertise." *Oeconomia: History, Methodology, Philosophy* 9 (1): 37-58.
- Delcey, Thomas, and Guillaume Noblet. 2021. "'The Eyes and Ears of the Agricultural Markets': A History of Information in Interwar Agricultural Economics." CHOPE Working Paper No. 2021-20. Center for the History of Political Economy, Durham, NC.
- Delcey, Thomas, and Francesco Sergi. 2019. "The Efficient Market Hypothesis and Rational Expectations. How Did They Meet and Live (Happily?) Ever After." HAL Working Paper No. hal-02187362. Centre pour la Communication Scientifique Directe, Lyon.
- Fama, Eugene F. 1965. "The Behavior of Stock-Market Prices." *The Journal of Business* 38 (1): 34-105.
- Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2): 28-30.
- Fama, Eugene F. 1975. "Short-Term Interest Rates as Predictors of Inflation." *The American Economic Review* 65 (3): 269-282.
- Hayek, Friedrich A. 1937. "Economics and Knowledge." *Economica* 4 (13): 33-54.
- Hayek, Friedrich A. 1945. "The Use of Knowledge in Society." *The American Economic Review* 35 (4): 519-530.
- Jovanovic, Franck. 2008. "The Construction of the Canonical History of Financial Economics." *History of Political Economy* 40 (2): 213-242.

- Mehrling, Perry. 2005. *Fischer Black and the Revolutionary Idea of Finance*. Hoboken, NJ: John Wiley & Sons.
- Modigliani, Franco, and Robert J. Shiller. 1973. "Inflation, Rational Expectations and the Term Structure of Interest Rates." *Economica* 40 (157): 12-43.
- Samuelson, Paul A. 1965. "Proof That Properly Anticipated Prices Fluctuate Randomly." *Industrial Management Review* 6 (2): 41-49.
- Samuelson, Paul A. 1973. "Proof That Properly Discounted Present Values of Assets Vibrate Randomly." *The Bell Journal of Economics and Management Science* 4 (2): 369-374.
- Sargent, Thomas J. 1972. "Rational Expectations and the Term Structure of Interest Rates." *Journal of Money, Credit and Banking* 4 (1): 74-97.
- Sargent, Thomas J. 1973. "Interest Rates and Prices in the Long Run: A Study of the Gibson Paradox." *Journal of Money, Credit and Banking* 5 (1): 385-449.
- Walter, Christian. 1996. "Une Histoire du Concept d'Efficiencie sur les Marchés Financiers." *Annales: Histoire, Sciences Sociales* 51 (4): 873-905.
- Working, Holbrook. 1935. "Differential Price Behavior as a Subject for Commodity Price Analysis." *Econometrica* 3 (4): 416-427.
- Working, Holbrook. 1949. "The Investigation of Economic Expectations." *The American Economic Review* 39 (3): 150-166.

Thomas Delcey received his PhD in economics at the Université Paris 1 Panthéon-Sorbonne. He currently holds a post-doctoral position at the Center for the History of Political Economy at Duke University. His main research area is the history of financial economics. Thomas is also engaged in a project with political philosophers on the recent use of the concepts of efficiency and neutrality by central bankers in Europe. Contact email: <thomas.delcey@duke.edu>

PHD THESIS SUMMARY:

**Otto Neurath and Ludwig von Mises: Philosophy, Politics,
and Economics in Viennese Late Enlightenment**

ALEXANDER LINSBICHLER

PhD in Philosophy, October 2020

University of Vienna

Logical empiricism and the Austrian School of economics are two of the internationally most influential intellectual movements with Viennese roots. By and large independently of each other, both have been subject to detailed historical and philosophical investigations for the last two decades. However, in spite of numerous connections and interactions between the two groups, their relationship has captured surprisingly sparse attention. My dissertation focuses on the many-faceted juxtaposition of two supposedly antagonistic champions of Viennese Late Enlightenment: logical empiricist Otto Neurath and Austrian economist Ludwig Mises. I rationally reconstruct and critically compare their epistemological, methodological, and economic positions and demonstrate that a closer look reveals more compatibilities and similarities than acknowledged by the received view and by the protagonists themselves.¹ Over and above the historiographic task of challenging and amending this received view, the analytic components of my thesis inform contemporary debates in philosophy, politics, economics, and other sciences.

¹ Milonakis and Fine, for instance, characterize Mises's praxeology as "the most anti-positivist and anti-empiricist approach to social science ever stated" (2004, 259), which *prima facie* does not square well with Neurath's 'empirical sociology' (1931; see also [1931] 1973 for the standard English translation of excerpts of the 1931 book) or his 'radical physicalism' ([1934] 1983). While the antithetical opposition in philosophy, methodology, science, and politics is usually treated as an implicit certainty, Boettke, echoing Sigmund (2017), eloquently voices the received view on the relation between logical empiricism and the Austrian School: "How actually would one engage in 'exact thinking during demented times'? One answer was provided by the Vienna Circle, the other was provided by Hayek" (Boettke 2018, 33; see also 293).

AUTHOR'S NOTE: I want to thank my thesis supervisors Elisabeth Nemeth and Friedrich Stadler, my thesis examiners Rainer Hegselmann and Scott Scheall, as well as Hasok Chang, Ivan Ferreira da Cunha, Erwin Dekker, Malachi Hacoheh, Karl Milford, Thomas Uebel, and all others who supported me by critically challenging my ideas and/or otherwise. I gratefully acknowledge funding from the Austrian Science Fund (FWF): W 1228-G18 and the Austrian Academy of Sciences.

Following an introduction and overview, chapter 2 reviews the existing literature on Neurath, Mises, and their encounter in the socialist calculation debates and questions the received view according to which Neurath and Mises are diametrically opposed in all respects. Admittedly, the socialist empiricist Neurath advanced calculation in kind for a ‘system of socialisation’ ([1921] 2004) whereas the classically liberal apriorist Mises devised the famous calculation argument against central planning. Yet, both scholars drew inspiration from the scientific and educational enterprise of Viennese Late Enlightenment, endorsed democracy, denied the possibility of meaningful monetary calculation under socialism, and sought to establish a viable notion of what is to be regarded as relevant and justified knowledge in the social sciences. The fact that Neurath and Mises also shared many philosophical, economic, and political opponents—including life-threatening totalitarian regimes—further motivates the more thorough analysis of their positions I prompt in chapter 2 (Linsbichler 2015).

Chapters 3 and 4 portray Felix Kaufmann as yet another idiosyncratic representative of the cultural milieu of interwar Vienna and as a mediator between the Vienna Circle and the Mises-Kreis. Kaufmann’s methodological writings, developed at the periphery of both logical empiricism and Austrian economics, facilitate understanding of their disagreements—some of them genuine and some of them merely apparent. His correspondence with Neurath indicates that what Kaufmann relayed to the Geistkreis and the Mises-Kreis as the doctrine of the Vienna Circle only captured overly reductionist, verificationist, and positivistic snippets of logical empiricism. Thereby, notwithstanding his other merits, Kaufmann contributed to the self-perception of many Austrian economists as anti-thetically opposed to logical empiricism (Linsbichler 2019; Linsbichler and Taghizadegan 2019a, 2019b).

The centrepieces of the thesis, chapters 5 to 8, employ conceptual tools of contemporary philosophy of science to identify and analyse three areas of hitherto neglected compatibilities and similarities between Neurath and Mises. First, I present an analytic version of Mises’s theory of human action which renders the apriorism of Austrian economics compatible with a logical empiricist stance (chapters 5 and 6); second, I consider their shared methodology of counter-factual reasoning (chapter 7); and third, I discuss common presuppositions and some consensual conclusions in the socialist calculation debates (chapter 8). The historical insights gained in these case studies in turn contribute to contemporary

philosophical debates on first principles in economics, logic, and mathematics, on thought-experiments and the use of unrealistic models, as well as on rationality, nudging, the role of knowledge in society, and presuppositions of assessments of social well-being.

Chapter 5 explicates and construes the aprioristic elements of Austrian economics, specifically the fundamental axiom of Mises' praxeology, as analytic instead of synthetic. The fundamental axiom, "man acts" (see, e.g., Mises 1962, 4), states that at least some human behaviour is purposeful, i.e., human individuals and only human individuals subjectively choose goals and apply means they subjectively consider expedient to attain these goals. Whereas the most prevalent view interprets Mises's fundamental axiom as a synthetic a priori judgment and has instigated many philosophers and economists to outright reject praxeology, I propose a shift from a synthetic fundamental axiom to an analytic one. Contrary to claims by many praxeologists, it is perfectly conceivable to explain human behavior employing alternatives to the fundamental axiom.² Neither direct observation nor intuition nor pure reason can rule out these alternatives conclusively, hence in the final analysis, the justification of the fundamental axiom is pragmatic. The ensuing conventionalist version of praxeology alleviates the charges of extreme apriorism against Austrian economics and makes praxeology more acceptable from a contemporary as well as from a logical empiricist perspective. One examiner pointedly described chapter 5 as 'saving praxeology from its originator' (Linsbichler 2017, 2021a).³

Logical empiricists' approval of analyticity and conventionalism in logic and mathematics is exemplified in chapter 6. Specifically, Neurath's brother-in-law, Hans Hahn, is portrayed as a pioneer of logical pluralism and of logical tolerance who adopted and adapted Russell's logicism and

² I draw analogies to the case of the parallel postulate of Euclidian geometry, which used to be deemed without alternative and synthetic a priori.

³ Although I identify several oft-neglected passages in Mises's writings which hint in the direction of analyticity and thus much less extreme apriorism, I certainly do not claim that Mises was a self-aware, full-fledged conventionalist. Rather, my constructive proposal aims at dispelling charges according to which praxeology is untenable because it relies on extreme apriorism. Regardless of details of the exegesis of Mises's epistemological deliberations, contemporary Austrian economists in Mises's tradition can continue their scientific endeavours without significantly altering the content of praxeology, but merely its epistemological status and the stance towards alternative research programmes. By contrast albeit in an equally constructive-minded spirit, Lipski (2021) suggests to reform praxeology by adding empirical content to the fundamental axiom to obtain a directly testable hypothesis, thereby dropping the aprioricity essential to Mises and most of his followers.

Wittgenstein's nominalism, and who anticipated a philosophy of mathematics made famous by Carnap (Linsbichler 2018). Moreover, Hahn's remarks on the nature of definitions render his conventionalism applicable to other purely analytic disciplines besides logic and mathematics, thus arguably also to praxeology.

Chapter 7 reconstructs the use of Neurath's 'scientific utopias' and Mises' 'imaginary constructions' as linchpins of thought experiments, thereby illustrating similarities in their methodology of counterfactual reasoning and their common groundwork to the then emerging subdiscipline comparative economic systems. The chapter also informs recent discussions on the epistemological status of thought experiments and unrealistic models. Specifically, I argue that Norton's (1996, 2004) argument view of thought experiments can account for new discoveries in ways Neurath anticipated, and further, I reformulate Häggqvist's (2009) model for thought experiments to highlight the role of alternatives and decisions in science and in public debate (Linsbichler and Cunha 2021).

Chapter 8 rationally reconstructs and critically compares the different and idiosyncratic conceptions of rationality defended by Neurath and Mises and suggests some consequent insights with respect to contemporary rationality wars, the socialist calculation debates, the foundations of welfare economics, and Viennese Late Enlightenment. The cautionary character of the latter is pinpointed by Neurath, foreshadowing a Hayekian theme: "Rationalism sees its chief insight in the clear recognition of the limits of actual insight" (Neurath [1913] 1983, 8). Considering Mises' deliberations on the rationality of individual action together with his denial of the possibility of rational action under socialism, I identify a tension: How can, as Mises maintains, all human actions be rational (in his sense of the term) and yet socialism preclude rational action in complex situations? Discussing problems of other solutions to this interpretational problem,⁴ I dissolve the tension by explicating Mises' sense of the terms 'rational' and 'action': as a result (and according to Mises), socialism precludes rational action because socialism precludes action. Chapter 8 subsequently highlights Neurath's and Mises' shared concern for the limits of rationality and for the potential of knowledge to improve decisions, and finally draws on Sugden's (2013) distinction between welfarist and contractarian perspectives to reveal hitherto overlooked compatibilities

⁴ Aside from Neurath, more recently O'Neill (1998), Salerno (1993), and Uebel (2018) at least implicitly dissolve the tension differently, namely by reading Mises as equating rationality with monetary maximization in the context of the calculation debates.

in the socialist calculation debates. Both Neurath and Mises reject monetary calculation, including most forms of cost-benefit-analysis as an evaluative standard on the social level, i.e., for the comparison of economic orders. Whereas Neurath enhances and champions calculation in kind as an alternative, Mises does not provide any workable evaluative standard. He regards calculation in kind as overly ponderous but does not offer principled objections against its use. In any case, Mises maintains that any (reasonable) evaluative standard on the social level strongly suggests the adoption of an economic order which provides meaningful money prices for monetary calculations on the part of acting individuals. As long as a by and large capitalistic economy prevails, both Mises and Neurath accept that individuals or individual firms voluntarily use monetary calculation accompanied by limited versions of calculation in kind, for instance so-called common good balance sheets (Linsbichler 2021c, 2021e).

My dissertation thesis is a starting point for further systematic reconstructions and critical comparisons of positions maintained in the logical empiricist tradition, on the one hand, and positions in the vicinity of Austrian economics, on the other.⁵ Chapter 9 indicates a number of suggestions for subsequent research, such as: (i) a re-evaluation of Austrian economists' stance opposing formal methods in the social sciences (Linsbichler 2021d); (ii) unearthing Carl Menger's, Karl Menger's, and Hahn's role in the early history of the principle of logical tolerance made famous by Carnap; (iii) an exploration of how, given the challenge of logical tolerance and logical pluralism, Mises's logical monism can be modified in order to safeguard the anti-racist conclusions he infers from it; (iv) further development and partial formalization of analytic praxeology as devised in chapter 5; (v) a history of proposals for universal basic income by scientific utopians; (vi) a reconstruction of Neurath's, Mises's, and Kelsen's thoughts on democracy and on the role of experts and education in a democratic social order. The lattermost topic notably indicates that many problems of philosophy and political economy debated in Viennese Late Enlightenment have not lost their significance in the 21st century.

REFERENCES

Boettke, Peter J. 2018. *F. A. Hayek. Economics, Political Economy and Social Philosophy*. London: Palgrave Macmillan.

⁵ The heterogeneity and complexity of logical empiricism beyond left-wing, positivistic reductionism has been re-discovered in recent decades. For an appreciation of the heterogeneity and complexity of the Austrian School of economics beyond aprioristic market fundamentalism, see e.g., Linsbichler (2020, 2021b, forthcoming).

- Häggqvist, Sören. 2009. "A Model for Thought Experiments." *Canadian Journal of Philosophy* 39 (1): 55–76.
- Linsbichler, Alexander. 2015. "Otto Neurath and Ludwig von Mises: The Socialist Calculation Debates and Beyond." In *Interactions in the History of Philosophy II*, edited by Burcin Ercan, 311–324. Istanbul: Delta Publishing.
- Linsbichler, Alexander. 2017. *Was Ludwig von Mises a Conventionalist?: A New Analysis of the Epistemology of the Austrian School of Economics*. Basingstoke: Palgrave Macmillan.
- Linsbichler, Alexander. 2018. "Was man aus Einflüssen machen kann—Hans Hahns Adaptierung von Russells Logizismus und Wittgensteins Nominalismus." *Contributions of the Austrian Wittgenstein Society / Beiträge der Österreichischen Wittgenstein Gesellschaft* XXVI: 138–140.
- Linsbichler, Alexander. 2019. "Felix Kaufmann: 'A Reasonable Positivist'?" In *Ernst Mach—Life, Work, Influence; Vienna Circle Institute Yearbook 22*, edited by Friedrich Stadler, 709–719. Cham: Springer.
- Linsbichler, Alexander. 2020. "Wieser, Friedrich Frh. von (1851–1926), Nationalökonom und Soziologe." In *Österreichisches Biographisches Lexikon 1815–1950*. Band XVI, Lieferung 71, 193–194. Vienna: Austrian Academy of Sciences Press.
- Linsbichler, Alexander. 2021a. "Austrian Economics Without Extreme Apriorism: Construing the Fundamental Axiom of Praxeology as Analytic." *Synthese* 198: 3359–3390.
- Linsbichler, Alexander. 2021b. "Philosophy of Austrian Economics." In *The Routledge Handbook of the Philosophy of Economics*, edited by Julian Reiss and Conrad Heilmann, 169–185. Abingdon: Routledge.
- Linsbichler, Alexander. 2021c. "Rationalities and Their Limits: Reconstructing Neurath's and Mises's Prerequisites in the Early Socialist Calculation Debates." *Research in the History of Economic Thought and Methodology (RHETM)* 39B: 95–128.
- Linsbichler, Alexander. 2021d. "Sprachgeist and Realisticness: The Troubled Relationship between (Austrian) Economics and Mathematics Revisited." Center for the History of Political Economy at Duke University Working Paper No. 2021–15. Duke University, Durham, NC.
- Linsbichler, Alexander. 2021e. "Viennese Late Enlightenment and the Early Socialist Calculation Debates: Rationalities and Their Limits." Center for the History of Political Economy at Duke University Working Paper No. 2021–16. Duke University, Durham, NC.
- Linsbichler, Alexander. Forthcoming. *Viel mehr als nur Ökonomie: Köpfe und Ideen der Österreichischen Schule der Nationalökonomie*. Wien: Böhlau.
- Linsbichler, Alexander, and Ivan Ferreira da Cunha. 2021. "Otto Neurath's Scientific Utopianism Revisited: A Refined Model for Utopias in Thought Experiments." Unpublished manuscript.
- Linsbichler, Alexander, and Rahim Taghizadegan. 2019a. "Commentaries to the Songs." In *Felix Kaufmann's Songs of the Mises-Kreis: Wiener Lieder zur Philosophie und Ökonomie*, edited by Rahim Taghizadegan and Huw Rhys James, 57–196. Wien: Mises.at.
- Linsbichler, Alexander, and Rahim Taghizadegan. 2019b. "Introduction to the New Edition." In *Felix Kaufmann's Songs of the Mises-Kreis: Wiener Lieder zur Philosophie und Ökonomie*, edited by Rahim Taghizadegan and Huw Rhys James, 11–16. Wien: Mises.at.
- Lipski, Jonas. 2021. "Austrian Economics Without Extreme Apriorism: A Critical Reply." *Synthese*, 199: 10331–10341.

- Milonakis, Dimitris, and Ben Fine. 2009. *From Political Economy to Economics: Method, the Social and the Historical in the Evolution of Economic Theory*. London: Routledge.
- Mises, Ludwig von. 1962. *The Ultimate Foundation of Economic Science. An Essay on Method*. Princeton: D. van Nostrand.
- Neurath, Otto. 1913. "The Lost Wanderers of Descartes and the Auxiliary Motive." In *Otto Neurath: Philosophical Papers 1913-1946*, edited by Robert S. Cohen and Marie Neurath, 1-12. Dordrecht: D. Reidel, 1983.
- Neurath, Otto. 1921. "A System of Socialisation." In *Otto Neurath. Economic Writings: Selections 1904-1945*, edited by Thomas Uebel and Robert S. Cohen, 345-370. Dordrecht: Kluwer, 2004.
- Neurath, Otto. 1931. *Empirische Soziologie: Der wissenschaftliche Gehalt der Geschichte und Nationalökonomie*. Wien: Springer.
- Neurath, Otto. 1934. "Radical Physicalism and the 'Real World'." In *Otto Neurath: Philosophical Papers 1913-1946*, edited by Robert S. Cohen and Marie Neurath, 100-114. Dordrecht: D. Reidel, 1983.
- Neurath, Otto. (1931) 1973. "Empirical Sociology." In *Empiricism and Sociology*, edited and translated by Marie Neurath and Robert S. Cohen, 319-421. Dordrecht: Reidel.
- Norton, John. 1996. "Are Thought Experiments Just What You Thought?" *Canadian Journal of Philosophy* 26 (3): 333-366.
- Norton, John. 2004. "Why Thought Experiments do not Transcend Empiricism." In *Contemporary Debates in the Philosophy of Science*, edited by Christopher Hitchcock, 44-66. Oxford: Blackwell.
- O'Neill, John. 1998. *The Market: Ethics, Knowledge and Politics*. London: Routledge.
- Salerno, Joseph T. 1993. "Mises and Hayek Dehomogenized." *The Review of Austrian Economics* 6 (2): 113-146.
- Sigmund, Karl. 2017. *Exact Thinking in Demented Times: The Vienna Circle and the Epic Quest for the Foundations of Science*. New York, NY: Ingram.
- Sugden, Robert. 2013. "The Behavioural Economist and the Social Planner: To Whom Should Behavioural Welfare Economics Be Addressed?" *Inquiry* 56 (5): 519-538.
- Uebel, Thomas. 2018. "Calculation in Kind and Substantive Rationality: Neurath, Weber, Kapp." *History of Political Economy* 50 (2): 289-320.

Alexander Linsbichler obtained his PhD in philosophy at the University of Vienna. He is currently a lecturer at the departments of philosophy and economics at the University of Vienna. His previous academic positions include visiting research appointments at Duke University (twice), at Universidade Federal de Santa Catarina, and at the University of Manchester as well as teaching positions at Central European University and Vienna University of Technology. Alexander's research interests include history and philosophy of science, general philosophy of science, philosophy of economics, philosophy of logic, philosophy of mathematics, political philosophy, history of political economy, history of analytic philosophy, philosophical logic, and model theory.

Contact e-mail: <alexander.linsbichler@univie.ac.at>