ERASMUS JOURNAL FOR PHILOSOPHY AND ECONOMICS

VOLUME 13, ISSUE 2, WINTER 2020

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>.

ISSN: 1876-9098

ERASMUS JOURNAL FOR PHILOSOPHY AND ECONOMICS
VOLUME 13, ISSUE 2, WINTER 2020

TABLE OF CONTENTS

ARTICLE

BOOK SYMPOSIUM on *How We Cooperate?*

INTERVIEW

SPECIAL CONTRIBUTION

BOOK REVIEWS

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 1–29.
https://doi.org/10.23941/ejpe.v13i2.443

Moral Community and Moral Order: Developing Buchanan's Multilevel Social Contract Theory

JAMES CATON
North Dakota State University

Abstract: This work aligns James Buchanan's theory of social contract with the structure of Michael Moehler's multilevel social contract. Most importantly, this work develops Buchanan's notions of *moral community* and *moral order*. It identifies moral community as the vehicle of escape from *moral anarchy*, where community is established upon a system of rules akin to James Buchanan's first-stage social contract. Moral order establishes the baseline treatment of non-members by members of a moral community and also provides a minimum standard for resolving disputes that are not resolved by the more robust social contract shared among community members. This work links the multilevel contract to polycentric social order, noting that polycentric systems may promote development of the moral order by enabling experimentation with and emulation of rules and rule systems made available by overlapping and adjacent institutions.

Keywords: contractarianism, social contract, political economy, Michael Moehler, James M. Buchanan

JEL Classification: B13, B15, B25, B31, B52, P50, P51

In his recently published work, *Minimal Morality: A Multilevel Social Contract Theory*, Michael Moehler (2018) argues that James Buchanan's approach to the social contract cannot include significant moral diversity. This is because the second stage of Buchanan's social contract—post-constitutional exchange—depends on the normative content of the first-stage contract, making Buchanan's formulation inappropriate for a pluralistic society. Moehler believes that the multistage contract is incompatible with a multilevel contract—a contract that limits the moral demands of a particular community on non-members adjacent to or coexisting within that community. He critiques Buchanan's formulation of the social contract on

AUTHOR'S NOTE: This article was greatly benefited by the comments of the referees. The author is indebted to Tatiana Romanovskaya for useful comments, and encouragement. All mistakes are my own.

the grounds that the history of the first-stage contract cannot be scrutinized by agents subject to it. Moehler argues that for deeply diverse societies, the resolution of conflicts derived from historical injustice cannot be facilitated by Buchanan's multistage contract. For Moehler, Buchanan's theory grants a normative preference for the status quo that can lead to implacable conflict in a society with significant moral diversity.

Although Buchanan considers more complex characterizations of human behavior, the core of his analysis in *The Limits of Liberty* ([1975] 2000), hereafter referred to as *Limits*, demands only that instrumentally rational agents be capable of choosing to submit reciprocally to the demands of the social contract. Moehler establishes prior assumptions concerning the prudent behavior required for cooperation in a morally diverse society. Moehler's work concerns elements of justice he believes to be requisite for the healthy functioning of a diverse, liberal society. He describes an ideal moral order under the heading of the *weak principle of universalization*. This principle includes a basic income guarantee that supports bargaining above some minimal level of income by instrumentally moral agents who follow a weak Kantian imperative to solve conflicts peaceably. Moehler recognizes that assumptions made in application of his multilevel social contract theory may be different from those he prefers (Moehler 2018, 161–162, 181–184).

For example, Moehler chooses to resolve the potential conflicts concerning the legitimacy of the status quo by "the introduction of the unconditional subsistence income [that] represents a viable productivist policy that minimizes destructive actions, administration costs, and the costs associated with free riding" (2018, 200). He also recognizes that, more generally, so long as bargaining agents expect to be made better off by negotiations under the status quo, they "may agree to employ the existing status quo as a basis for conflict resolution in order to ensure the benefits of peaceful long-term cooperation at least in the future, as suggested by James Buchanan" (2018, 162).

Moehler's approach has much in common with Buchanan's framing. Both follow a contractarian approach. However, unlike Moehler, Buchanan intentionally avoids committing his agents to a Kantian categorical imperative. Instead, he develops an incentive-compatible escape from Hobbesian anarchy without deviating from the assumption of instrumental rationality *before* the social contract has been established. After establishment of the social contract that defines a community, members submit themselves to a structure of rules with the expectation of a reciprocal submission by others. Optimizing over a longer time horizon, instrumentally rational agents choose to operate under a civic morality (Congleton 2018). This reciprocal submission to given ethical and moral criteria cannot itself lead to an inclusive social contract amongst morally diverse agents. On these grounds, Moehler critiques Buchanan's ([1975] 2000) contractarian approach. *Limits*, however, was part of a more general project that started at least a decade before the publication of the book. This project concerned the social contract and individual ethics. Perhaps due to its nascency, Buchanan's later development of this project is omitted from Moehler's analysis.

Together with his earlier work, Buchanan's later work provides a multilevel theory of the social contract that employs a robust formulation of human agency. As in *Limits*, agents may follow rules that constrain their behavior, limiting short-term gains, with the expectation that others will follow the same set of rules. Buchanan even goes as far as to claim that the members of a *moral community* self-identify with their community and with the set of beliefs entailed in community membership. In other words, Buchanan claims, community members express collective intentionality (Searle 1995, 2005). These agents exist within exclusive moral communities supported by a civic morality (Buchanan 1965, [1981] 2001a). Each member acts in accordance with the community's social contract with the expectation that other members do the same. This is made possible by the exclusivity of membership. Buchanan also observes a more general *moral order*, but fails to explain how this moral order might arise.

I develop Buchanan's multilevel social contract theory in a manner coherent with his multistage contract, thereby showing that Buchanan's later work is compatible with his earlier work. In developing Buchanan's theory, I will show that it is possible to establish an evolutionary theory of the social contract that is compatible with rational choice theory, and that generates outcomes comparable to a Kantian approach while relying on less burdensome assumptions about human behavior. I will use Buchanan's multistage contract to explain the development of moral community.

Buchanan's ([1975] 2000) two-stage social contract uses rational choice theory to explain the formation of communities around an initial social contract coherent with the allocation of resources and power present in *moral anarchy*. Short-sighted utility maximization in Hobbesian anarchy gives way to a civic morality that includes a shared expectation of reciprocation among community members (Buchanan 1965, 3: social state 5; Congleton 2018, 40). When “two persons accept limits to their own freedom of action […] [t]he first leap out of the anarchistic jungle has been taken” (Buchanan [1975] 2000, 77). Buchanan scales this logic from agreement between two members to agreement between all community members.[1]

The same formulation holds for the development of moral order between and within moral communities that have otherwise escaped moral anarchy. This second level of the social contract develops out of the need to settle and minimize intercommunity conflict. This analytical nesting generates a multilevel theory that only requires a disposition toward rational norm-following of the community members and participants to conflict that cannot be settled according to a first-level social contract.

The argument will proceed as follows. First, I will introduce Buchanan’s multistage framework. I contextualize the multistage framework in light of its relevance to Buchanan’s work on moral community. I follow with a summary of Moehler’s multilevel framework. I argue that moral order develops and spreads along similar lines as a consequence of the interaction between moral communities. This nested analysis confronts Gaus’ argument that “[t]he rational strategy in large groups is to refrain from investing in norm change” (2018, 130). This is because the analysis takes moral communities as inputs that facilitate the development of a moral order. Viewing agent identity in the light of community membership allows for the conceptualization of a meta-social contract that binds interacting moral communities: the moral order.

BUCHANAN’S PATH OUT OF ANARCHY

> One essential problem that arises with Buchanan’s two-stage contract theory is that the normative content of the first-stage constitutional contract forms the basis for the second-stage post-constitutional contract. […] For Buchanan, this feature of his social contract theory is unproblematic, because ‘[t]he status quo defines that which exists. Hence, regardless of its history, it must be evaluated as if it were legitimate contractually.’ In other words, Buchanan’s political social contract theory simply assigns normative authority to the status quo, and, more importantly, makes the normative content of the second-stage contract dependent upon it.
>
> — Moehler (2018, 159)

[1] For larger communities, see Buchanan’s discussion of “Defection and Enforcement” ([1975] 2000, 83–88).

Buchanan ([1975] 2000) presents a two-stage social contract, noting that economic theory has typically dealt with the second stage, which he refers to as post-constitutional exchange. In the first stage, violence has not been contained by a social contract. Violence might be used to repossess resources from those less able to defend themselves, leading to costly investment in defense due to anticipation of predation (Buchanan [1975] 2000, 69–77). In the second stage, with ownership delineated, agents may engage in welfare-improving exchanges and agreements that are subject to the precedent constitution. Buchanan's goal was to provide an economic explanation for the development of institutions that undergird social cooperation and support the second-stage contract comprised of economic exchange. Here, I will concentrate on the first stage and its relation to moral community. In my later presentation, I will presume that the very first moral community must have developed by this process and that every other moral community either develops by this process, exists within and is supported by an existing moral community, or is the result of a split of an existing moral community.

The initial development of cooperation in *Limits* does not depend on a common moral frame. It begins in a world of Hobbesian anarchy where life is "nasty, brutish, and short" (Hobbes [1651] 1968, 186). In this setting, the strong may plunder the weak and battle amongst themselves. Agents are instrumentally rational, meaning that they are not constrained by an ethical disposition that is defined by a categorical imperative. Assuming that interactions are repeated, it may benefit both the plundered and the plundering to develop arrangements that make both better off (Olson 1993). Buchanan admits that, in this theoretical setting:

> The disarmament contract that may be negotiated may be something similar to the slave contract, in which the 'weak' agree to produce goods for the 'strong' in exchange for being allowed to retain something over and above bare subsistence, which they may be unable to secure in the anarchistic setting. (Buchanan [1975] 2000, 78)

Hobbes' solution to this dilemma is for the members of society to adopt a common morality—in the form of a universal social contract—and to submit to a sovereign who is tasked with the administration of that contract. Buchanan escapes the Hobbesian dilemma by a naturally occurring incentive structure. Like Hobbes' solution, *Limits* presents only the social contract—indicating a shared morality—of a single community. Absent this social contract, Hobbesian anarchy predominates. If agents were to return to such an anarchical state, they would operate according to purely instrumental rationality where behavior is only constrained by access to resources, especially for defense and coercion.

This is only the starting point for social relations. Once behavior is submitted to the rules of the social contract, opportunistic behavior that would threaten a return to anarchy is quelled by the expectation of mutual defection. As the human agent participates in a community, he or she necessarily chooses "between separate *rules* for behavior" and not "between separate *acts* in particular circumstance" (Buchanan 1965, 2). For Buchanan, an individual's choice to "adopt the moral law or the expediency criterion as an ethical rule surely depends upon his own predictions about the behavior of others" (1965, 2–3). In his description of the status quo, Buchanan recognizes that enforcement operates adjacent to "ethical constraints on individual behavior" ([1975] 2000, 99). The weaker those internal constrains, the greater the costs of enforcement required to correct behavior that deviates from the status quo. This can be corrected by bargaining over the status quo to align it with renegotiation expectations. The contract thus evolves with its standards being internalized by approving participants. Otherwise, an increased level of costly enforcement will be required to maintain the social contract and avoid a return to anarchy.

It is only by the development and sustainment of a shared understanding of one's position in the imminent hierarchy of social positions, and the *deontic powers* associated with these positions, that a network of actors can move out of Hobbesian anarchy.[2] Although Buchanan concentrates on a political theory of social contract in *Limits*, that discussion is supported by an ethical understanding of the human agent that is consistent with his earlier work (Buchanan 1965). To submit to a social contract, then, is to submit to a set of rules defining rights, duties, and obligations concerning one's role in society (Searle 1995, 2005, 2006). Development out of anarchy is facilitated by a mutual recognition of the social contract. The contract is held together, at least initially, by a commitment to a civic morality: a reciprocal expectation of commitment among community members (Buchanan 1965; Congleton 2018). Otherwise, cooperation might fail due to opportunistic behavior not bound by social rules. By common adherence to a civic morality, community members interact

[2] Just as Buchanan recognizes bargaining as playing a role in the evolution of a social contract, John Searle argues that a social contract exists anywhere "you have a community of people talking to each other, performing speech acts" (2005, 2). See also the *deontic operators* presented by Crawford and Ostrom (1995).

according to a shared structure of expectations that facilitates cooperation (Buchanan [1975] 2000, 114). As members begin to take for granted the status quo embodied in and including the social contract, the identity of members becomes bound in a moral community (Buchanan [1981] 2001a, 188). Mutual expectation increasingly takes the form of wholehearted submission to the community's social contract.

MORAL COMMUNITY AND MORAL ANARCHY

The initial stability provided by Buchanan's formulation of the social contract in *Limits* enables the development of moral community as it forms a basis for bargaining within the contract.[3] Moral community logically precedes moral order. This being the case, it is not surprising that Buchanan did not distinguish between moral community and moral order until several years after *Limits* was published. To clarify the meaning of moral order, which I later elaborate, it will help to first distinguish between moral anarchy and moral community.

Taken as a positive demonstration, rather than a normative formulation, the problem is precisely how one might construct a theoretical escape from moral anarchy. Not only must this description allow for an escape, it must also explain how the social contract keeps moral anarchy at bay. In *Limits,* moral anarchy is overcome by Buchanan's development of incentive-compatible equilibrium arrangements. Once instrumentally rational agents develop incentive-compatible relations amongst themselves, the long-term result is the development of shared practices that lead agents to resist change in strategy—unlike instrumentally rational agents whose mode of behavior is not constrained by a shared structure of rules. Strategy bounded by social rules develops through a process of bargaining (Bourdieu 1990, 122–134). When participants successfully bargain, the status quo is moved closer to "renegotiation expectations"—terms that negotiating parties will voluntarily accept—making the adherence to the social contract less costly and, therefore, less dependent upon enforcement via coercion (Buchanan [1975] 2000, 98).

Buchanan's price-theoretic analysis bootstraps the development of moral community out of moral anarchy. Once a shared rule structure has stabilized through the development of incentive-compatible strategies, that structure guides behavior within the moral community. Once the bargaining over rules by instrumentally rational agents has led to a

[3] At this point of the development of his theory, Buchanan used the term 'community' without the descriptor 'moral'.

satisfactory arrangement given access to violence and other resources, community is the inevitable result. Community members submit their rationality to the shared rules and values of the community.[4] These agents optimize, but they do so within the constraints of the shared frame of their moral community and the margin of influence that they may manifest over that framework by renegotiation, whether formal or informal.

Membership in the community may thus be valued as an approximation of the expected benefits from the stability of community structure. A conception of the common good is tangible at the level of the moral community. Members share a common conception of the deontic powers associated with membership, identifying themselves with the community (Buchanan [1981] 2001a, 188; Searle 2005, 2006). The result is that inside the moral community, members form a collective ensemble by submission to a robust shared rule structure that coalesces with their private lives. The good of the members is aligned and even identified with the good of the community. Members procure a bundle of goods for which they are willing to incur the costs of membership that entails shared beliefs and behavioral constraints. The social space between communities, at worst, exists as a moral anarchy when there is no shared moral structure aside from the null set and, at best, operates as a moral order with norms shared between communities, analogous to Moehler's *minimal morality*.

MOEHLER'S MULTILEVEL SOCIAL CONTRACT

Moehler critiques Buchanan on grounds that a multistage social contract depends upon a single shared morality. In modern liberal societies, we interact with individuals holding diverse beliefs and moral commitments. In these societies, individuals cannot demand or expect acceptance of the full set of their own beliefs in interactions with those who are not members of their community. Under these circumstances, the Hobbesian solution of a single morality and a single sovereign is insufficient. This may seem innocuous, but some beliefs concerning justice may be conflicting. Having only a single social contract—as opposed to allowing for multiple

[4] In this respect, agents who operate within the rule structure of their community benefit in a manner similar to Gigerenzer and Brighton's (2009) description of the *homo heuristicus* agent who intentionally ignores some information. Like Smith (2003) and Dekker and Remic (2019), the approach here concentrates on shared rules. Agents within a community intentionally ignore some strategies as those strategies would conflict with membership within a community: the loss of membership would deprive the agent of the bundle of goods made accessible by membership.

instances in the form of distinct communities—does not allow for the development of inclusive moral commitments that transcend a particular community and facilitate conflict resolution through bargaining between agents with distinct, and perhaps opposing, beliefs.

Moehler's multilevel theory, for the second-level contract, relies on one core principle (the weak principle of universalization). This principle "constitutes the 'second contract' into which the members of society enter" where agents strive for peaceable conflict resolution according to a "morality in the form of 'each according to her basic needs and above this level according to her relative bargaining power'" (Moehler 2018, 18). The weak principle entails two conditions:

> *First*, it [the social contract theory] must ensure that, in each instance, agents can defend their interests maximally based on their actual capacities in the world in which they live, ensuring that agents receive a share of the goods in dispute that is proportional to their relative bargaining power. *Second*, and as a potential constraint on such behavior, it must ensure that agents can maintain their existence as separate agents and satisfy their basic human needs as a basis for conflict resolution (minimum standard of living), if the goods that are in dispute permit it. Any viable principle of conflict resolution that can ensure stable peaceful long-term cooperation among rational prudential agents in the real world must satisfy these conditions. (Moehler 2020, 49–50; emphasis mine)

The first condition of the principle is compatible with Buchanan's formulation of the social contract in *Limits*. The resultant distribution depends upon the resources controlled by each agent and the judgment employed over these resources.[5] Moehler's second condition is more demanding than Buchanan's framework. Buchanan's bargaining agents operate initially using instrumental rationality. They are capable of threatening a return to moral anarchy to increase their leverage in bargaining. Moehler's agents, in contradistinction to this possibility, use instrumental morality where members "have an overarching interest in ensuring peaceful long-term cooperation" (2020, 57). Any threat to return to moral anarchy lies outside the bounds of the weak principle of universalization.

Moehler's weak principle of universalization is a modified Kantian categorical imperative. He refers to those acting according to this principle as *homo prudens*. The weak principle must be accepted by all members

[5] Here, I am drawing specifically from Frank Knight's (1921) emphasis on the entrepreneur's exercise of judgment over resources in his control. Not coincidentally, Knight was Buchanan's advisor during his graduate studies (Wagner 2017).

of society to maintain peace. The theoretical limitations of the instrumentally moral agent constrain the damage that can be done by Hobbes' *Foole* (Gaus 2013; Moehler 2014). At the first level of morality, agents are presumed to have internalized a robust local morality. At the second level of morality, which handles cases where an accepted set of solutions has not yet been developed, instrumentally moral agents bargain under only the minimal constraint of Moehler's weak principle of universalization (Moehler 2018).

Moehler's second level operates in two scenarios. In the first case, two members of the same moral community, to use Buchanan's language, may be involved in conflict that cannot be straightforwardly solved by the community's social contract. The thick morality of the community simply cannot be applied to this category of cases. The development of a solution to such conflict entails an attempt by either party to attain maximum value given the minimal constraint provided by Moehler. This is simple enough to imagine within a given community where all actors share a common moral frame. And, in the real world, they would be nested in a common community of interacting individuals capable of observing the interaction and of forming judgments in light of the strategies used by the other agents. These judgments would then frame future interactions between these bargaining individuals and the rest of the community.

The more difficult scenario occurs when conflict exists between members of different communities. The weak principle of universalization places constraints on the extent to which two interacting communities might differ. If, for example, one or both individuals pertinent to conflict are from a community where "eradication of human beings or certain members of society [is believed] to be the overarching goal", then Moehler's theory "could not harmonize the interactions among agents on terms which all members of society could agree" (2020, 60). In order for the theory to be applicable, an action from any participant must be bound by the weak principle of universalization.

This highlights a significant distinction between Buchanan and Moehler. Buchanan concerns himself with harmonization that results from bargaining alone. Buchanan's purpose is not to present a theory of justice, even if one is implied by his theory's acceptance of the status quo. Outcomes quite often will cohere with the end state promoted by the weak principle of universalization. Yet for Buchanan, if a stronger party does not submit his or her behavior to the contract of a moral community and if he or she expects no net benefit from establishing any sort of cooperation with a weaker party, then actors remain in moral anarchy. In the final state of analysis, no cooperation occurs. Whatever plundering or murdering a sufficiently strong party had intended will occur so long as the expected benefits of the action exceed the expected costs. Before an incentive-compatible social contract is developed, the theorist cannot, as the saying goes, dispute preferences (Stigler and Becker 1977).

In reflecting upon institutional change, Buchanan observed:

> In economists' terminology, institutional-constitutional change operates upon the constraints within which persons maximize their own utilities; such change does not require that there be major shifts in the utility functions themselves. (Buchanan [1981] 2001a, 201)

Of course, Buchanan recognized elsewhere that individuals can engage in *personal* development that might even include radical transformation (Buchanan [1979] 1999), but his concern in his work on the social contract was *institutional* development. After interacting agents develop a social contract where the more powerful agree to constrain their capacity for violence—or concomitant with that development—Buchanan's agents take on a substantive ethical dimension. Then, community members can generate moral and constitutional artifacts that members can reflect and act upon in effort to transform the community, its purpose, and their roles in it. By this process moral diversity can be supported by the development of a moral order.

BUCHANAN, MOEHLER, AND KANTIAN COMMITMENT

Buchanan's formulation of the social contract omits the weak principle of universalization. Throughout his work, Buchanan intentionally avoids such a Kantian commitment when explaining the development and sustainment of social cooperation (1965). Buchanan describes an ongoing bargaining process that motivates buy-in from those self-interested actors subject to the social contract who may consider moral anarchy—a state where individuals are treated as means to ends, with no moral parameters constraining this treatment—as a viable alternative to the status quo.

Buchanan intentionally avoids "external ethical criteria [...] imposed on the existing structure" that "tended to distract effort and attention from the less romantic but more productive approach involved in working out possible compromise modifications that would be agreeable to large numbers of persons in the community" ([1975] 2000, 111). A return to moral anarchy is included in the option set available to Buchanan's agents in the bargaining process. It is a risk that must be considered by all parties involved. From this vantage point, history is littered with constitutional moments that take the status quo for granted but, by the very existence of bargaining over the social contract, do not treat it as immutable. Rather than demand historical justice, Buchanan's agents accept that they can, at best, express influence over the evolution of the social contract.

Moehler critiques this approach by Buchanan. Moehler's agents demand recompense for injustice across generations, as this affects the incumbent distribution of resources. Otherwise, they might also consider the return to anarchy a viable option, which would violate the weak principle of universalization. Buchanan's multilevel social contract does not directly provide the justice demanded by Moehler. It does not necessarily forbid such attempts, but provides them no special status in the bargaining process. Taken as a whole, both theories exhibit significant overlap. Moehler recognizes many of the features presented by Buchanan, viewing them as cases where the weak principle of universalization is violated.

With the purpose of describing social evolution in mind, Buchanan argues that gains from peace may be valued independently from social history and may themselves be sufficient to offset animosity derived from an initial injustice. The only requirement of the contract is that agents value their own positions—with those positions' incumbent mix of wealth, rights, and duties—well enough to temper each other's demands for historical justice. This does not, however, prevent agents from presenting utility maximizing demands in the form of claims about justice. These claims may demand a social contract with a set of rights "insupportable in anything that might resemble genuine anarchistic struggle", and so "when presented under the disguise of justice [modifications to the social contract] tend to attract support from those elements of the community whose primary motivation is to arrange preferred redistributions of rights among *others*" (Buchanan [1975] 2000, 104). For Buchanan, this demand for historical justice may influence the development of a social contract, but this is not an inevitable outcome. Moehler's second level strictly demands that the weak principle of universalization is adhered to. He asserts that "[i]f the members of society do not regard the status quo to be justified, then they may demand compensation first before they fully accept the demands of the weak principle of universalization" (Moehler 2018, 162). For Moehler, an initial administration of distributive justice via something comparable to a basic income guarantee is a more obvious path to maintaining submission to the weak principle, which supports the peace required for a liberal society.

SOCIAL CONTRACT: STAGES *AND* LEVELS

Moehler's more serious concern with Buchanan's framework is that it lacked a well-developed theory of a multilevel social contract comparable to Buchanan's two-stage social contract theory. Presumably, this is the reason for Moehler's focus on the framework presented in *Limits*. Although a multilevel social contract is absent from *Limits*, multilevel analysis appears several years later in Buchanan's Abbot Memorial Lecture, "Moral Community, Moral Order, and Moral Anarchy" ([1981] 2001a) and in the shorter "Moral Community and Moral Order: The Intensive and Extensive Limits of Interaction" ([1983] 2001b). In the first of these lectures, Buchanan acknowledges in a footnote ([1981] 2001a, 187n1) that the work contributes to the same project as *Limits*. Buchanan's work presents a *positive*, rather than a *normative*, multilevel theory of social contract.

One might argue that Moehler's theory indicates the bounds within which a social contract may operate absent resort to unsanctioned violence. Moehler's formulation concerning the requirements for a society of *homo prudens*—provision of a minimum level of income—goes beyond a general description of the problem. It cannot consider situations where, faced with a decision between significant loss—for example, death—and rebellion against the standards set by the social contract, individuals may well choose rebellion.

Still, there is no escaping the economic logic of anarchy without accepting the status quo as a frame of reference in Buchanan's framework. As Moehler points out, Buchanan implicitly "assigns normative authority to the status quo" that is generated from the initial distribution of resources in anarchy (Moehler 2018, 159). He does not, as Moehler, assert the principles by which a pluralist liberal moral order might be peacefully sustained. Rather, he explains how social order might arise from moral anarchy.

Next, I will elaborate Buchanan's theory in light of the rich structure provided by Moehler. Unlike Moehler, I will not emphasize the weak principle of universalization. Like Buchanan, I presume that return to moral anarchy is always an option for individuals who bargain over the social contract. In doing this, I present a theory of a process that describes the move from moral anarchy to a world with moral community and, eventually, moral order.

Buchanan defines the moral order in terms of binary interactions and relations—equality defined by abstract rules—as opposed to roles whose deontic powers stem from the hierarchy of the moral community. Under the moral order, individuals regard one another as legal and moral equals. The moral order necessarily represents the minimum standard by which agents from differing communities interact. It is analogous to Moehler's second-level social contract, a minimal morality, and includes Moehler's particular formulation as *one* possible manifestation of the moral order.

Buchanan elaborates the system of rules defining a moral order using his visit to Austria as an example:

> I did not qualify for membership in the Viennese or Austrian moral community at all. But I was able to survive well by a knowledge of and adherence to a system of rules that involved a mutual respect for the rights of property, that of my own and those of persons with whom I had dealing! It is easy to imagine the difficulties I might have encountered in a genuinely 'foreign' land that was not characterized by such agreed-on rules of behavior and in which, quite literally, I should have to depend upon the genuine 'morality' of others to survive. (Buchanan [1983] 2001b, 209)

The moral order is the domain of interaction subject to the minimal set of rules of the second-level social contract. This includes not only rules, but also the expectations derived from these rules, and the interaction facilitated by this expectation. The moral order is distinct from moral community in that the rules that support the moral order allow individuals to "treat each other as moral reciprocals" (Buchanan [1981] 2001a, 189). Without a shared moral order that allows for peaceable interaction amongst relative strangers, one is entirely dependent upon the moral attitude of community members toward outsiders.

Buchanan's notion of moral order must be developed with special reference to the robust description provided in *Limits.* All that is required for a particular instance of the social contract to be adopted is "that this assignment is mutually accepted" so that "mutual gains may be secured from the consequent reduction in defense and predation effort" (Buchanan [1975] 2000, 78). As Gerald Gaus observes, "Hobbes's problem remains our problem, even if we recoil at his solution" (2013, 278). That problem is to "resolve the 'foundational crisis' of morality" (D'Agostino, Gaus, and Thrasher 2017). This is true for the social contract governing behavior within *and* between communities. When individuals from different communities interact, there exists a greater possibility of opportunism among bargaining parties. Without something like the weak principle of universalization shared between bargaining parties, there is great potential for theft and destruction. The moral order is a shared social scaffolding that mitigates the occurrence of opportunism, which would otherwise occur in a moral anarchy.

Moral order can only develop once moral communities have been formed by the process outlined above. The development of the moral community informs our understanding of the moral order. Where Moehler proposes that the weak principle of universalization cannot be derived from the first-level contract, my framework holds a shared minimal morality—Buchanan's moral order—as an *artifact* of moral communities. The treatment of community members in cases where the social contract of the moral community cannot facilitate conflict resolution provides a baseline for the treatment of outsiders by community members.

As in Moehler's formulation, agents subject to a first-level contract—members of a moral community—share a thick moral frame. The second level exists, as Moehler suggests, in cases where the social contract of a given moral community is unable to solve a conflict because (1) the contract is ill-suited to resolve the conflict, (2) the agents in disagreement are subject to two distinct social contracts whose dissimilarities do not allow either contract to facilitate sufficient resolution of the disagreement, or (3) one of the agents subject to the conflict has no moral community and the thick moral frame of the other agent does not present a solution acceptable or applicable to the non-member (see Figure 1).

Moehler asserts that the second-level contract cannot be derived from the first-level contract. Otherwise, conflicts that are not facilitated by the first-level contract would lead to resolution consistent with anarchy. This claim can be clarified by elaboration on case (1) above. If the membership in a moral community is itself valued by the bargaining parties, then norms within the community concerning violence among members will at least be submitted to by bargaining parties. The solution reached will be subject to at least the most primitive confines of the community's social contract. Thus, bargaining under conditions of ambiguity by members who value their positions will likely develop the community's social contract subject to Moehler's weak principle of universalization.

Second-level morality is relevant for intra-communal conflict that cannot be resolved by first-level morality. It represents a baseline for the treatment of all members within that community in cases where the

TYPES OF CONFLICT

CASE	INTERACTION TYPE	DESCRIPTION	RESULT
(1)	Intra-communal	Conflict between community members not resolved by the existing terms of a thick shared moral frame.	Defer to the community's most primitive principles supporting conflict resolution.
(2)	Inter-communal	Conflict between members of different communities not resolved due to a lack of thick shared moral frame.	Defer to principles commonly held between communities.
(3)	Extra-communal	Conflict between a community member and an individual who is not a member of any community. The non-member lacks any thick moral frame.	Defer to the community's standard treatment of non-members.

Figure 1: Three cases where the social contract of a moral community cannot resolve a conflict.

shared thick moral frame fails to facilitate the resolution of conflict. Likewise, second-level morality indicates the standard treatment of non-members. In this sense, the moral order promoted by a community is endogenous to its most basic shared moral presuppositions. If conflict is resolved by violence, on the other hand, then the community has either failed to consult its most essential moral presuppositions or those presuppositions accept subjugation of the weak by the strong.

Case (2) is more difficult. Its resolution can inform case (1). In the first case, members bargain over the structure of the social contract and each member's position under it. Membership in the same moral community facilitates bargaining as each member accepts the existing contract as status quo. The transformation of the social contract prevents conflicts from leading to deterioration of the social contract. Disregard for rules that bind the members' behavior, especially in regard to unsanctioned violence—that is, violence not condoned by the social contract, especially not exercised under particular circumstances that legitimate its use—threaten the integrity of the social contract. When conflicting parties are not members of the same community, they do not necessarily have the same primitive set of moral presuppositions upon which to rely in forming expectations. This is because these conflicting parties do not share the same thick moral frame. If both parties share a thin moral frame in the form of a minimal set of moral presuppositions, then they could engage in bargaining over these terms. The development of a solution provides a basis for future cases of conflict resolution between these two communities.

Buchanan reflects on this distinction by noting that within a moral community that lacks a moral order, "[s]ince the individual person in such a setting thinks of himself as a member of this community rather than as an individual, he will more readily acquiesce in what would seem overtly unfair treatment under the moral order" ([1981] 2001a, 195). If the interaction within a moral community is not also framed by the impersonal standards of the moral order, the boundaries of the moral community are moral anarchy, providing all the more reason to acquiesce to what those of us with modern sensibilities would consider violations of, for example, the rule of law.

The second level also reflects the treatment of non-members acting within existing moral communities who together participate in a shared moral order. Non-members have no position within the moral community's hierarchy and therefore—when attempting to order their interaction with community members—are benefited neither by an ongoing conversation within the moral hierarchy, nor by a robust, shared moral frame. A community's treatment of non-members reflects the community's most basic moral presuppositions concerning the interaction with other human agents and, as mentioned above, likely indicates the bounds of treatment that members are willing to endure for the good of the community. In the case of Moehler's framework, for example, communities that also accept the weak principle of universalization for conflict not successfully mediated at the first level are protected from moral anarchy by the second-level social contract. Members *and* non-members are afforded this protection.

Difficulty occurs, however, if bargaining at this second level breaks down. Under conditions where parties are unable to agree on the terms of resolution, dispute may descend into violence. Similarly, Moehler defines his principle as 'weak' because he acknowledges that members of a community may find themselves in conflict with non-members, despite a preference for the opposite, if violence from the outside party cannot be effectively mitigated. This is consistent with Moehler's concern that unwillingness to participate in second-level bargaining by disgruntled community members could lead to a breakdown in the social contract, say, by violent revolution. Neither does Buchanan provide a *Foole*-proof remedy for this problem except to note, like Moehler, that the development of the social contract is bound by the welfare outcomes expected among members given a return to moral anarchy. In order to avoid moral anarchy, both participants to a conflict must expect that they can be made better off by peaceful negotiation than by initiating or continuing a pattern of violence. So long as both parties are committed to engaging in mutually beneficial interaction, Moehler's weak principle of universalization holds. Barring mutual acceptance of the weak principle, the stronger party might maintain the moral order by attempting to mitigate the destruction of non-cooperators, perhaps by promoting an institutional transformation that also transforms behavior of the defecting parties. The moral order is maintained.

Case (3) includes the interaction of a community member with an individual who is not a member of any community (extra-communal interaction). Once either a single community identifies the fundamental principles that guide the development of its social contract, or diverse communities develop a shared moral order that enables members of these communities to interact with one another, it is possible for individuals to escape moral anarchy without belonging to a moral community. These agents, who are not members of any community, freeride on the moral order developed by existing moral communities.

MORAL COMMUNITY WITHOUT MORAL ORDER

We can imagine a case where a moral community exists around a social contract that has fully mitigated violence for cases handled by the social contract but not necessarily for those outside it. Suppose that feuding individuals resort to violence when agreement cannot be reached via the social contract. In this case, there exists no moral order. The world outside of the bounds of the social contract exists in moral anarchy. We might call such a community a *predatory community*.[6] Such a community has not developed a belief that human life, let alone human liberty, ought not to be aggressed against without just cause. That is, community members in their conflicts among one another violate the weak principle of universalization. Parts of the domain of this community reaches into the depths of moral anarchy.

Since such a community defies our modern sensibilities, it is useful to include an example. Peter T. Leeson (2014) describes such a community

[6] One might prefer the less affective 'amoral community', however the author interprets this term as deceptively neutral. Such a community is not amoral in the objective sense since, as Buchanan describes, it involves "the ways that persons act and feel toward one another" ([1981] 2001a, 187). Neither is the community 'amoral' in the normative sense, if we mean 'amoral' to have a meaning distinct from and more favorable than 'immoral'.

in the example of the Indian Khond society. Leeson argues that the system of ritualized human sacrifice that existed in this society served as a means to wealth destruction, and that this system prevented intercommunity conflict by limiting incentives for intercommunal plundering. British Major Samuel Macpherson observed that while the system performed well in providing "order and security" within each community, "beyond all [communities] is discord and confusion" (Macpherson 1865, 81; cited in Leeson 2014, 149–150). This should be no surprise since, conceptually, moral anarchy forms the boundaries of such a community. The subject of Khondian human sacrifice by a given tribe was usually not a member of the tribe, but there was no restriction upon sacrificing even a member of one's own tribe.

Leeson gives the Khond society as an example where rational choice theory explains why a society might fail to escape from a suboptimal equilibrium. The system had lasted, Leeson argues, because "it was also socially productive" in that "the wealth lost in violent clashes without human sacrifice exceeds that which is destroyed via human sacrifice" (2004, 162–163). In this presentation, accumulation of wealth is an attractor for conflict from other tribes. While it might have been possible for one tribe to dominate the others under different circumstances, the system of human sacrifice led to the exportation of wealth. The sacrificial subject, a *meriah*, was purchased for the purpose of human sacrifice. Most often, these individuals were not from a tribe in the Khond society. Thus, the system led to an outflow of wealth that limited incentives for, and therefore the level of, intertribal plundering.

This incentive compatibility is necessary to explain the functioning of the Khond society, but it is not sufficient for this purpose. The system of human sacrifice was deeply embedded in the social contract in the form of religious beliefs and practices:

> Konds believed their fate rested in the hands of *Tari Penu* — the malevolent earth goddess to whom they offered *meriahs*. To 'obtain abundant crops, to avert calamity, and to insure prosperity in every way' they required her favor. *Tari* craved the blood of sacrificial human victims and 'caused all kinds of afflictions and death if she was not satisfied,' most notably 'through war and natural calamities'. (Leeson 2014, 158)

Participation in cultural and religious practices within Khond society reinforced the social contract that "underlay a close identity between the ecclesiastical and temporal interests of the tribesmen" (Gangte 2017, 116). Priests participated directly while other members of the Khond society contributed to the purchase of the *meriah*. Once the sacrifice was complete "the crowd would rush to the victim and stripped the flesh from his bones" with the intention of mixing the flesh of the sacrificed with the soil where the tribe planted (Gangte 2017, 117).

While rational choice theory helps us to explain why heinous equilibria might emerge, additional tools are required to understand how such a system fits within the general structure of a social contract. A theory of *morals by agreement* should be capable of explaining the liberal organization described by Moehler, as much as it should be capable of situating the system of Khond society, even if such a society represents a failure in moral development. The social contract commonly allowed for each party's access to force to adjudicate intertribal conflict. Moehler's normative approach prevents such an application since a society organized around a system of human sacrifice considered "the eradication of human beings or certain members of society to be the overarching goal" (Moehler 2020, 60).

Considering his purpose, Moehler is correct. However, the framework elaborated here can still bring into clearer view the structure of such a society and how a moral community might evolve out of such a local equilibrium. Moral order is indicative of the lower bound of treatment between interacting moral communities, with the null set being moral anarchy. The lower bound for treatment of those falling outside the protection of the social contract took the form of sacrifice of innocent non-members. Although "[i]n practice they were nearly always non-Konds" (Leeson 2014, 151), by definition, the standard of treatment of non-members also indicated the lower bound of treatment for members. The same second-level contract governing treatment of non-members also mediates conflict not resolved by the first-level contract. Buchanan defines moral anarchy as a setting where "each person treats other persons exclusively as means to further his own ends or objectives" ([1981] 2001a, 190). The systematic sacrifice of humans and, potentially, even community members for maintenance of the social order falls within this definition.

The elements of moral anarchy present in each tribe's social contract introduced a moral chasm between the communities in Khond society, and an insecurity that could potentially threaten the members of a tribe since, as Leeson notes, "[i]n principle *meriahs* could be persons of any age, sex, race, or caste" (2014, 151). The sacrifices came from outside the society, Leeson argues, because the stability provided by this system occurred as a result of wealth destruction that disincentivized war. The system didn't extinguish moral anarchy; it simply mitigated its detrimental effects by means of wealth destruction while still allowing for behavior consistent with moral anarchy under particular circumstances.

Khond society was only able to exit the equilibrium when something analogous to the weak principle of universalization was provided by importation of the British legal system. In this case, the weak principle of universalization that alleviated intertribal conflict was not derived from or developed within an existing social contract. Macpherson coordinated a new arrangement between several tribes where British authorities offered to administer justice. He had observed that the Khonds "most anxiously desire of us justice — not betwixt man and man, which their own institutions can afford, but betwixt tribes and their divisions" (Macpherson 1865, 178; cited in Leeson 2014, 161). He offered a substitute for the system of human sacrifice. Intertribal conflict no longer needed to be governed by a system of vying alliances threatening and engaging in war. Instead, the tribes by mutually submitting to British legal rule could abandon both war and the system of human sacrifice that indicated an absence of moral order among and between tribal communities.[7]

We observe how emulating principles and experimenting with their application can improve the functioning of societies. While the development of moral order could have been applied to just a single tribe, the ability of that order to govern interaction between communities required that at least two tribes agree to change their manner of interaction with one another (Vanberg and Buchanan 1988, 152). A society's exit from the system dependent upon human sacrifice required that another means be substituted for maintaining order between tribes. Macpherson offered British legal administration for a small number of tribes willing to exit the system. Participants in this experiment received protection from intercommunal aggression and were therefore able to opt out of the wealth destruction entailed in the system of human sacrifice.

The only means of maintaining moral diversity in the face of a *community* that systematically implements and approves of aggression against innocent members is for a competing system of morals to be capable of withstanding the exercise of force from that community. Otherwise, moral diversity may be extinguished by parties with access to violence. Macpherson was able to offer this option to interested tribes

[7] These Khond tribes preferred to import the British legal system instead of relying on the status quo system that depended upon human sacrifice.

because he drew his power from outside the Khond society. The potential span of association and cooperation improved as a result. Once several communities succeeded in improving their relations by this means, "soon other communities 'spontaneously proffered to relinquish the sacrifice, mainly on the condition of obtaining protection and justice, and actually pledged themselves accordingly'" (C. R. 1848, 275; cited in Leeson 2014, 161–162). Bargaining at the second level of the social contract was initially facilitated by Macpherson, and was quickly internalized throughout Khond society, transforming their moral communities in the process.

Multilevel and Polycentric Orders

The example of the transformation of the Khond society also illustrates how the multilevel and polycentric frameworks inform one another. British actors and legal elements successfully interacted with the Khond society in a manner that led to its transformation. That is, British institutions outcompeted existing endogenously formed institutions (Boettke, Coyne, and Leeson 2008). The presence of alternatives allowed for Khond tribes to adopt an alternate system for adjudicating conflict, especially intertribal conflict. The resultant moral order allowed individuals to engage one another with the expectation that plundering was no longer an option.

Moehler notes that inconsistency is bound to be present within a polycentric order and finds this problematic if there is to exist a coherent, well-functioning multilevel social contract:

> As a result of such abstraction, the agents may not understand the relevance and normative force of the principles justified, as is often suggested with regard to Kant's categorical imperative, which leads to a problem of (in)stability. Second, if the inhabitants of society are held constant, then moral rules can be justified that are valid only for certain subgroups of society, which leads to a polycentric moral order with restricted although potentially partially overlapping jurisdictions. Conceptually, *such a polycentric moral order cannot ensure stability of cooperation because, in the worst case, moral interactions may arise for which no moral rules are justified for all parties to a conflict,* in particular if the parties belong to different subgroups of society. (Moehler 2020, 45; emphasis mine)

Macpherson's discomfort with the Khonds and his attempts to curtail the system of human sacrifice indicate the incompatibility of British institutions with Khondian predatory communities. Moehler hints at a way of resolving this tension as “conceptually it [the polycentric order] is merely an intermediate step for defining a moral system that can ensure stability of cooperation in deeply morally diverse societies” (2020, 45–46). It is by no means an insignificant step in analysis to provide a path to agreement. This is part of my intention in developing Buchanan’s framework. A polycentric system provides a more reliable basis for allowing communities to develop and adopt criteria that promote a moral order and, thus, a multilevel social contract. The growing overlap between the Khond society and British practices led to a transformation of moral communities in the Khond society away from its status quo by integration of an Anglo moral order. The piecemeal development of a liberal order by the Khonds would have been highly unlikely given their starting point. Predatory communities have no obvious incentive to maintain moral diversity in the case of conflict with morally diverse agents. Key to the transformation, a small number of tribes, with protection from the British, succeeded in *exiting* the system of human sacrifice (Kukathas 2003), serving as exemplar for other tribes that wished to emulate their integration of British legal institutions.[8]

This sort of integration highlights the manner in which members of a society might adopt a system of rules that, implicitly or explicitly, coheres with Moehler’s weak principle of universalization. Without identifying a process that tends to lead the development of the social contract in the direction of the weak principle, such an analysis is subject to uncertainty that could unnecessarily limit its usefulness. Van Schoelandt (2019) notes that the existence of overlapping jurisdictions in a polycentric order could facilitate the adoption of second-level morality when first-level morality fails, and vice versa.[9] Noting this, Moehler reflects that his theory applies to agents “who have, all things considered, an overarching interest in securing peaceful long-term cooperation” (2018, 18). This outcome seems, to this author, akin to an equilibrium state. Allowing polycentric order to inform the development of the social contract illustrates how such a state might be reached.

[8] Dekker (2016) and Dekker and Kuchař (2016) refer to tradeable exemplary goods. One might, in line with their following of Hannah Arendt, think of either (1) the British legal system as an exemplary institution, or (2) the tribes that successfully integrate the system as exemplary communities for other Khond tribes. Similarly, instead of subjecting civil relations to Sharia law, Qatar maintained separate civil courts modelled after the ‘Romano-Germanic’ system upon the exit of the British in 1971 (Hamzeh 1994).

[9] On polycentricity and political organization, see also Polanyi ([1951] 1998), Ostrom ([1991] 2014), Aligica (2014), and Aligica and Tarko (2012).

Finding what these agreements should be and how they can be structured to be inclusive may not be a straightforward process. By noting the possibility of overlapping social contracts associated with a polycentric order—that is, a morally diverse social order—a multilevel theory of social contract allows for a process of experimentation and emulation in regard to rules and rule structures. A polycentric order increases the number of combinations that might be tested by a community and adopted in the social contract. Presuming that Moehler's instrumentally moral agents are looking for cost minimizing means of resolving conflict, these agents will search through this combinatorial space in order to find or generate rule structures that can facilitate resolution. As we observed with the development of the Khond society, the existence of or potential for overlapping social contracts can enable a society to exit a suboptimal equilibrium where conflicts cannot be solved according to the weak principle of universalization. It also provides reason to temper Buchanan's pessimistic concern, in reflecting upon social conditions in the United States, that the moral order might unravel (Buchanan [1981] 2001a, 196–198).

The multilevel framework, with explicit inclusion of polycentric order, allows for an explanation of how certain institutions might spread across diverse societies and, in the process, constructively interact with social contracts of diverse communities. One might use the framework to analyze, for example, different episodes in European history, including: the spread of Roman law in diverse social orders under the Roman Principate; the role of Catholic institutions in maintaining the remnants of that law for communities across Western Europe during the Dark Ages; and the significance of European legal fragmentation in facilitating the Protestant Reformation or the liberty required for post-Enlightenment intellectual developments. Although Moehler's intention was to provide a theory that explains how a morally pluralistic society with "agents [who] may hold irreconcilable moral ideals" (2018, 1) can function, a more general interpretation of the multilevel framework, otherwise compatible with Moehler's framework, sheds light on the source of social dysfunction and the path to ameliorating that dysfunction without presuming a sole Hobbesian sovereign.

CONCLUSION

While it is not correct to claim that every moral community must develop out of moral anarchy, the moral community solves the problem of moral anarchy and some moral community must solve this problem *before* a moral order develops.[10] Here Buchanan diverges from Moehler as Moehler's intention is not to provide an evolutionary account of the formation of social contract. Still, Moehler recognizes alternative cases that do not fall within the purview of his own analysis.

In *Limits*, Buchanan presents the hardest case. Outside of moral communities, in a world absent moral order, agreements depend purely upon the incentive structure present in moral anarchy. The rationality of the initial set of agreements is eventually embedded in the artifacts of rules and beliefs of the moral community that emerge from it.[11] Absent the morality of the community, what predominates is a moral anarchy whose outcomes are guided directly by access to violence and which is comprised of instrumentally rational agents lacking a common moral frame distinct from the null set. Absent a shared moral order, moral communities that are entirely distinct in terms of overlapping membership must follow a similar course of development described in the initial formation of a community from moral anarchy. Lacking a shared moral order that attributes worth even to non-members, a community may not qualify for the descriptor 'moral', at least not in the strict sense. Interaction between communities that lack a shared moral order occurs in a sea of moral anarchy where conflict may swiftly descend into violence.

The moral community eliminates moral anarchy within the confines of the community. When a minimal set of shared norms neither exists, nor is developed between communities, the moral order collapses into a moral anarchy (Munger 2020) that will be present in any conflict not resolved by the first-level contract. Such disagreements are resolved by might, as this is the nature of human relations in moral anarchy. Moral anarchy predominates if the moral order deteriorates, or never existed, between communities lacking tight overlap. Moral anarchy may even be a feature in certain corners of a moral community's social contract, as exemplified by predatory communities.

A refined moral order allows the intercommunal interaction to progress beyond moral anarchy. A moral order might be developed by an intensive process of introspection—for example, by consideration of

[10] The Mengerian tradition takes a similar approach in describing the emergence of institutions. In this light, the development of moral community and moral order is informed by Menger's causal-genetic description of the evolution of money (Menger [1871] 2007, [1883] 1985).

[11] Similarly, Vincent Ostrom (2006) refers to artifacts of governance that are generated in the process of participation in institutions of governance.

fundamental moral or legal principles—by importation and emulation, as in the case of the Khonds. By the moral order, conflict between neighboring communities is mediated by a shared rule structure, for example, the tolerance exemplified by the liberal moral order (Mises [1949] 1996, 146, 148, 152). The moral order can serve as the basis for the resolution of intracommunal conflict where the first-level contract fails at this task, or may resolve conflict between individuals from communities with differing first-level contracts *if* they at least participate in the same moral order. As with the modern liberal order, this enables individuals to live, if they so wish, outside or on the margins of any particular moral community, freeriding in some sense on the moral infrastructures of existing communities.

In all, I have presented insights from Moehler's multilevel theory of the social contract by using his work and the framework presented by Buchanan to mutually inform one another. Concerned about incentive compatibility, Moehler constrains his analysis to his agent *homo prudens*, and therefore includes, for specific empirical conditions, the binding constraint of a basic income guarantee that enables behavior typified by *homo prudens*. Although consistent with the structure of Moehler's framework, my development of Buchanan's social contract theory is concerned with a different dimension of this problem. The multilevel social contract theory developed here is a strictly positive theory of coordination of diverse actors subject to a multilevel social contract. This theory does not preclude failure through anything analogous to the weak principle of universalization. It seeks to describe how cooperation within and between communities can exist in spite of the real and ever-present threat of a return to moral anarchy.

REFERENCES

Aligica, Paul D. 2014. *Institutional Diversity and Political Economy: The Ostroms and Beyond*. Oxford: Oxford University Press.

Aligica, Paul D., and Vlad Tarko. 2012. "Polycentricity: From Polanyi to Ostrom, and Beyond." *Governance* 25 (2): 237–262.

Boettke, Peter J., Christopher J. Coyne, and Peter T. Leeson. 2008. "Institutional Stickiness and the New Development Economics." *The American Journal of Economics and Sociology* 67 (2): 331–358.

Bourdieu, Pierre. 1990. *The Logic of Practice*. Stanford, CA: Stanford University Press.

Buchanan, James M. 1965. "Ethical Rules, Expected Values, and Large Numbers." *Ethics* 76 (1): 1–13.

Buchanan, James M. (1979) 1999. "Natural and Artifactual Man." In *The Collected Works of James Buchanan. Volume 1. The Logical Foundations of Constitutional Liberty,* edited by Geoffrey Brennan, Hartmut Kliemt, and Robert D. Tollison, 246–259. Indianapolis, IN: Liberty Fund.

Buchanan, James M. (1975) 2000. *The Collected Works of James M. Buchanan. Volume 7. The Limits of Liberty: Between Anarchy and Leviathan*. Edited by Hartmut Kliemt. Indianapolis, IN: Liberty Fund.

Buchanan, James M. (1981) 2001a. "Moral Community, Moral Order, and Moral Anarchy." In *The Collected Works of James M. Buchanan. Volume 17. Moral Science and Moral Order*, edited by Hartmut Kliemt, 187–201. Indianapolis, IN: Liberty Fund.

Buchanan, James M. (1983) 2001b. "Moral Community and Moral Order: The Intensive and Extensive Limits of Interaction." In *The Collected Works of James M. Buchanan. Volume 17. Moral Science and Moral Order*, edited by Hartmut Kliemt, 202–210. Indianapolis, IN: Liberty Fund.

C. R. 1848. "The Khonds — Abolition of Human Sacrifice and Female Infanticide." *Calcutta Review* 10: 273–341.

Congleton, Roger D. 2018. "Buchanan on Ethics and Self-interest in Politics: A Contradiction or Reconciliation?" In *Buchanan's Tensions: Reexamining the Political Economy and Philosophy of James M. Buchanan*, edited by Peter J. Boettke, and Solomon Stein, 35–50. Arlington, VA: Mercatus Center.

Crawford, Sue E. S., and Elinor Ostrom. 1995. "A Grammar of Institutions." *The American Political Science Review* 89 (3): 582–600.

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 August 20, 2019. https://plato.stanford.edu/entries/contractarianism-contemporary/.

Dekker, Erwin. 2016. "Exemplary Goods: Exemplars as Judgment Devices." *Valuation Studies* 4 (2): 103–124.

Dekker, Erwin, and Pavel Kuchař. 2016. "Exemplary Goods: The Product as Economic Variable." *Journal of Contextual Economics: Schmollers Jahrbuch* 136 (3): 237–255.

Dekker, Erwin, and Blaž Remic. 2019. "Two Types of Ecological Rationality: Or How to Best Combine Psychology and Economics." *Journal of Economic Methodology* 26 (4): 291–306.

Gangte, Lalrameng K. 2017. "Human Sacrifice among the Khonds of Orissa C. 1836–1861: A Study." *Mizoram University Journal of Humanities & Social Sciences* 3 (1): 114–125.

Gaus, Gerald. 2013. "Hobbesian Contractarianism—Orthodox and Revisionist." In *The Bloomsbury Companion to Hobbes*, edited by Sharon A. Lloyd, 263–278. London: Bloomsbury Academic.

Gaus, Gerald. 2018. "It Can't be Rational Choice All the Way Down: Comprehensive Hobbesianism and the Origins of Moral Order." In *Buchanan's Tensions: Reexamining the Political Economy and Philosophy of James M. Buchanan*, edited by Peter J. Boettke, and Solomon Stein, 117–145. Arlington, VA: Mercatus Center.

Gigerenzer, Gerd, and Henry Brighton. 2009. "Homo Heuristicus: Why Biased Minds Make Better Inferences." *Topics in Cognitive Science* 1 (1): 107–143.

Hamzeh, A. Nizar. 1994. "Qatar: The Duality of the Legal System." *Middle Eastern Studies* 30 (1): 79–90.

Hobbes, Thomas. (1651) 1968. *Leviathan*. Edited with an introduction by Crawford B. Macpherson. Harmondsworth: Penguin.

Knight, Frank H. 1921. *Risk, Uncertainty, and Profit*. New York, NY: Houghton Mifflin Company.

Kukathas, Chandran. 2003. *The Liberal Archipelago: A Theory of Diversity and Freedom.* Oxford: Oxford University Press.

Leeson, Peter T. 2014. "Human Sacrifice." *Review of Behavioral Economics* 1 (1–2): 137–165.

Macpherson, Samuel C. 1865. *Memorials of Service in India*. Edited by William Macpherson. London: John Murray.

Menger, Carl. (1883) 1985. *Investigations into the Method of the Social Sciences with Special Reference to Economics*. Translated by Francis J. Nock. Edited by Louis Schneider. New York, NY: New York University Press.

Menger, Carl. (1871) 2007. *Principles of Economics*. Translated by James Dingwall, and Bert F. Hoselitz. Auburn, AL: Ludwig von Mises Institute.

Mises, Ludwig von. (1949) 1996. *Human Action: A Treatise on Economics. In 4 Volumes.* Edited by Bettina Bien Greaves. Indianapolis, IN: Liberty Fund.

Moehler, Michael. 2014. "The Scope of Instrumental Morality." *Philosophical Studies* 167 (2): 431–451.

Moehler, Michael. 2018. *Minimal Morality: A Multilevel Social Contract Theory*. Oxford: Oxford University Press.

Moehler, Michael. 2020. *Contractarianism*. Cambridge: Cambridge University Press.

Munger, Michael. 2020. "Moral Community and Moral Order: Buchanan's Theory of Obligation." *Public Choice* 183 (3–4): 509–521.

Olson, Mancur. 1993. "Dictatorship, Democracy, and Development." *The American Political Science Review* 87 (3): 567–576.

Ostrom, Vincent. 2006. "The 2005 John Gaus Lecture. Citizen-Sovereigns: The Source of Contestability, the Rule of Law, and the Conduct of Public Entrepreneurship." *PS: Political Science and Politics* 39 (1): 13–17.

Ostrom, Vincent. (1991) 2014. "Polycentricity: The Structural Basis of Self-Governing Systems." In *Choice, Rules and Collective Action: The Ostroms on the Study of Institutions and Governance*, edited by Filippo Sabetti, and Paul Dragos Aligica, 45–60. Colchester: ECPR Press.

Polanyi, Michael. (1951) 1998. *The Logic of Liberty: Reflections and Rejoinders*. Indianapolis, IN: Liberty Fund.

Searle, John R. 1995. *The Construction of Social Reality*. New York, NY: The Free Press.

Searle, John R. 2005. "What is an Institution?" *Journal of Institutional Economics* 1 (1): 1–22.

Searle, John R. 2006. "Social Ontology: Some Basic Principles." *Anthropological Theory* 6 (1): 12–29.

Smith, Vernon L. 2003. "Constructivist and Ecological Rationality in Economics." *The American Economic Review* 93 (3): 465–508.

Stigler, George J., and Garry S. Becker. 1977. "De Gustibus Non Est Disputandum." *The American Economic Review* 67 (2): 76–90.

Van Schoelandt, Chad. 2019. "Between Traditional and Minimal Moralities." *Analysis* 79 (1): 128–140.

Vanberg, Victor, and James M. Buchanan. 1988. "Rational Choice and Moral Order." *Analyse & Kritik* 10 (2): 138–160.

Wagner, Richard E. 2017. *James M. Buchanan and Liberal Political Economy: A Rational Reconstruction*. Lanham, MD: Lexington Books.

James Caton is an economist and professor in the Department of Agribusiness and Applied Economics at North Dakota State University, where he is also a Faculty Fellow at the Center for the Study of Public Choice and Private Enterprise. His research focuses on entrepreneurship and monetary economics. He has published in scholarly journals including the *Southern Economic Journal*, the *Journal of Entrepreneurship and Public Policy* and *The Journal of Artificial Societies and Social Simulation*. He is a Fellow with the Sound Money Project at the American Institute for Economic Research.
Contact e-mail: <james.caton@ndsu.edu>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 30–42.
https://doi.org/10.23941/ejpe.v13i2.513

Kantian Kantian Optimization

Matthew Braham
Universität Hamburg

Martin van Hees
Vrije Universiteit Amsterdam

I. Introduction

In their *Theory of Games and Economic Behavior*, von Neumann and Morgenstern refer to a game-theoretic solution as a “standard of behavior” ([1944] 1953, 41). If we apply this description to all game-theoretic solution concepts and interpret the notion of a ‘standard of behavior’ as a *norm*, we can say that game theory is the study of norm-constrained behavior. Given the rich variety of ethical theories of norms, it is surprising that so much of game theory is dominated by one particular norm: Nash equilibrium. Other solution concepts may describe other norms which are worth studying.

In *How We Cooperate: A Theory of Kantian Optimization*, John Roemer (2019) sets out to develop and defend an alternative account of norm-constrained behavior. His idea is to apply Kantian moral reasoning to provide us with a new theory of social cooperation. Specifically, he tries to make use of the Categorical Imperative (CI) in an optimization model that guarantees mutually beneficial states of affairs in archetypical social dilemmas such as recycling, volunteering in times of war (‘doing one’s part’), soldiers protecting comrades in battle, voting, paying taxes, tipping, and charitable giving. Basically, Roemer wants to model and explain under what conditions we can solve the two major problems that afflict Nash equilibrium (16). These are the tragedy of the commons and the free-rider problem, which concern inefficiencies in the presence of negative and positive externalities, respectively. Roemer’s central thesis is:

Authors’ Note: We are very grateful to the editors for their helpful comments and suggestions.

> Kantian optimization ‘solves’ what must appear as the two greatest failures of Nash optimization, from the viewpoint of human welfare. (16)

Methodologically, Roemer achieves his result not by tampering with the canonical concepts of preferences and utility of rational choice theory that underpin economics, but rather by distinguishing between two different kinds of optimization strategy that are possible, given the structure of rational choice and our preferences. The two types correspond with two different questions we may raise (12):

> **Nash optimizer**. “Given the strategy chosen by my opponent, what is the best strategy for me?”
>
> **Kantian optimizer**. “What is the strategy I would like both of us to play?”

Roemer argues that the answers to these questions will usually differ. For him, social cooperation is all about the latter form of optimization. He grounds this distinction in a synthesis of work in evolutionary psychology and linguistics (Tomasello 2014, 2016) and social ontology (Gilbert 1990; Bratman 1992). The idea that emerges from this synthesis is that as a fact of evolution, humans are a “cooperative species” (1), meaning that we have evolved natural capacities ‘to do things together’ and to understand the value of doing so. We are able to form complex systems of language, behavior, and social interaction through which we can share a “union of interests” (4) and according to which we are able to judge that doing things in this way is both individually and mutually advantageous. Thus, when faced with social dilemmas, our instincts and our thoughts are not necessarily Nash optimized at all. Rather, we may understand that Kantian thinking and optimization can be more advantageous.

The heart of Roemer’s claim is that what makes Kantian optimization specifically ‘Kantian’ is that it sufficiently resembles the fundamental feature of the CI, namely, *universalization*. In Roemer’s phrasing:

> *Take those actions you would will be universalized.* (13, emphasis added)

Roemer does not require that such a universalization be governed by altruism; in fact, far from it. The trick, rather, is to pull social cooperation out of an individualistic hat. The Kantian optimizer is still an individualistic being to the extent that she optimizes the choice of a common strategy—a strategy which is played by all—but which would be best for her. The Kantian optimizer does not consider the payoffs of others and each agent need only know their own preferences. What is required, however, is that each individual expects others to behave in a like manner and that this is based on trust or past experiences (13).

For our contribution to this symposium, we shall ignore Roemer's contribution to the study of human cooperation. The alternative solution concept that he develops is a serious challenge to orthodoxy in economics and game theory. The rigorous formal analysis as well as its application to market economies makes it a profound contribution to both normative economics and formal ethics. However, rather than expanding upon the relevance of the analysis, or its relation to other game-theoretic unorthodoxies—a discussion of the relation with models of team reasoning (Bacharach 2006) is regrettably missing—we will hone in on its theoretical embedding and, in particular, on its 'Kantian' credentials.

Roemer's use of 'Kantianism' follows an established tradition among economists. According to this tradition, an agent follows Kantian morality if she acts under the assumption that others will do the same thing that she does, and if she tries to maximize her utility under that constraint. Economists frequently overlook the fact that this interpretation of Kant differs from the core of Kantian ethical theory: the CI is about the universalization of an agent's maxims rather than her actions and it does not refer to utility maximization. Although Roemer admits that we should not afford too much importance to the reference to Kant (13), it would be interesting to examine whether his new solution concept can be grounded in such a way that it becomes compatible with the Kantian perspective in ethics.

We shall suggest a way of strengthening the Kantian pedigree of Roemer's approach. First, in section II, we will focus on the distinction between actions and maxims and explain the importance of that distinction for the Kantian perspective. It is true that Kantian optimization sometimes yields the same conclusion as a run-of-the-mill application of Kant's CI, but it can also yield both false positives (a defence of immoral behavior) and false negatives (the rejection of permissible behavior). Partly drawing on our earlier work, we then (in section III) give an interpretation of maxims that brings Roemer's analysis closer to Kantian ethics. Finally, we wrap up our analysis with a short conclusion (section IV).

II. IS KANTIAN OPTIMIZATION KANTIAN?

Roemer's main reason for attributing Kantianism to his solution concept is its reference to universalization.[1] As said, Roemer is quite aware that his theory is only loosely 'Kantian', but he nevertheless takes the Kantian optimization condition to be a "natural interpretation" of Kant's CI (viii). Also, he chooses a "'Kantian' nomenclature" because "there is a history of using it in economics" (13).[2]

Of course, in many cases the application of a standard Kantian argument will yield an outcome that coincides with Roemer's analysis. To abide with the usual moral injunctions against theft, deceit, murder, etc., is to adopt a course of action that we all strictly prefer to one in which everyone is willing to transgress those norms. Moreover, to follow such norms means to not be tempted to change our behavior if it happens to improve our personal situation: neither the 'white lie' nor the 'perfect fraud' is an option in Kantian morality.

A fundamental difference between Kant and Roemer is that, for Kant, to act morally is to act autonomously in accordance with those *maxims* that satisfy the 'Moral Law' (for which the CI is a test). Roemer is not concerned with maxims but with actions. Moreover, in Roemer's account, a course of action is moral if we derive some advantage from it being universally adopted. This is more akin to the thought of Hobbes or Hume than it is to that of Kant. For Kant, 'advantage' at best plays an indirect role in morality; that is, it may simply make it easier to follow the moral law (Kant [1797] 1996, 519, 6:388).

Roughly speaking, Kant's CI examines whether the *underlying* reason of a person's action—the *maxim*—is one to which everyone could possibly subscribe. It is a twofold test. In the first step, the CI checks for the existence of a possible world in which everyone could act on the basis of that maxim. If such a possible world does indeed exist, a second step checks to see if an agent that adopts the maxim *can will* that world into existence. If so, acting on the basis of that maxim is morally admissible and not so otherwise.

Thus formulated, the CI is notoriously ambiguous and philosophers have spilled copious amounts of ink in their efforts to interpret it. Yet

[1] To simplify the presentation, we restrict our analysis to games with a common diagonal, allowing us to take simple Kantian equilibrium to be the relevant solution concept (cf. 23, Proposition 2.1).

[2] In an endnote to page 13, Roemer actually suggests that his approach is closer to Kant's 'Hypothetical Imperative' and that his use of the term 'Kantian' is "for its suggestive meaning and [I] do not wish to imply that there is a deeper, Kantian justification of my proposal" (220n7).

despite this ambiguity, two features of the CI stand out: (a) it tests one's maxims rather than one's actions; (b) it focuses on the *possibility* of everyone acting on the same maxim rather than on the *advantages* that we may derive from it.

Same Action, Different Maxims

One implication of Kant's focus on maxims is that one and the same action can be appraised differently depending on how its underlying maxim is formulated. Roemer's Kantian optimization does not capture this distinctive feature of Kantian morality. The relevance of this can be illustrated by way of the tragedy of the commons. If we exclusively focus on actions, then the farmer who brings his herd to the overgrazed commons because he needs extra earnings, however meagre they may be, to care for the well-being of his family presumably acts in accordance with the moral law. Here Kantian optimization leads to a false negative: the behavior is incorrectly condemned as a wrong. It would be correctly so rejected, if, say, the farmer lets his herd graze because he wants to make an extra buck regardless of the circumstances.

The possibility of false negatives makes it rather clear that we cannot always condemn or blame an agent for not being 'cooperative'. False positives are also possible. The original illustration of the Prisoner's Dilemma game—the case with two prisoners who are offered a deal by the District Attorney—can serve as an example of such a false positive. Assume the prisoners are members of a criminal gang and have in fact committed the crimes they are accused of. Acting on the basis of the maxim of loyalty, they both deny their guilt (play 'Cooperate'). Thus, they both play Kantian equilibrium strategies but theirs is not a play that a Kantian would be likely to endorse.

One can object, of course, that we can ignore these false negatives and false positives because their possibility merely underscores that we have to be careful in describing the moral context of the game at hand. A game to which we apply the Kantian solution concept—the Prisoner's Dilemma in the current discussion—is assumed to describe what we have called elsewhere the *moral field*: it specifies all the relevant moral features of the situation (Braham and van Hees 2012, 611). To refer to the criminal nature of the organization or to the actions that led to the prisoners' arrest means bringing in morally relevant features that are not captured by the game at hand. Yet, if we were to expand the scope of the game, and thus the moral field, we may very well see that within the resulting 'larger' game of which the Prisoner's Dilemma forms a part, cooperation between the two

	L	F
L	$(0,0)$	$(3,3)$
F	$(-1,-1)$	$(0,0)$

Table 1: The Tango game.

prisoners may fail to form a Kantian equilibrium and therefore may not be justifiable. However, whereas such an expansion may indeed rule out some incorrect judgements, without further argument we cannot be sure that it will *always* do so.

Same Maxim, Different Actions

A different problem arises from the possibility that a maxim can be associated with different actions. This is the case if an individual can act upon a maxim in different ways, but also when the very same maxim corresponds with different actions for different individuals. The latter occurs, for instance, if the provision of a public good requires different inputs from different individuals because it necessitates a division of labour. The optimal outcome will then result only if all individuals act differently and within their domain of expertise. This poses no problem for Kantian morality (if all agents intend to bring about a public good) but it may complicate the application of Kantian optimization.

To see this, consider what we call the 'Tango game' (Table 1); a two-person game in which the players have two strategies, *Lead* (L) and *Follow* (F). The row player specializes in L while the column player specializes in F. The players' respective utility functions are the same. The worst outcome ensues if they both try to perform the role they were not specialized to do, while the two next preferred outcomes are those in which one of them deviates from her specialization, and their most preferred outcome is the one in which they both act on the basis of their specialization.

The Pareto-optimal play (L,F) is the unique Nash equilibrium, whereas the two Kantian equilibria are suboptimal. In the Tango game, Kantian optimization thus generates a false positive ('act the same way'). Being 'cooperative' in this context, however, means to 'act differently'.

One could argue that this observation is a mere semantic sleight of hand in that we are simply mis-describing the actions. Suppose we redescribe the players' actions as *Specialization* (S) and *Non-specialization* (N), respectively. This leads to the 'Modified Tango game' (Table 2).

Now, the problem vanishes and Kantian optimization yields the morally desirable outcome. Just as the problems following from multiple maxims suggested a move to a different type of modelling, so too the problem

	S	N
S	$(3,3)$	$(0,0)$
N	$(0,0)$	$(-1,-1)$

Table 2: The Modified Tango game.

arising from multiple actions instantiating the very same maxim may be solved via re-modelling. But while this may indeed work for this particular game, it is a somewhat ad hoc solution. Why would the second game, rather than the first, describe the situation correctly? Can we simply decide how to describe the agents' actions? As Roemer notes (28), Kantian optimization requires specifying when different individuals' strategies are the same, which may not always be obvious.

Preferences and Morality

Roemer emphasizes (13) that the Kantian optimizer is only trying to realize her own preferences as well as possible. While being altruistic is compatible with Kantian optimization, it is not at all necessary for it. Crucial is that the realized outcome be the most preferred one, the character of the preferences themselves is not relevant. The difference between Kantian and Nash equilibrium lies in the comparison between their respective outcomes and not with the way in which they themselves are compared, which is preference-based. Yet the preference-based comparison does not square well with an essential characteristic of Kantian ethics. For Kant, the CI is about the *possibility* of universalizing a maxim. It is not about the desirability of the consequences of universally adopted actions or maxims.

III. Re-Kanting Roemer

Roemer's project suggests the value of bringing Kantianism and the welfare consequentialism of economics closer together. But, can we bring them even closer? That is, is there a game-theoretic analysis of the CI that focuses on maxims but which draws on Roemer's interpretation of Kantian optimization? We shall argue that such an analysis is indeed possible.[3]

To do so we have to unpack the formulation of the CI—that is closest to Roemer's idea of Kantianism—which is known as the *Formula of Universal Law* (FUL) version of the CI:

[3] Here, we will use some ideas from our earlier work (Braham and van Hees 2015) but which deviate from it in the way we connect maxims with preferences. The latter idea is motivated by Amartya Sen's (1974) early suggestion to model morality in terms of meta-rankings.

> **FUL.** Act only in accordance with that maxim through which you can at the same time will that it become a universal law. (Kant [1785] 1996, 73, 4:421)[4]

FUL has two constituent parts: the aforementioned maxims and the two tests of the universalizability of the maxims. We start with the concept of a maxim. One option, which we adopt, is to view a maxim as a rule of conduct that refers to a person's intentions across a range of circumstances.[5] We then take a maxim to be about which states of affairs are to be picked out by conduct whenever certain circumstances arise. Regimenting it a little, a maxim is a tripartite relation of the form:

> an agent will do α if β in order to ϕ,

where α ranges over actions, β over circumstances, and ϕ over states of affairs.

According to such a conception, a maxim consists of two intentions: an *act*-intention (performance of some action α) and an *outcome*-intention (realization of some state of affairs ϕ). Here, we can use the revealed-preference interpretation of utility functions according to which an agent's preference describes a hypothetical choice that she faces. These hypothetical choices can in turn be understood as describing her intentions: they specify a certain choice (α) for a state of affairs (ϕ) in a possible choice situation (β). By this interpretation 'i prefers x over y' means 'if the agent were to have a choice between x and y, then she would choose x', which in turn can be interpreted as 'i intends to choose x if the choice is between x and y'. Taking a maxim to be a collection of such conditional intentions, we arrive at a conceptual link between Kantian maxims and utility functions. In Kantian ethics the admissibility of people's behavior does not depend on the assessment of the outcomes of their actions but on the motivation underlying their behavior. By interpreting preferences and the corresponding utility functions in terms of that motivation rather than in terms of an assessment, we can apply the economic apparatus.

[4] As is well-known, Kant also provided us with a number of other less formalistic formulations in the *Groundwork of the Metaphysics of Morals*, which he believed to be equivalent. These are known as the formulas of 'Humanity', of 'Autonomy', and 'Kingdom of Ends'.

[5] This conception goes back to O'Neill (1975, 34–42). See also Westphal (2011, 111). There are other possible interpretations of this approach. For a recent and comprehensive analysis of what Kantian maxims are, see Herissone-Kelly (2018).

	t_1	t_2
s_1	x	y
s_2	z	v

Table 3: An example game form h.

Indeed, whereas such a route may require a stretch of imagination for the conventional Kantian, it does fit neatly within economic theory. To indicate the outlines of such 'Kantian economics', let us start with a game form h that models the situation that contains all the ingredients of a game (players, strategies, outcomes) except for the preferences of the players. A game g is, then, defined as a game form h plus a preference profile $\mathbf{u} = (u_1, \ldots, u_n)$. Say the game form h is as in Table 3.

Thus the game form is 'part of' a Prisoner's Dilemma if the outcomes correspond with the time that each prisoner has to spend in prison and if the prisoners have the intention to reduce their prison time as much as possible. In a 'Battle of the Sexes' game, the outcomes describe ways of spending the evening and the partners intend to be together although both would opt for different things to do together, etc. We denote a utility function of an agent i that is associated with a particular maxim m as $u_{i,m}$. Accordingly, a profile in which each individual utility function is associated with the same maxim m is denoted as $\mathbf{u}_m$.

We can now turn to the formulation of the CI. We will focus only on that part of it that is closest to Roemer's analysis and which Christine Korsgaard (1996, 93) calls the 'Practical Contradiction Interpretation'.[6] In doing so we sidestep the issue of ascertaining which maxims can be universally adopted at all, and simply assume that the information is given exogenously.

We say that a practical contradiction arises if the universal adoption of a maxim would lead to a state of affairs that is at odds with the maxim. To see how this works, let h be the game form and let $\mathcal{M}$ be the non-empty set of all maxims that can be adopted universally in h. For any $m \in \mathcal{M}$, g_m denotes the game $(h, \mathbf{u}_m)$ that describes the universal adoption of m. To simplify the analysis, we assume that the game g_m associated with the universal adoption of a maxim m always has a unique and pure Nash equilibrium, the outcome of which is denoted by x^*_m.[7]

[6] The two tests that the CI is taken to comprise are commonly referred to as the Contradiction in Conception (CC) and the Contradiction in the Will (CW) tests (O'Neill 1975). For our interpretation of the CI, we draw upon our earlier work (Braham and van Hees 2015). Note, however, that what we take to be the CW test is interpreted as the CC test by Korsgaard.

[7] The assumption simplifies the presentation because it avoids the need to introduce preferences over lotteries or set-preferences.

Whereas a maxim in $\mathcal{M}$ is one that *can* be a universal law, this does not yet mean that we can rationally *will* it to become a universal law. In the *Groundwork of the Metaphysics of Morals*, Kant expands on FUL as follows:

> Some actions are so constituted that their maxim cannot even be *thought* without contradiction as a universal law of nature, far less could one *will* that it *should* become such. In the case of others that inner impossibility is indeed not to be found, but it is still impossible to *will* that their maxim be raised to the universality of a law of nature because such a will would contradict itself. (Kant [1785] 1996, 75, 4:424)

Given our interpretation of individual preferences, we say that a rational agent *can will* the universal adoption of a maxim m if, and only if, the outcome x^*_m resulting from the universal adoption of the maxim m is, according to m, indeed the outcome that he intends to choose in any pairwise comparison with the outcome resulting from the universal adoption of any other maxim. Or, more succinctly, it rules out the possibility of a rational agent acting on the basis of a maxim m that tells her not to act on it if everyone were to do so. This possibility is the conflict within a person's will that we take to be excluded by Kant's FUL.

We can illustrate this framework with a two-person version of the tragedy of the commons. Assume that the farmers have only two maxims available to them, which for simplicity's sake, we call *individual* and *collective*. The *individual* maxim is 'unconditional self-interest'; the *collective* maxim conditions behavior on the social optimum. Assuming with Kant ([1785] 1996, 74–75, 4:423) that universal self-interested behavior is feasible, their adoption of the individual maxim yields the Prisoner's Dilemma, whereas a game in which the cooperative outcome is the only Nash equilibrium would result if they both chose the collective maxim. Next we construct games in which the players adopt maxims. To do so, we use the notion of a Kantian game form.

Kantian game form. Given a game form h with associated $\mathcal{M}$, the Kantian game form $\hat{h}$ is a particular game form in which:

1. Each individual strategy set is $\mathcal{M}$.
2. The outcome of a play $(m, \ldots, m) \in \mathcal{M} \times \ldots \times \mathcal{M}$ is x^*_m.

Each combination of a Kantian game form and a preference profile $\mathbf{u}_m$ associated with a maxim m yields a unique game $(\hat{h}, \mathbf{u}_m)$. We can now bring Roemer's idea of Kantian optimization into harmony with Kantian ethical theory. We call this '*Kantian* Kantian Optimization', and formulate it as follows:

> ***Kantian* Kantian optimization**. Given a Kantian game form $\hat{h}$, acting on the basis of maxim m is morally admissible if, and only if, (a) $m \in \mathcal{M}$, and (b) the play $(m, \ldots, m)$ is a simple Kantian equilibrium of the game $(\hat{h}, \mathbf{u}_m)$.

Using again the tragedy of the commons illustration, let $\mathcal{M} = \{m^i, m^c\}$ and let $(h, \mathbf{u}_{m^i})$ be a Prisoner's Dilemma game with the unique Nash equilibrium (D, D) of both players defecting. Conversely, a universal adoption of the collective maxim yields a 'Prisoner's Harmony' game $(h, \mathbf{u}_{m^c})$ with a unique Nash equilibrium (C, C) of both players cooperating. Turning to the Kantian game form $\hat{h}$ in which the strategies of the agents are m^i and m^c, we next examine the agents' assessments of the various games resulting from the universal adoption of a maxim. We see that in terms of their Prisoner's-Dilemma preferences (describing maxim m^i), the Prisoner's-Harmony outcome $x^*_{m^c}$ is ranked higher than the Prisoner's-Dilemma outcome $x^*_{m^i}$. Hence, (m^i, m^i) is not a simple Kantian equilibrium of the game $(\hat{h}, \mathbf{u}_{m^i})$. It is for that reason that acting on the basis of the individual maxim is said to be inadmissible. On the other hand, the play (m^c, m^c) is a simple Kantian equilibrium in the game in which m^c rather than m^i describes the preferences of the agents, that is, in $(\hat{h}, \mathbf{u}_{m^c})$. Acting on the basis of m^c is, therefore, admissible.

IV. CONCLUSION

We have suggested that Roemer's account of Kantian optimization can be brought closer to Kantian ethical theory by making certain suitable assumptions about the interpretation of maxims. Kantian optimization can then be seen as forming a solution of particular games, namely, games in which agents choose maxims on the basis of which they will act. Since such games have a very specific nature which do not coincide with the games in which Roemer analyses Kantian optimization, we called the resulting account '*Kantian* Kantian optimization'.

How exactly does this strengthen the Kantian pedigree of Roemer's solution concept? By definition, *Kantian* Kantian optimization is an instance of Kantian optimization, but the converse need not always be true. Kantian optimization may fail to be truly Kantian because there may be no maxim that, if universally adopted, would lead to the cooperative outcome. A natural next step in our analysis would be to examine the class of games and maxims in which Kantian optimization reaches the same verdict as *Kantian* Kantian optimization. This is important for welfare economists as it will provide guidance as to how to achieve socially optimal outcomes in the morally right kind of way—which is what Kantian morality is all about. Thereby, it will introduce a dimension of moral proceduralism that is generally lacking. In this way, Roemer's *How We Cooperate* has opened up a wholly new avenue of theoretical possibilities.

REFERENCES

Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. Edited by Natalie Gold and Robert Sugden. Princeton, NJ: Princeton University Press.

Braham, Matthew, and Martin van Hees. 2012. "An Anatomy of Moral Responsibility." *Mind* 121 (483): 601–634.

Braham, Matthew, and Martin van Hees. 2015. "The Formula of Universal Law: A Reconstruction." *Erkenntnis* 80 (2): 243–260.

Bratman, Michael E. 1992. "Shared Cooperative Activity." *The Philosophical Review* 101 (2): 327–341.

Gilbert, Margaret. 1990. "Walking Together: A Paradigmatic Social Phenomenon." *Midwest Studies in Philosophy* 15 (1): 1–14.

Herissone-Kelly, Peter. 2018. *Kant on Maxims and Moral Motivation: A New Interpretation*. Cham: Springer.

Kant, Immanuel. (1785) 1996. "Groundwork of the Metaphysics of Morals." In *Practical Philosophy*, edited by Mary J. Gregor, 37–108. Cambridge: Cambridge University Press.

Kant, Immanuel. (1797) 1996. "The Metaphysics of Morals." In *Practical Philosophy*, edited by Mary J. Gregor, 353–604. Cambridge: Cambridge University Press.

Korsgaard, Christine M. 1996. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

O'Neill, Onora. 1975. *Acting on Principle: An Essay on Kantian Ethics*. New York, NY: Columbia University Press.

Roemer, John E. 2019. *How We Cooperate: A Theory of Kantian Optimization*. New Haven, CT: Yale University Press.

Sen, Amartya K. 1974. "Choice, Orderings and Morality." In *Practical Reason: Papers and Discussions*, edited by Stephen Körner, 54–67. Oxford: Blackwell.

Tomasello, Michael. 2014. *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.

Tomasello, Michael. 2016. *A Natural History of Human Morality*. Cambridge, MA: Harvard University Press.

von Neumann, John, and Oskar Morgenstern. (1944) 1953. *Theory of Games and Economic Behavior*. 3rd edition. Princeton, NJ: Princeton University Press.

Westphal, Kenneth R. 2011. "Practical Reason: Categorical Imperative, Maxims, Laws." In *Immanuel Kant: Key Concepts*, edited by Will Dudley and Kristina Engelhard, 103–119. Durham: Acumen.

Matthew Braham is Professor of Practical Philosophy at Universität Hamburg, Germany. He is Deputy Speaker of the Graduate Program in Collective Decision-Making. He has published in moral and political philosophy, social choice, and game theory.
Contact e-mail: <matthew.braham@uni-hamburg.de>

Martin van Hees is Professor of Ethics at VU Amsterdam. His research interests concern the foundations of (moral and political) liberalism and he has worked on theories of freedom, the analysis of moral responsibility, and quality of life assessments. He is a former editor of *Economics & Philosophy*.
Contact e-mail: <m.van.hees@vu.nl>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 43–84.
https://doi.org/10.23941/ejpe.v13i2.514

Normative Aspects of Kantian Equilibrium

Itai Sher
University of Massachusetts Amherst

I. Introduction

This paper concerns John Roemer's new book *How We Cooperate: A Theory of Kantian Optimization* (2019). The book provides a solution concept for games, which is an alternative to the standard economist's concept of Nash equilibrium. Roemer names the new solution concept *Kantian equilibrium.* Roemer explains the reason for the name—"I invoke Immanuel Kant here because of his categorical and hypothetical imperatives, which state that one should take those actions one would like to see universalized" (13)—but Roemer disclaims any very detailed relation to Kant's moral philosophy, writing: "I use the term for its suggestive meaning and do not wish to imply that there is a deeper, Kantian justification of my proposal" (220n7).

The basic idea behind Kantian equilibrium is that in a cooperative situation everyone asks: 'What would be best *for me* if everyone were to do it?' When everyone answers in the same way, then that is what everyone does. There are variants of this idea that can be applied to cases when everyone does not answer in the same way.

Kantian equilibrium contrasts with Nash equilibrium. In Nash equilibrium, one chooses *one's own* strategy to maximize *one's own* utility *holding others' strategies fixed* at the equilibrium. In contrast, in Kantian equilibrium, one chooses the *common strategy to be adopted by everyone* to maximize *one's own* utility.

A basic theme of *How We Cooperate* is that the economic literature conflates altruism and cooperation. To explain cooperation, rather than dropping economic theory's reliance on self-interest and allowing altruism, we should drop economic theory's traditional model of optimization. We should keep the assumption of self-interest and replace Nash opti-

Author's Note: I am grateful to Leah Boustan, Scott Cunningham, David Clingingsmith, Scott Imberman, Juan Moreno-Cruz, Romans Pancs, Paolo Piacquadio, and Peter Sher for helpful comments and discussions relating to this paper. I am also grateful to the editors Hein Duijf, Akshath Jitendranath, Marina Uzunova for helpful comments.

mization with Kantian optimization: agents should not be assumed to hold others' actions fixed when optimizing their own, but should rather think of others' choices as part of the optimization.

We can think of Kantian equilibrium either as a *descriptive* or a *prescriptive* concept; that is, it may describe how people *do* behave or how they *should* behave. Or it could be both descriptive and prescriptive. Roemer writes:

> I intend the concept of simple Kantian equilibrium to be both a positive and a normative concept: positive because I believe it is a good model of many real instances of cooperation, and normative because I believe that the observation 'we must all hang together, or ... we shall all hang separately' makes good sense as a recommendation for action in such situations. (215)

This paper will focus on the normative, rather than the positive, aspect of Kantian equilibrium. The basic position for which I will argue is that Kantian equilibrium is an important idea but it faces both technical and non-technical challenges, which need to be overcome if it is to be successful.

Section II focuses on the technical issues and sections III–V focus on the non-technical issues. The two parts are related as the points made in sections III–V build on the formal points made in section II. Proofs of the propositions in the technical section are in the Appendix.

The three technical issues concern existence, efficiency, and strategic equivalence. First, Kantian equilibrium may not exist. This leads to the question: what is an integrated normative approach to interactions modeled as games that leads to prescriptions both when Kantian equilibrium exists and when it fails to exist? Second, while Roemer documents important cases in which Kantian equilibria are efficient and Nash equilibria are not, it is also easy to construct examples of inefficient Kantian equilibria. This matters insofar as, in the book, efficiency plays an important role in justifying Kantian equilibrium. Third, by relabeling strategies, it is possible to construct strategically equivalent games whose Kantian equilibria differ, whereas it is not possible to do this for Nash equilibrium. In many settings, especially when there is a common way of measuring strategic choices, this is not necessarily a problem but it does imply that the informational requirements for Kantian equilibrium are stronger than the informational requirements for Nash equilibrium: Kantian equilibrium does not just depend on preference data, but rather we need some privileged way of measuring strategic choices, and moreover this particular choice of measurement must have a normative justification. In cooperative social interactions in which different people who are cooperating make different types of choices, a common way of measuring the strategic decisions of different players may not be available. For example, this might occur when the leaders of a political party and their supporters cooperate, each group taking a different type of action. The general problem is conceptual: the advice 'do the same thing you would like everyone to do' does not cover all instances of cooperation, because in many such instances, different cooperators are differently situated, so that everyone *doing the same thing* is not an option. We do successfully cooperate in situations in which different people are differently situated, and ultimately we need a theory of cooperation that accommodates such situations. The variants of Kantian equilibrium introduced to address this issue do not address it in a general way.

The non-technical challenges to Kantian equilibrium center on the basic normative justification for playing Kantian equilibrium. Roemer emphasizes that Kantian equilibrium can be founded in self-interest and trust, writing:

> Playing the strategy that one would like everyone to play is, for me, motivated by the common knowledge assumption [...] and trust, not by a concern for the welfare of the group as a whole. It entails a recognition that cooperation can make *me* better off (incidentally, it makes all of us better off). But that parenthetical fact is not or *need not be* the motivation for my playing 'cooperatively.' (34–35)

Roemer argues that Kantian equilibrium is founded in self-interest and trust. I argue that whereas trust is important for Nash equilibrium—because if the other happens to deviate from their equilibrium strategy, your equilibrium strategy may no longer be a best response—the solution concept of Kantian equilibrium does not provide any formalization of the reason that trust matters. More importantly, I argue that Kantian equilibrium cannot have a foundation on the basis of trust and self-interest alone. It must be founded on some moral idea that goes beyond self-interest. While, as I mentioned above, Roemer disclaims a precise connection to Kantian deontology, it is useful to make a comparison. In the same way that the categorical imperative cannot be justified on the basis of pure self-interest, neither can Kantian equilibrium. Some appeal must be made to other moral notions such as fairness, solidarity,[1] or concern for others. While I do not take a stance on the precise nature of the justification for playing one's Kantian equilibrium strategy, in section V, I discuss the possibility of founding Kantian equilibrium in morality.

It is important to observe that, at times, Roemer seems to write as though Kantian equilibrium is justified on the basis of moral considerations. For example, Roemer connects Kantian equilibrium to what Elster (2017) referred to as quasi-moral norms,[2] writes of a slogan associated with Kantian equilibrium that "I do not object to calling this a moral code" (132), and refers to Kantian equilibria as potentially providing "ethically convincing prescriptions, if the characterization of [Kantian] equilibrium [...] appeals as a property of fairness to the individuals in the society" (216).

Despite these apparent appeals to morality, Roemer talks about founding Kantian equilibrium on self-interest, and it is difficult to see how self-interest can provide a foundation for the morality of cooperation. The resolution for this apparent tension seems to be the view that we can derive versions of the apparently moral notions by combining self-interest with a new kind of optimization. As I shall argue below, I do not think this is correct: the sort of cooperation embodied in Kantian equilibrium cannot be justified by combining self-interest with a different model of optimization. Rather, I think that agents must appeal to independent moral considerations in order to justify playing their part in a Kantian equilibrium. The role of morality in Roemer's theory is a critical issue and I discuss it further in section V.III, which closes the paper.

I want to emphasize that my aim in this paper is not to refute Kantian equilibrium, nor to argue that Nash equilibrium is superior to Kantian equilibrium. Nash equilibrium is a very well-established solution concept that has been extensively studied, and both its strengths and weaknesses are well-known. In contrast, Kantian equilibrium is a new solution concept and my purpose here is to pose some challenges for Kantian equilibrium

[1] Roemer does discuss the importance of solidarity to Kantian equilibrium, but views solidarity as compatible with pure self-interest. He also discusses connections to fairness. I will have more to say about this below.

[2] Roemer explains a quasi-moral norm as a norm:

> [...] that is motivated by wanting to do the right thing. But the 'right thing' is defined in large part by what others do. [...] I cooperate because I *see others* taking the cooperative action. A *moral norm* is, in contrast, unconditional. [...] Because I believe that trust is a necessary condition, I view cooperation as a quasi-moral norm, for trust is established by observing that others are taking the cooperative action or have taken similarly cooperative actions in the past. (9)

and to discuss its interpretation in particular in connection to its normative aspects. I view my most important point as being that a player attempting to justify Kantian equilibrium play must appeal to moral—and not just self-interested—considerations. Thus, I suggest a different interpretation of Kantian equilibrium than the one in *How We Cooperate*. While I sometimes compare Kantian and Nash equilibrium and judge Nash equilibrium more favorably on some dimensions, my aim is not to come to a verdict on which, if any, of the two solution concepts theorists should employ; indeed, as I think Roemer would agree, the answer may depend on the setting—or, in a single setting, it may be informative to compare them. In *How We Cooperate*, Roemer has done a remarkably impressive job of developing Kantian equilibrium and applying it to a rich array of economic and social settings. I think that Kantian equilibrium is an important contribution, and I hope and expect that it will receive much attention.

II. Framework and Formal Properties

This section introduces Kantian equilibrium and discusses some of its virtues and shortcomings. In particular, I present the definition of simple Kantian equilibrium and contrast it with Nash equilibrium (section II.I), I discuss existence of Kantian equilibrium and its failure (section II.II), variants of simple Kantian equilibrium, such as multiplicative, additive, and φ-Kantian equilibrium (section II.III), the efficiency of Kantian equilibrium and lack thereof (section II.IV), and the interpersonal comparisons of *strategies* on which the notion of Kantian equilibrium relies (section II.V).

II.I. Simple Kantian Equilibrium vs Nash Equilibrium

Consider a game with n players, a common strategy space S, from which each player chooses a strategy, and a set of utility functions $V^i \colon S^n \to \mathbb{R}$ for each player $i = 1, \ldots, n$. Let $[n] = \{1, \ldots, n\}$ be the set of agents.

Definition 1. *A* ***strategy*** $s^* \in S$ *is a* ***simple Kantian equilibrium*** *if:*

$$\forall i \in [n], \forall s \in S : \quad V^i(s^*, \ldots, s^*) \geq V^i(s, \ldots, s) \qquad (1)$$

That is, s^* is a simple Kantian equilibrium if every player choosing s^* is better for each player than every player choosing any other strategy s. Roemer's definition of Kantian equilibrium, applied to games in which all players have the same set of strategies, defines a *strategy* to be a simple Kantian equilibrium, whereas usually, when talking about solution concepts, we think of equilibria in terms of *strategy profiles*. We can however extend the definition to strategy profiles. Define a *strategy profile* $\mathbf{s}^* = (s_1^*, s_2^*, \ldots, s_n^*) \in S^n$ to be a simple Kantian equilibrium if there exists $s^* \in S$ such that

$$s^* = s_1^* = s_2^* = \cdots = s_n^* \tag{2}$$

and $(s^*, \ldots, s^*)$ satisfies (1).

Let us contrast Kantian with Nash equilibrium. For any strategy profile

$$\mathbf{s} = (s_1, \ldots, s_{i-1}, s_i, s_{i+1}, \ldots, s_n)$$

and, for any strategy $s_i' \in S$, the strategy profile

$$\left(s_i', \mathbf{s}_{-i}\right) = \left(s_1, \ldots, s_{i-1}, s_i', s_{i+1}, \ldots, s_n\right)$$

is the result of replacing s_i by s_i' in $\mathbf{s}$.

Definition 2. *A strategy profile* $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ *is a* ***Nash equilibrium*** *if:*

$$\forall i \in [n], \forall s_i \in S: \quad V^i(\mathbf{s}^*) \geq V^i\left(s_i, \mathbf{s}_{-i}^*\right)$$

The difference between Nash and Kantian equilibrium is that in a Nash equilibrium, each agent chooses the strategy that maximizes their own utility, holding everyone else's strategy fixed (at the Nash equilibrium profile), whereas, in a Kantian equilibrium, each player selects the strategy that would maximize her own utility if *everyone* were to use it. The strategy only counts as a Kantian equilibrium if, using this method, all agents conclude that the same strategy is best.

II.II. Existence of Simple Kantian Equilibrium

One could generalize the concept of Kantian equilibrium to relax the requirement that all players prefer the same common strategy. Consider the following solution concept—not in Roemer's book.

Definition 3. *A strategy profile* $(s_1^*, \ldots, s_n^*)$ *is a* ***subjective Kantian equilibrium*** *if:*

$$\forall i \in [n], \forall s \in S: \quad V^i\left(s_i^*, \ldots, s_i^*\right) \geq V^i(s, \ldots, s)$$

A subjective Kantian equilibrium is a strategy profile such that each player chooses the strategy that she would like everyone to choose if everyone were to choose the same strategy. However, subjective Kantian equilibrium does not require that everyone who reasons in this way actually ends up choosing the same strategy.

If we add to subjective Kantian equilibrium the requirement that, reasoning in this way, all players settle on the same desired strategy—in other words, if we add to subjective Kantian equilibrium the assumption that the same commonly adopted strategy is preferred by everyone, requirement (2)—then subjective Kantian equilibrium becomes simple Kantian equilibrium.

The above makes it clear why in general a simple Kantian equilibrium will often not exist. While subjective Kantian equilibrium exists quite broadly—as long as the optimization problem

$$\max_{s\in S} V^i (s,\dots,s) \tag{3}$$

has a solution for all players i—that solution will typically not satisfy (2). Indeed, it would be a coincidence if each person i solving problem (3) were to come up with the same solution $s_i^* = s^*$. Hence, there will typically be no simple Kantian equilibrium.

Let us contrast this with Nash equilibrium. Suppose that S is a compact convex subset of $\mathbb{R}^m$ such that, for each i, $V^i(s_i, \mathbf{s}_{-i})$ is continuous in $(s_i, \mathbf{s}_{-i})$, and quasi-concave in s_i. Then, a pure-strategy Nash equilibrium exists (Debreu 1952; Fan 1952; Glicksberg 1952). Under the same conditions, a subjective Kantian equilibrium exists.[3] But simple Kantian equilibria will rarely exist. For simplicity, continue to assume that S is a compact convex subset of $\mathbb{R}^m$, and assume moreover that each of the V^i functions is continuous and strictly concave. Then, the simple Kantian equilibrium will be unique if it exists. The existence of simple Kantian equilibrium will then require:

$$\forall i,j \in [n]: \quad \arg\max_{s\in S} V^i(s,\dots,s) = \arg\max_{s\in S} V^j(s,\dots,s) \tag{4}$$

If condition (4) initially holds, then there will be an arbitrarily small perturbation of the V^i functions that preserves strict concavity and continuity but upsets condition (4), and so undoes the existence of simple Kantian equilibrium. So simple Kantian equilibrium, even when it exists, is not robust.

[3] For a subjective Kantian equilibrium, the assumption that $V^i(s_i, \mathbf{s}_{-i})$ is quasi-concave in s_i is not necessary.

One condition that Roemer gives for existence of simple Kantian equilibrium is a *common diagonal* condition (23, Proposition 2.1):

$$\forall i, j \in [n]: \quad V^i(s, \ldots, s) = V^j(s, \ldots, s) \tag{5}$$

It is obvious why this guarantees existence: in particular, (5) implies (4). However, notice that condition (5), like condition (4), is not robust: a small perturbation of the utility functions undoes it.

II.III. Multiplicative, Additive, and φ-Kantian Equilibrium

Roemer is well aware of the non-existence problem and indeed uses it to motivate variants of Kantian equilibrium (41–43). Suppose that the strategy space is $S = [0, \infty)$. Define a *Kantian variation* to be a function $\varphi : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}_+$ such that $\varphi(s, 1) = s$ for all $s \in S$.[4]

Definition 4. *A strategy profile $(s_1^*, \ldots, s_n^*)$ is a φ-**Kantian equilibrium** if:*

$$\forall i \in [n], \forall r \in \mathbb{R}: \quad V^i(s_1^*, \ldots, s_n^*) \geq V^i(\varphi(s_1^*, r), \ldots, \varphi(s_n^*, r)) \tag{6}$$

Two special cases of φ-Kantian equilibrium are multiplicative and additive Kantian equilibrium. In the case of multiplicative Kantian equilibrium, the Kantian variation is:

$$\varphi(s, r) = \max\{s \cdot r, 0\} \tag{7}$$

And in the case of additive Kantian equilibrium, the Kantian variation is:

$$\varphi(s, r) = \max\{s + r - 1, 0\} \tag{8}$$

In the case of multiplicative Kantian equilibrium, condition (6) simplifies to:[5]

$$\forall i \in [n], \forall r \in \mathbb{R}_+: \quad V^i(s_1^*, \ldots, s_n^*) \geq V^i(r \cdot s_1^*, \ldots, r \cdot s_n^*)$$

[4] Roemer also assumes that a Kantian variation φ must be such that $\varphi(s, r)$ is increasing and concave in r, but I relax this requirement because it is not important for my purposes.

[5] One problem with multiplicative equilibrium formulated in this way is that $(s_1, \ldots, s_n) = (0, \ldots, 0)$ is always a multiplicative Kantian equilibrium because, for all r, $(r \cdot 0, \ldots, r \cdot 0) = (0, \ldots, 0)$. Thus we should really restrict attention to interior multiplicative Kantian equilibria: that is, $(s_1, \ldots, s_n)$ where $s_i > 0$ for all i.

In the case of additive Kantian equilibrium, condition (6) simplifies to:

$$\begin{aligned}&\forall i \in [n], \forall r \in \mathbb{R}: \\ &V^i\left(s_1^*, \ldots, s_n^*\right) \geq V^i\left(\max\left\{s_1^* + r, 0\right\}, \ldots, \max\left\{s_n^* + r, 0\right\}\right)\end{aligned}$$

The basic idea is that we start from a given strategy profile $\mathbf{s}$, and ask whether there is some one-dimensional deviation from that profile that someone thinks is desirable, where the nature of the deviation is determined by the Kantian variation φ. If everyone agrees that the optimal such deviation is *no* deviation, then we declare $\mathbf{s}$ to be a φ-Kantian equilibrium.

We have the following relation between φ-Kantian equilibrium and simple Kantian equilibrium.

Proposition 1. *Suppose that, for all $s \in S$, $\{\varphi(s, r) : r \in \mathbb{R}\} = S$. Then, $\mathbf{s}^* = \left(s_1^*, s_2^*, \ldots, s_n^*\right)$ is a simple Kantian equilibrium if and only if (1) $\mathbf{s}^*$ is a φ-Kantian equilibrium, and (2) $s_1^* = s_2^* = \cdots = s_n^*$.*[6]

In particular, observe that both the variations (7) and (8) satisfy the assumptions of the proposition, so that the proposition applies to both additive and multiplicative Kantian equilibrium.[7]

Proposition 1 establishes that the property of being a φ-Kantian equilibrium is (under a weak assumption) easier to satisfy than the property of being a simple Kantian equilibrium. Roemer establishes the existence of multiplicative Kantian equilibria for a class of production economies (108, Proposition 7.1), and also in production economies for a broader class of φ-Kantian equilibria (110, Proposition 7.3).

One can also construct settings in which φ-Kantian equilibria fail to exist under conditions under which Nash equilibria exist. Define a *two-player zero-sum game* to be a game with two players such that:

$$\forall \mathbf{s} \in S^2: \quad V^1(\mathbf{s}) + V^2(\mathbf{s}) = 0$$

Proposition 2 below establishes the non-existence of Kantian equilibria in zero-sum games. One feature of zero-sum games is that all outcomes of the game are Pareto efficient (relative to the outcomes that are feasible in the game). As explained in footnote 8, and established formally in the Ap-

[6] This proposition is related to Roemer's Proposition 3.6 (50), which specifically concerns production economies.

[7] To be more precise, the multiplicative equilibrium satisfies the assumptions of the proposition except when $s_i^* = 0$ for some i. So, in the case of multiplicative Kantian equilibrium, the proposition applies to all *interior* equilibria. See footnote 5.

pendix, Proposition 2 can be generalized to games in which all outcomes are Pareto efficient.[8] For example, it applies to any game in which some positively valued resource must be distributed among a group of agents, and the outcomes of the game consist of different ways of dividing the resource among the n agents without throwing any of it away.

Proposition 2. *Let* $([2], S, (V^1, V^2))$ *be a two-person zero-sum game.*

(i) Suppose that there exist $s, s' \in S$ *such that* $V^1(s,s) \neq V^1(s',s')$. *Then a simple Kantian equilibrium does not exist in this game.*

(ii) Suppose that:

$$\forall (s_1, s_2) \in S^2, \exists r \in \mathbb{R}: \quad V^1(s_1, s_2) \neq V^1(\varphi(s_1, r), \varphi(s_2, r)) \tag{9}$$

Then a φ*-Kantian equilibrium does not exist in this game.*

It is natural to observe that zero-sum games are poor candidates for Kantian equilibria because the motivation of Kantian equilibrium essentially involves cooperation, and zero-sum games are inimical to cooperation. However, the point here is just to highlight a problem related to the failure of existence of Kantian equilibrium, and the dependence of existence on the structure of preferences. The problem is that the theory provides a non-empty solution concept only for certain kinds of preferences and not for others. How should Kantian optimizers behave in settings which don't allow for much cooperation? Saying that they simply revert to Nash reasoning does not give us a unified normative theory of behavior across domains.

II.IV. Efficiency

Quite a few of the results in *How We Cooperate* establish that Kantian equilibrium leads to efficient outcomes when Nash equilibrium does not. Continue to assume that $S = \mathbb{R}_+$. Say that a game is *strictly increasing* if, for all i, V^i is strictly increasing in the strategies of all other players $j \neq i$, and *strictly decreasing* if, for all i, V^i is strictly decreasing in the strategies of all other players $j \neq i$. A game is *strictly monotone* if it is either strictly increasing or strictly decreasing. In particular, any simple, multiplicative,

[8] One can generalize Proposition 2 to n-person games G. Instead of assuming that G is zero-sum, assume that the outcome of every strategy profile is Pareto efficient (relative to the set of feasible outcomes in the game). In part (i), assume that there exist strategies s, s', and a player $i \in [n]$, such that $V^i(s, \ldots, s) \neq V^i(s', \ldots, s')$. In part (ii), instead of (9), assume that, for all profiles $(s_1, \ldots, s_n) \in S^n$, there exist an $i \in [n]$, and an $r \in \mathbb{R}$, such that $V^i(s_1, \ldots, s_n) \neq V^i(\varphi(s_1, r), \ldots, \varphi(s_n, r))$.

or additive Kantian equilibrium in a strictly monotone game is Pareto efficient (23, Proposition 2.1; 42, Proposition 3.1; 43, Proposition 3.2). With an additional condition on the Kantian variation φ, a φ-Kantian equilibrium of a strictly monotone game is also Pareto efficient (79, Proposition 4.5). In contrast, in any strictly monotone, continuously differentiable, quasi-economic game,[9] any interior[10] Nash equilibrium is inefficient (44, Proposition 3.3).[11] The significance of strictly monotone games is that they represent situations in which there are positive or negative externalities that take a particularly simple form. The book also contains efficiency results with regard to other specific games.

What is the significance of these results? One thought is that efficiency is in some sense constitutive of successful cooperation. For example, it might be thought that cooperation consists essentially in realizing mutual gains, so that efficiency is necessary and sufficient for successful cooperation. That is, in an inefficient outcome, there are mutual gains that have not been realized, but that can be realized; in an efficient outcome, there are no mutual gains *involving everyone*, and any further movement will amount to a loss for someone.

However, associating efficiency with successful cooperation is misleading: efficiency is neither necessary nor sufficient for the success of cooperation. It is not sufficient because efficiency is compatible with very unequal outcomes, in which one party takes all or almost all of the gains for herself. Nor is it necessary, because it is possible to have quite successful cooperation without realizing *all* mutual gains. Ultimately, the justification for Kantian equilibrium, if it is to capture the idea of cooperation, must be more than just that it leads to efficient outcomes.

It is also important to note that Kantian equilibria can fail to be efficient. Roemer shows that the Battle of the Sexes game, which violates the monotonicity assumption of Roemer's Proposition 2.1, has an inefficient simple Kantian equilibrium (27, Proposition 2.3).[12] Roemer also shows that a failure of efficiency can occur in the presence of altruism (see the discussion on 87; this is a consequence of Proposition 5.3 on 85). I now illustrate the possibility of inefficient equilibria in the context of a simple example. This example can be interpreted in terms of altruistic prefer-

[9] A game is *quasi-economic* if (1) the common strategy space is $S = \mathbb{R}_+$, (2) for all $\mathbf{s}_{-i}$, $V^i(s_i, \mathbf{s}_{-i})$ is quasi-concave in s_i, and (3) $V^i(s_i, \mathbf{s}_{-i}) \to -\infty$ as $s_i \to +\infty$.

[10] Given the common strategy space $S = \mathbb{R}_+$, a Nash equilibrium $(s_1^*, \ldots, s_n^*)$ is *interior* if $s_i^* > 0$ for all agents i.

[11] See also the conditions imposed on Kantian variations I mentioned in footnote 4.

[12] The existence of an inefficient Kantian equilibrium in the Battle of the Sexes depends on the precise parameter values of the game.

ences, but it need not be interpreted in terms of altruism, because one way of interpreting the payoff functions in (10) below is as giving monetary payoffs, which, for any strategy profile, are the same for both players.

Consider a two-player game with strategy space $S = \mathbb{R}_{++}$, and suppose that each player $i = 1, 2$ has the utility function:[13]

$$V^i(s_1, s_2) = \ln(2) + \ln(s_1 s_2) - s_1 - 2s_2 \tag{10}$$

Thus the players have *identical* utility functions over outcomes but control different variables. Player 1 controls strategy s_1 and player 2 controls strategy s_2. The reason for including $\ln(2)$ in the utility function will become evident below (in section II.V).

Then, to solve for a simple Kantian equilibrium, we find s^* that solves:

$$\max_{s} \ \ln(2) + 2\ln(s) - 3s$$

The unique simple Kantian equilibrium is $s^* = 2/3$, and the utility for each player at this Kantian equilibrium is $\ln(8/9) - 2 \approx -2.12$. Note that $s_1^* = s_2^* = 2/3$ is also a multiplicative Kantian equilibrium, and an additive Kantian equilibrium. Note, however, that if, instead of s^*, player 1 chose $s_1 = 1$ and player 2 chose $s_2 = 1/2$, then both players would receive a utility of -2, which is better.

The example shows:

Proposition 3. *(i) It is possible that the unique simple Kantian equilibrium of a game is inefficient.*[14] *(ii) Both multiplicative and additive Kantian equilibria can be inefficient.*[15]

Thus Kantian equilibrium does not provide a general solution to the problem of inefficient equilibria. Indeed, as the above example shows, one can construct very simple games in which Kantian equilibria fail to be efficient.

Moreover, the above example is particularly troubling. Consider the following interpretation. Two individuals face individual decision problems. Each must choose a positive real number. Player 1's utility function is $U^1(s_1) = \ln(s_1) - s_1$. Player 2's utility function is $U^2(s_2) = \ln(s_2) - 2s_2$. Suppose that each player solves their own problem individually. Now suppose that nothing changes but that each player completely internalizes

[13] The strategy space $\mathbb{R}_{++}$ is not closed but it could just as well be $[\varepsilon, +\infty)$ for some small $\varepsilon > 0$.

[14] This part of the proposition also follows from Roemer's Proposition 2.3 (27).

[15] In the game studied above, there exists an inefficient multiplicative Kantian equilibrium, but there also exists another efficient multiplicative Kantian equilibrium.

the other's interests to the extent that it becomes their own (see section III.I on altruism below), so that, noting that $\ln(s_1 s_2) = \ln(s_1) + \ln(s_2)$, each person's utility function becomes $V^i = U^1 + U^2$ for $i = 1, 2$.[16] What would be the best thing for the players to do in this case? It seems clear that each player should simply do as they were doing prior to the altruistic transformation: player 1 should simply maximize U^1 and so choose $s_1 = 1$, and player 2 should maximize U^2 and select $s_2 = 1/2$. The constraint $s_1 = s_2$, which generates the inefficient equilibrium $s_1^* = s_2^* = 2/3$, seems completely unmotivated. Likewise, starting from the strategy profile $(s_1^*, s_2^*) = (2/3, 2/3)$ (which is both an additive and a multiplicative Kantian equilibrium), and considering only joint deviations in line with some Kantian variation also seems completely unmotivated. What this example suggests is that not only does Kantian equilibrium lead to inefficient outcomes in certain circumstances, but also that the reasoning it recommends can sometimes seem quite unwarranted and the conditions under which, and the reasons for which, it is warranted need to be made clearer.

Note finally that in contrast to the Battle of the Sexes, in which the simple Kantian equilibrium Pareto dominates all Nash equilibria (Roemer's Proposition 2.3 on 27), in the example above, the unique Nash equilibrium Pareto dominates the unique simple Kantian equilibrium.

II.V. Strategic Equivalence and Interpersonal Comparability of Strategies

Roemer writes:

> The reader should note the formal similarity between multiplicative Kantian and Nash equilibrium. Both use ordinal preferences only. Each considers a counterfactual: with Nash reasoning, the counterfactual is that I alone change my strategy, whereas in Kantian reasoning, I imagine that all players change their strategies in a prescribed way. (42)[17]

It is not only the ordinality that the two notions have in common but also the lack of need for interpersonal comparison of utilities in verifying the equilibrium criterion.

However, Kantian equilibrium is fundamentally different than Nash equilibrium. In particular, as I argue in this section, it requires *cardinality and interpersonal comparison of strategies* and violates certain traditional

[16] This differs from (10) by the constant $\ln(2)$, but the addition of a constant doesn't really change anything.

[17] The quote presumably applies to other kinds of Kantian equilibria as well.

criteria of *strategic equivalence*. Roemer briefly discusses the point on 28, and related points elsewhere (40, 48–49). That is not necessarily bad—we can often measure strategies on a common scale; certainly it is often easier to measure strategies interpersonally than to do the same for utilities. But a rationale and an interpretation of these features is required.

Consider two strategic games, G and $\hat{G}$, with the same player set $[n]$, and such that within each game all players have a common strategy set:

$$G = \left([n], S, \left(V^i\right)_{i\in[n]}\right) \qquad \text{and} \qquad \hat{G} = \left([n], \hat{S}, \left(\hat{V}^i\right)_{i\in[n]}\right)$$

Call game $\hat{G}$ a *relabeling* of game G if there exists a collection of functions $\mathbf{f} = (f^i)_{i\in[n]}$, called the *relabeling profile*, such that: (1) for all players $j \in [n]$, $f^j\colon S \to \hat{S}$ is a bijection, and, (2) for all $\mathbf{s} = (s_1, \ldots, s_n) \in S^n$ and all $i \in [n]$, $\hat{V}^i(\mathbf{f}(\mathbf{s})) = V^i(\mathbf{s})$, where $\mathbf{f}(\mathbf{s}) = \left(f^1(s_1), \ldots, f^n(s_n)\right)$. Call a relabeling profile $\mathbf{f}$ *positive linear* if, for each of the functions f^i, there exists $\alpha^i > 0$ such that $f^i(s_i) = \alpha^i \cdot s_i$, for all $s_i \in S$. The following sort of result is well known.[18]

Proposition 4. *Let $\hat{G}$ be a relabeling of G with relabeling profile* $\mathbf{f}$*. Then* $\mathbf{s}^*$ *is a Nash equilibrium of G if and only if* $\mathbf{f}(\mathbf{s}^*)$ *is a Nash equilibrium of $\hat{G}$.*

The result applies to Nash equilibrium but one might think that it should apply more generally to any reasonable solution concept insofar as it seems that a relabeling of strategies should have no impact on the solution in essential respects. So if s_i is relabeled as s'_i and s_i was part of an equilibrium prior to the relabeling, s'_i should be part of a corresponding equilibrium in the relabeled game. I will have more to say about this below.

Now consider a two-player game with strategy space $S = \mathbb{R}_{++}$ and suppose that each player $i = 1, 2$ has the utility function:[19]

$$\hat{V}^i(s_1, s_2) = \ln(s_1 s_2) - s_1 - s_2 \tag{11}$$

Then the simple Kantian equilibrium is the solution to:

$$\max_s \; 2\ln(s) - 2s$$

[18] This result applies to pure-strategy equilibria, but a similar result applies to mixed equilibria. See, for example, Gabarró, García, and Serna (2011) for more details.

[19] The strategy space $\mathbb{R}_{++}$ is not closed but it could just as well be $[\varepsilon, +\infty)$ for some small $\varepsilon > 0$.

This expression is maximized at $s = 1$ and each agent gets a utility of -2. Note that this is also a multiplicative and an additive Kantian equilibrium, and it is Pareto efficient.

Recall the game $G = ([2], \mathbb{R}_{++}, (V^1, V^2))$ from the previous section (section II.IV) with utility functions (10), and let $\hat{G} = \left([2], \mathbb{R}_{++}, \left(\hat{V}^1, \hat{V}^2\right)\right)$ be the game just described with utility functions (11). Consider the positive linear relabeling $\mathbf{f} = (f^1, f^2)$ for which $f^1(s_1) = s_1$ and $f^2(s_2) = 2s_2$. This transforms the game given by the utility functions V^i in (10) into the game given by the utility functions $\hat{V}^i$ in (11).[20] Notice that while $(s_1^*, s_2^*) = (2/3, 2/3)$ is the unique simple Kantian equilibrium in G and also an additive and a multiplicative equilibrium in G, the relabeled strategy profile $\mathbf{f}(s_1^*, s_2^*) = (2/3, 4/3)$ is neither a simple nor an additive Kantian equilibrium in $\hat{G}$.[21] $\mathbf{f}(s_1^*, s_2^*)$ is, however, a multiplicative Kantian equilibrium of $\hat{G}$ (because the relabeling profile is positive linear). Notice, however, that $\mathbf{f}(s_1^*, s_2^*)$ is an inefficient multiplicative Kantian equilibrium of $\hat{G}$ that is dominated by the multiplicative Kantian equilibrium $(s_1^{**}, s_1^{**}) = (1,1)$ in $\hat{G}$.[22] However, if instead one applied the nonlinear transformation $\tilde{\mathbf{f}} = \left(\tilde{f}^1, \tilde{f}^2\right)$ with $\tilde{f}^1(s_1) = \sqrt{s_1}$ and $\tilde{f}^2(s_2) = s_2$ to the game G, then $\tilde{\mathbf{f}}(s_1^*, s_2^*)$ is not a multiplicative Kantian equilibrium in the resulting game.[23] More generally, we have:

Proposition 5. *The following three results hold:*

(i) Simple Kantian equilibrium, additive Kantian equilibrium, and multiplicative Kantian equilibrium are not in general preserved under the

[20] In particular, observe that the transformations (f^1, f^2) of strategies induce the transformations $\hat{V}^i$ of the utility functions V^i. To confirm this, observe that when we plug in the transformed strategy profile $(f^1(s_1), f^2(s_2))$ into the transformed utility function $\hat{V}^i$, using the fact that $\ln(s_1 \cdot 2s_2) = \ln(2) + \ln(s_1 s_2)$, we recover the original utility function, as required:

$$\hat{V}^i\left(f^1(s_1), f^2(s_2)\right) = \hat{V}^i(s_1, 2s_2) = \ln(2) + \ln(s_1 s_2) - s_1 - 2s_2 = V^i(s_1, s_2)$$

[21] Clearly $(2/3, 4/3)$ cannot be a simple Kantian equilibrium because $2/3 \neq 4/3$. Observe that:

$$\frac{\mathrm{d}}{\mathrm{d}r}\bigg|_{r=0}\left[\ln\left(\frac{2}{3}+r\right)+\ln\left(\frac{4}{3}+r\right)-\left(\frac{2}{3}+r\right)-\left(\frac{4}{3}+r\right)\right] = \frac{3}{2}+\frac{3}{4}-2 = \frac{1}{4} \neq 0$$

It follows that $(2/3, 4/3)$ is not an additive Kantian equilibrium.

[22] (s_1^{**}, s_1^{**}) is also a simple and an additive Kantian equilibrium of $\hat{G}$.

[23] In particular, consider the two-player game with strategy space $\mathbb{R}_{++}$ and utility functions:

$$\tilde{V}^i(s_1, s_2) = \ln(2) + 2\ln(s_1) + \ln(s_2) - s_1^2 - 2s_2$$

relabeling of strategies. That is, for each of these types of Kantian equilibria, there exists a game G, and a relabeling $\hat{G}$ with relabeling profile $\mathbf{f}$*, such that, for some Kantian equilibrium* $\mathbf{s}^*$ *of* G*,* $\mathbf{f}(\mathbf{s}^*)$ *is not a Kantian equilibrium of* $\hat{G}$*.*

(ii) *If $\hat{G}$ is a relabeling of G with positive linear relabeling profile* $\mathbf{f}$ *such that, for some i and j, $f^i \neq f^j$, then, for every simple Kantian equilibrium* $\mathbf{s}^* = (s^*, \dots, s^*)$ *of G with $s^* > 0$,* $\mathbf{f}(\mathbf{s}^*)$ *is not a simple Kantian equilibrium of* $\hat{G}$*.*

(iii) *If $\hat{G}$ is a relabeling of G with positive linear relabeling profile* $\mathbf{f}$ *such that $S = \hat{S} = \mathbb{R}_+$, then* $\mathbf{s}^*$ *is a multiplicative Kantian equilibrium of G if and only if* $\mathbf{f}(\mathbf{s}^*)$ *is a multiplicative Kantian equilibrium of* $\hat{G}$*.*[24]

It is instructive to contrast Proposition 5 with Proposition 4. Nash equilibrium is invariant to relabeling whereas Kantian equilibrium is not.

This is not necessarily a decisive objection to Kantian equilibrium: different solution concepts may have different informational requirements. But it does mean that there are some suppressed principles that must determine what the *right* way of measuring strategies is. These principles ought to be made explicit. If we are just given a game abstractly via its utility functions, as in (10), we don't know whether it has been presented in such a way that the solution concept of Kantian equilibrium can be applied. This contrasts with Nash equilibrium, for which utility information is sufficient. In some cases, such as many examples in *How We Cooperate*, it may be obvious that different agents' strategies are measured in

Observe that $\tilde{V}^i\left(\tilde{\mathbf{f}}(s_1, s_2)\right) = \tilde{V}^i\left(\sqrt{s_1}, s_2\right) = V^i(s_1, s_2)$, and we have:

$$\begin{aligned}
\left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=1} \tilde{V}^i\left(r\tilde{f}^1(s_1^*), r\tilde{f}^2(s_2^*)\right) &= \left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=1} \tilde{V}^i\left(r\sqrt{\frac{2}{3}}, r\frac{2}{3}\right) \\
&= \left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=1} \left[\ln(2) + 2\ln\left(r\sqrt{\frac{2}{3}}\right) + \ln\left(r\frac{2}{3}\right) - \right. \\
&\qquad \left. - \left(r\sqrt{\frac{2}{3}}\right)^2 - 2r\frac{2}{3}\right] \\
&= 2 + 1 - 2\cdot\frac{2}{3} - 2\cdot\frac{2}{3} = \frac{1}{3} \neq 0
\end{aligned}$$

This implies that $\tilde{\mathbf{f}}(s_1^*, s_2^*)$ is not a multiplicative Kantian equilibrium of the relabling $\tilde{G}$ of G corresponding to $\tilde{\mathbf{f}}$.

[24] Part (iii) can be generalized. It holds if there exists a $k > 0$ such that, for all $i \in [n]$ and for all $r > 0$, $f^i(rs_1, \dots, rs_n) = r^k f^i(s_1, \dots, s_n)$, or, in other words, if all the f^i functions are homogeneous to the same degree.

the same natural units, and we may take this canonical way of measuring strategies as an input that is necessary for analyzing the game via Kantian equilibrium. However, cooperation is not restricted to situations in which the strategy spaces of different players are the same. Sometimes different players in the game have to make different kinds of choices, and it is not clear how the theory would extend to such cases.

Part (iii) of the theorem shows that if, for each player, one can choose a *privileged ratio scale* on which to measure the players' strategies, then interpersonal comparisons of strategy spaces are not required for *multiplicative* Kantian equilibrium. Part (ii) shows that the same is not true for simple Kantian equilibrium. But notice that a given underlying reality can in general be measured using multiple non-equivalent scales. So there has to be some *choice* of scale even in the best case. In some cases, there may be an obvious natural choice, and in others not.

In the examples above, the relabelings f^i were allowed to be idiosyncratic to individuals. One might wonder what happens if we restrict attention to relabelings that are the same for all individuals. Say that a relabeling $\hat{G}$ of G is *uniform* if the corresponding relabeling profile $\mathbf{f} = (f_i)_{i \in [n]}$ is such that, for all $i, j \in [n]$, $f^i = f^j$. With respect to uniform relabelings, we then have:

Proposition 6. *The following two results hold:*

(i) If $\hat{G}$ is a uniform relabeling of G with relabeling profile $\mathbf{f}$, then $\mathbf{s}^*$ is a simple Kantian equilibrium of G if and only if $\mathbf{f}(\mathbf{s}^*)$ is a simple Kantian equilibrium of $\hat{G}$.[25]

(ii) For both additive and multiplicative Kantian equilibria, there exists a game G and a uniform relabeling $\hat{G}$ with relabeling profile $\mathbf{f}$, such that for some Kantian equilibrium $\mathbf{s}^*$ of G, $\mathbf{f}(\mathbf{s}^*)$ is not a Kantian equilibrium of $\hat{G}$.

Notice that while simple Kantian equilibria are preserved under uniform relabelings, Nash equilibria are also preserved under *nonuniform* relabelings. So, again, Nash equilibria are preserved under a broader class of intuitively 'strategically irrelevant' transformations. Additive and multiplicative Kantian equilibria are not even in general preserved under uniform relabelings.

[25] I am grateful to Marina Uzunova for suggesting this part of the proposition.

In general, the important lesson that emerges in this section is that Kantian equilibrium does *not* just depend on utility information, but also on some *normatively privileged measurement of strategies.*

III. Kantian Optimization Cannot Be Justified in Terms of Self-Interest

This section argues that Kantian optimization cannot be justified in terms of self-interest. Section III.I discusses altruism, as opposed to self-interest, while sections III.II and III.III argue that Kantian equilibrium cannot be justified purely in terms of self-interest.

III.I. Altruism

In motivating the Kantian equilibrium approach to cooperation, Roemer contrasts it with two other approaches that are common in economics: (1) a foundation for cooperation in terms of *altruism*, and (2) a foundation for cooperation in terms of *far-sighted self-interest* and repeated interaction. With respect to (2), Roemer writes:

> Until behavioral economics came along, the main way of explaining cooperation—which here can be defined as the overcoming of the Pareto inefficient Nash equilibria that standardly occur in games—was to view cooperation as a *Nash* equilibrium of a complex game with many stages. (7)

Roemer argues against both of these approaches. Here I will focus on the first approach in terms of altruism. I mention in passing that I take issue with the characterization of the problem of cooperation in the above quotation for the reasons that I gave in section II.IV.

It is worthwhile to start by saying a word about what altruism is. Richard Kraut (2020), for example, writes: “Behavior is normally described as altruistic when it is motivated by a desire to benefit someone other than oneself for that person’s sake.” Kraut’s definition is in terms of *behavior* and *motives.* In contrast, economists often talk in terms of altruistic *preferences* (noting that in economic theory, preferences and behavior are typically taken to be closely related, even definitionally).[26] Motives and preferences are related but distinct concepts. In the case of allocating some good among different individuals, we may represent i’s altruistic

[26] Viewing behavior and preferences as definitionally related amounts to a flaw in economic theory, in my view.

preferences by a utility function of the form

$$U^i(\mathbf{x}) = u^i\left(x^i\right) + \alpha^i \sum_{j \neq i} u^j\left(x^j\right) \tag{12}$$

where x^i is the amount of the good allocated to agent i, $\mathbf{x} = (x^1, \dots, x^n)$ is the entire allocation, and $u^i\left(x^i\right)$ is a measure of the value of x^i to person i. α^i measures the extent to which i weighs the interests of others, with $\alpha^i = 0$ corresponding to pure selfishness, and $\alpha^i = 1$ corresponding to pure altruism.[27]

Translating between a formal representation (12) and its meaning with regard to altruism is not as straightforward as it may appear.[28] I briefly mention a few relevant issues that I don't have space to expand on here. Suppose that U^i represents i's decision utility: that is, the function whose maximization determines or represents the decisions that i would make in various circumstances. That leaves open the different question of whether i's interests or well-being is represented by U^i or u^i (or something else). Also, it leaves open the question of what i's *reasons* are for choosing so as to maximize the altruistic objective U^i. Is it because helping others makes i feel good? Is it because i cares about other people? Is it because i feels a moral duty to help others? Exploring these questions would take us too far afield, but it is important to keep in mind that the simple utility representation in (12) leaves open important questions about the nature of altruism.[29]

III.II. Self-Interest vs Altruism as Bases for Cooperation

Roemer is critical of altruism as a basis for cooperation. He writes: "*Altruism* and cooperation are frequently confounded in the literature" (5). And, further:

> My claim is that the ability to cooperate for reasons of self-interest is less demanding than the prescription to care about others. I believe that it is easier to explain the many examples of human cooperation from an assumption that people learn that cooperation can further their own interests than to explain those examples by altruism. (5)

[27] Note that in order for (12) to make sense from i's point of view, the utility functions u^i and $(u^j)_{j \neq i}$ cannot represent merely ordinal preferences, but rather must have cardinal significance, and, moreover, must be interpersonally comparable.

[28] See Roemer's related discussion of different interpretations of altruism on 93–94.

[29] Sen (1977) discusses themes related to those in this section.

This claim is both descriptive and normative. It is descriptive insofar as it makes a claim about what *does* motivate people, but it is normative insofar as it claims that self-interest *can* provide a *justification* for cooperation, specifically via Kantian equilibrium.

Let us then consider the claim that Kantian equilibrium is founded on self-interest rather than on altruism. It is not clear in what sense self-interest can serve as a foundation for cooperation in Kantian equilibrium. It is clear how self-interest can serve as a basis for cooperation in the repeated-game foundation of cooperation: "an individual has self-interested preferences but helps another individual as part of a Nash equilibrium in a game with stages, or a repeated game, in an equilibrium with reciprocation" (93). However, this is not the self-interested foundation that Roemer advocates. Roemer advocates, rather, Kantian equilibrium, and not Nash equilibrium in a game with multiple stages as the means to cooperation.

Explaining how self-interest founds cooperation, Roemer writes:

> *Solidarity* is defined as 'a union of purpose, sympathies, or interests among the members of a group' (*American Heritage Dictionary*). [...] Solidarity, so construed, is not the cooperative action that the individuals take but rather a characterization of their objective situation: namely, that all are in the same boat and understand that fact. I take 'a union of interests' to mean that we are all in the same situation and have common preferences. It does not mean we are altruistic toward each other. Granted, one might interpret 'a union of ... sympathies' to mean altruism, but I focus rather on 'a union of purpose or interests.' (4)

And:

> The key point is that cooperation of an extensive kind can be undertaken because it is in the interest of *each*, not because each cares about others. I am skeptical that humans can, on a mass scale, have deep concern for others whom they have not even met, and so to base grand humanitarian projects on such a psychological propensity is risky. I do, however, believe that humans quite generally have common interests and that it is natural to pursue these cooperatively. [...] It seems that the safer *general* strategy is to rely on the underlying motive of self-interest, active in cooperation, rather than on love for others, active in altruism. (5)

But does the formal framework of Kantian equilibrium validate the claim that self-interested motivation can lead to cooperation? Consider a person's Kantian optimization problem:

$$\max_{s \in S} V^i(s, \ldots, s) \tag{13}$$

It may seem that in solving this problem, the agent is acting out of self-interest rather than altruism, because it is only the agent's own utility function V^i that is being optimized and not a utility function like $\sum_j V^j$ (or a weighted sum), which takes account of all agents' utilities.

But this appearance of the embodiment of self-interest in (13) is not straightforward for a number of reasons. First, in the general abstract formulation of (13), we don't know what the utility function V^i is, and hence whether it in fact involves altruistic considerations.[30] Second, what is being optimized is not the strategy that agent i will choose—which i has control over—but the strategies that *all* agents will choose, including the strategies that other agents j, and not i, control. Why should we think of an agent who is simply pursuing their own self-interest as optimizing over actions that they themselves do not control?[31] Is it because this choice is to be understood as the result of an agreement reached by the different agents over the actions that they jointly control, or as the result of a social norm?[32] If so, what is to enforce the agreement or norm? Punishments or other incentives? If so, we are back to something like the far-sighted repeated-game account of cooperation. It is true that the Kantian equilibrium is the agreement that one would self-interestedly want everyone to reach *if* facing the constraint that everyone choose the same strategy, but what would bind the agent to this constraint? If it is a sense of fairness or solidaristic duty to the group, then the motive has a moral aspect and is not purely self-interested.

Third, the Kantian equilibrium requires not just that s^* maximize $V^i(s, \ldots, s)$ for the agent i on whom we are focused but that s^* also maximize $V^j(s, \ldots, s)$ for all $j \neq i$. If s^* maximizes $V^i(s, \ldots, s)$ but not $V^j(s, \ldots, s)$, then s^* is not a Kantian equilibrium. So in fact the criterion involves maximization of *all* agents' utility functions, and indeed in a symmetric way. So in what sense is the Kantian equilibrium criterion

[30] Indeed, this possibility is explored in chapter 5 of *How We Cooperate*.

[31] A similar question might be posed for an agent not maximizing their own self-interest, but rather some other objective. See the discussion in section V below.

[32] See the discussion on 21–22 of *How We Cooperate*. There, Roemer claims that it is actually Kantian optimization that determines the norms. But even if that were so, the questions that follow in the text above about what enforces the norms still have the same force.

self-interested? The formal criterion appeals to the interests or objectives V^i of *all* agents, not just a single agent i. Unlike Nash equilibrium, which also appeals to maximization of all V^i functions, but which can be interpreted in a self-interested way, each agent does not maximize subject to the others' choices, but rather all agents' interests are simultaneously maximized subject to some self-imposed constraint. Intuitively, the Kantian equilibrium criterion seems to be concerned with the maximization of everyone's interests.

III.III. Nash Optimization vs Kantian Optimization

The book often frames the distinction between traditional economics and the project it proposes as the difference between Nash optimization and Kantian optimization. Under Nash optimization, other players' strategies are taken as given, whereas under Kantian optimization, optimization is simultaneously over all people's strategies. The book advocates Kantian optimization.[33]

One criticism of Kantian optimization is that when optimizing any objective, one should optimize over the actions that one can control. The reason that, in Nash optimization, the actions of others are held fixed is that one has no control over the actions of others. Analogously, if we are not talking about a game in which there are other players, but rather a decision problem, one should optimize over the aspects of the situation that one can control. That one should optimize over what one can control is the *reason* that actions of others are held fixed in Nash equilibrium. Indeed, even under weaker solution concepts such as rationalizability (Bernheim 1984; Pearce 1984),[34] agents are thought to maximize against their (possibly mistaken) beliefs as to what others will do (where those beliefs are constrained by common knowledge of rationality). More generally, if we allow for the possibility that others make mistakes, then if an agent assumes that others will play specific strategies—rational or not—the agent

[33] Ideas like Kantian optimization have been put forward before. It is not uncommon for people to suggest cooperation in the Prisoner's Dilemma game because that is what one would like everyone to do. A common criticism is that this recommendation involves magical thinking because to be a rational prescription it would need to implicitly presuppose that one player deciding to cooperate will *cause* the other player to cooperate, which is false. For a criticism of such arguments, see Dekel and Gul (1997). At 21–22, Roemer says that his argument—what he calls "Method Two" (19)—does not invoke such magical thinking and is distinct from it. As I shall argue below, there is no good argument for invoking only self-interest in favor of taking the cooperative action in the (one-shot) Prisoner's Dilemma.

[34] Rationalizability is a solution concept that encodes the consequences of common knowledge of rationality but does not require that agents make correct predictions about the behavior of others.

should optimize, holding fixed their beliefs about others' strategies; and if one merely has probabilistic beliefs over others' strategies—rational or not—one should optimize an expectation given those beliefs. In no case does one maximize over things—controlled by other people or by nature—that one oneself does not control.

I now go over this argument a little more formally. Suppose that $\mathbf{s}^*$ is a Kantian equilibrium (of any kind: simple, additive, multiplicative, φ). That is consistent with the possibility that, for some $s_i' \in S$:

$$V^i\left(s_i', \mathbf{s}_{-i}^*\right) > V^i\left(s_i^*, \mathbf{s}_{-i}^*\right) \tag{14}$$

And indeed Kantian equilibria will often allow (14) to occur.[35] If some person i expects everyone else to play as in $\mathbf{s}^*$ and is purely self-interested, then why shouldn't such a person choose s_i' rather than s_i^* from a self-interested perspective? If i expects others to play some other strategy profile $\mathbf{s}_{-i}$, why shouldn't i select whichever strategy s_i it is that maximizes $V^i\left(s_i, \mathbf{s}_{-i}\right)$? If i has probabilistic beliefs p_{-i} over the strategies of the others, why shouldn't i select whichever strategy s_i it is that maximizes $\sum_{\mathbf{s}_{-i}} u_i\left(s_i, \mathbf{s}_{-i}\right) \cdot p_{-i}\left(\mathbf{s}_{-i}\right)$? It seems that if there is an argument for choosing s_i^*, it cannot just appeal to self-interest; it must appeal to other notions: either solidarity, or fairness, or altruism, or something else. But all of these concepts, *including solidarity*, are moral concepts that in some sense go beyond mere self-interest. It may be that Kantian equilibrium identifies what it is for a person to be *doing their part*. But if this is so, then the justification for doing one's part—the argument that one *should* do one's part—must go beyond mere appeal to one's self-interest and must appeal to some moral considerations.

One might reply that the above argument is question-begging and that it starts off by privileging Nash optimization over Kantian optimization, whereas that is what is at issue here. But I don't think it is question-begging. Nash optimization and Kantian optimization are technical terms, and what one really needs to appeal to are *reasons* to play in one way or another. I have been arguing that, from a purely self-interested perspective, there are no good reasons to play Kantian equilibrium; one must rather appeal to moral reasons in order to justify Kantian play.

[35] In games for which the Nash equilibria are inefficient and Kantian equilibria are efficient, a violation of the form (14) will always occur for some player i at any Kantian equilibrium. See the results discussed in the beginning of section II.IV above.

	Stag	Hare
Stag	(2,2)	(0,1)
Hare	(1,0)	(1,1)

Table 1: The Stag Hunt.

IV. The Problem of Trust

Roemer emphasizes that trust is a key ingredient, along with self-interest, for Kantian equilibrium. He writes: “One often thinks of trust as key in cooperative situations [. . .]. I think of trust as induced by the assumptions of common knowledge and common capacity” (20). The discussion of trust in sections 2.1 and 9.3 of *How We Cooperate* is interesting. However, one problem with the notion of Kantian equilibrium is that it does not provide any formalization of the reason that trust is important.

It will be useful to contrast Kantian equilibrium with Nash equilibrium for the purpose of evaluating trust. Consider the Stag Hunt game (Table 1). The explanation of this game comes from Jean-Jacques Rousseau:

> If a deer was to be taken every one saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs. (Rousseau [1755] 1923, 209–210)

In the above game, the action *Stag* corresponds to staying at one’s post, which, if done by both players, will cause the stag to be caught, yielding a payoff of 2 for each player. The action *Hare* corresponds to chasing the hare, which will cause an agent to catch the hare but the other player, if he stays at his post, to catch nothing. It is assumed that catching the hare is less good than having a share of the stag.

The cooperative outcome in this game is (*Stag*, *Stag*), and it is also a Nash equilibrium. It is clear why, from the standpoint of Nash equilibrium, two players who were playing this game would need to trust one another. It is only worthwhile for Ann to play *Stag* if she expects Bob to play *Stag* as well. If Bob were to deviate and play *Hare* (perhaps because he too didn’t trust Ann), *Stag* would lead to a low payoff for Ann and Ann would be better off playing *Hare* as well.[36]

In contrast, consider the Prisoner’s Dilemma game in Table 2. Here, the dominant strategy is for players to defect, but mutual cooperation

[36] Both (*Stag*, *Stag*) and (*Hare*, *Hare*) are Nash equilibria of the Stag Hunt.

	Cooperate	Defect
Cooperate	$(0,0)$	$(-0.5,1)$
Defect	$(1,-0.5)$	$(-0.25,-0.25)$

Table 2: The Prisoner's Dilemma.

Pareto dominates mutual defection. Let us consider the Kantian equilibrium of the *mixed extension* of this game, that is, the Kantian equilibrium of the game in which the players choose mixed strategies, so that the strategy choices are the probabilities of playing cooperate. The payoff to each player if both players choose the same probability p of cooperating is:

$$\left[0\cdot p^2\right]-\left[0.5\cdot p\left(1-p\right)\right]+\left[1\cdot\left(1-p\right)p\right]-\left[0.25\cdot\left(1-p\right)^2\right]$$

This expression simplifies to:

$$0.5p\left(1-p\right)-0.25\left(1-p\right)^2$$

The Kantian equilibrium is the probability p^* of cooperation that solves:

$$\max_p\left[0.5p\left(1-p\right)-0.25\left(1-p\right)^2\right]$$

The solution is:

$$p^*=\frac{2}{3}$$

(See Proposition 2.2 in Roemer 2019, 25.)

The question is: if players are to play the Kantian equilibrium, why should Ann care about whether Bob cooperates in this game? More precisely, why should Ann base her decision on the assumption that Bob cooperates? That is, why should she make a different decision if she expects Bob to cooperate and play p^* than if she does not? Notice that if Ann and Bob both play p^*, in the Kantian equilibrium of the Prisoner's Dilemma, Ann's payoff is:

$$\frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{3}-\frac{1}{4}\cdot\left(\frac{1}{3}\right)^2=\frac{1}{12}$$

In contrast, if Bob deviates to his best reply and plays *Defect*, then in playing p^*, Ann's expected payoff would be lowered from 1/12 to:

$$-\frac{2}{3}\cdot\frac{1}{2}-\frac{1}{3}\cdot\frac{1}{4}=-\frac{5}{12}$$

So Ann depends on Bob to play p^* in order to maintain her payoff. But notice that no matter what Bob does—whether Bob cooperates with probability 1, or defects with probability 1, or cooperates with probability p^*, or with any other probability p—Ann will be better off if she defects than if she cooperates. So why should Ann's decision to play p^* hinge on Bob playing precisely p^* rather than something else? Ann has an incentive to defect if Bob defects, but she also has an incentive to defect if Bob plays p^*.

One might say that the reason that Ann should only play p^* if Bob does is that it is not fair for Ann to bind herself to her part of the Kantian equilibrium if Bob does not do his part, harming Ann as a consequence. But notice that the appeal to *fairness* is a moral appeal, not a self-interested appeal. Alternatively, one might say that the reason is that if Bob does not do his part, then the *collective* goal of coordinating on p^* is not met, but this is a collective, and not a purely individual goal. Whatever the reason, it is not *formalized* as part of the solution of Kantian equilibrium: there is no formalism for how one might condition one's play on the basis of expected fairness of the other player or on the expected success of the collective goal. In the case of Nash equilibrium, the notion of a *best response* formalizes the dependence of one player's choice on another's. In the case of Kantian equilibrium, there is no corresponding notion formalizing this dependence. This is especially clear in the case of *simple* Kantian equilibrium. Lacking an account of how behavior is to be conditioned on fair play by the other, or solidarity by the other, it is not clear why Ann should do her part only if she expects Bob to do his.[37] And certainly, from a purely self-interested perspective, there is no reason why Ann should stick with the Kantian equilibrium if and only if she expects Bob to do so.

I have discussed the Nash equilibrium of the Stag Hunt and the Kantian equilibrium of the Prisoner's Dilemma. To round out the discussion, let us consider the Nash equilibrium of the Prisoner's Dilemma and the Kantian equilibrium of the Stag Hunt. (*Defect*, *Defect*) is the Nash equilibrium of the Prisoner's Dilemma. This equilibrium does not require trust, as the best response is *Defect* regardless of what the other player does; there is no dependence of the best response on the other's strategy. So, Nash equilibrium does not require trust in every game; but as we have seen above in connection to the Stag Hunt, Nash equilibrium is *compati-*

[37] In section 9.3 (134–136), Roemer discusses this, stating that people are *conditional cooperators* who cooperate if they expect a high enough proportion of others to cooperate, but I think the ideas found there could benefit from a stronger foundation.

ble with the importance of trust.[38] But a proponent of Nash equilibrium would not say that the (*Defect*, *Defect*) equilibrium depends on trust. In contrast, Roemer would want to say that the Kantian equilibrium of the Prisoner's Dilemma relies on trust. But, again, as we have seen, there is no justification for this claim. Finally, observe that the unique simple Kantian equilibrium of the Stag Hunt is the pure strategy equilibrium (*Stag*, *Stag*). This was also the (non-unique) Nash equilibrium strategy profile that we discussed above. However, whereas in the case of Nash equilibrium, playing *Stag* requires trust because the *best response* to *Hare* is *Hare* rather than *Stag*, so one needs to know what the other is doing to know what one should do, the Kantian equilibrium of (*Stag*, *Stag*) does not appeal to the notion of a best response. So, it is not clear how the Kantian equilibrium of (*Stag*, *Stag*) depends on trust, because just as in the Prisoner's Dilemma there is no formalism in Kantian equilibrium that makes its prescription conditional on an expectation of what the other player will do.

V. A Moral Justification for Kantian Equilibrium

In sections III and IV, I have argued that Kantian equilibrium cannot be given a purely self-interested justification. That is, there do not exist purely self-interested reasons for an agent to play their part in a Kantian equilibrium. I want to clarify that here I am not talking about the psychology of Kantian equilibrium, which may make it appealing or natural for people to play their part in a Kantian equilibrium (for a discussion of the psychology, see Elster 2017), but rather about the way a player might validly justify play of their Kantian equilibrium strategy as a basis for cooperation.

A justification for playing Kantian equilibrium requires appeal to some moral considerations. In this section, I discuss the possibility of a moral foundation for Kantian equilibrium. I also discuss the connection to *collective intentions* and *team agency*, which is related.

V.I. Morality

To think about the foundation for Kantian equilibrium, it is important to distinguish between two types of question:

[38] Note that I do not need to assume that whenever the best response depends on the other player's strategy, this is always naturally interpreted in terms of trust. I claim only that in some games, like the Stag Hunt, it is natural to interpret the game with reference to trust.

(1) **Individual question.** What should an individual do unilaterally in order to further a given objective O? What should an individual do unilaterally to obey duties D or respond to reasons R?

(2) **Social question.** What is the best thing for a group to do collectively in order to further a given objective O? What sorts of institutions and norms should groups employ to best fulfill collective duties D or respond to reasons R?

I will initially focus on the furthering of an objective O rather than obeying duties D or responding to reasons R. The objective O can be either selfish or moral (or anything else). For example, if we take the selfish objective (from Bob's point of view) of furthering Bob's interests, versions of the first question are: 'What can Bob do, holding others' behavior fixed, to best further Bob's interests?', and 'What should Bob do unilaterally, given Bob's beliefs about how others will behave, to best further Bob's interests?'. Versions of the second question are: 'What social arrangement best furthers Bob's interests?', and 'What can everyone do collectively to best further Bob's interests?'. If we take the objective O to be the moral objective associated with utilitarianism—maximizing aggregate utility—then one version of the first question corresponds to a kind of act-utilitarianism: 'What can Bob do, holding others' behavior fixed, to maximize aggregate utility?'. And a version of the second question is: 'What can people do collectively to maximize aggregate utility?'.

Kantian equilibrium, like some other moral ideals, seems to operate both at the social and individual levels, so that it implicates both types of question above. The scheme that I am about to describe can be viewed as an instance of *team reasoning*, which I shall discuss in section V.II. It can be natural to first ask the social question and then use the answer to address the individual question. In particular, first we ask the social question: how should a group act cooperatively so as to best achieve everyone's goals? Suppose that the strategy profile $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ is the strategy profile that is best from the collective standpoint. Perhaps it best embodies a fair scheme of cooperation. Then, at the *individual* level, each agent i has a *moral* reason to do their part—namely to select s_i^*—in the cooperative scheme. The strategy profile $\mathbf{s}^*$ is determined by social considerations, but each individual i is then enjoined to select s_i^*, which is the part of the scheme that they can control. Notice that, crucially, each agent has a *moral* reason to select s_i^*, not merely a self-interested reason: the individual has reasons to do her part in a larger cooperative enterprise, which affects her interests and also those of others, not just to further her own narrow interests. If she only cared about her own personal interests—rather than also about doing her part in the cooperative scheme—she would have no reason not to deviate from the collective plan in any way that benefited her.

Note that the pattern of reasoning described in the previous paragraph is not unique to Kantian equilibrium. We could use similar reasoning with regard to other moral theories. For example, we could use the same approach with regard to utilitarianism. We may first ask the social question: which norms, institutions, or habits would maximize the utilitarian objective $\sum_i V^i$? Then, on an individual level, we may enjoin each individual to do their part in the utilitarian scheme. A scheme of *cooperative utilitarianism* along these lines was advocated by Regan (1980).

Let us consider Kantian equilibrium specifically. What are the objectives, duties, and reasons that might justify a person behaving according to the Kantian equilibrium prescription? Rather than seeking to maximize the sum of these utilities, $\sum_i V^i$, we attempt to maximize each utility function V^i individually, either because the utility functions V^i are not interpersonally comparable or because we think that maximizing the utility functions individually is a better ideal. However, in general, it is not possible to maximize all V^i functions simultaneously: there is a trade-off between the different objectives V^i. The way that Kantian equilibrium attempts to resolve this trade-off is by limiting the class of admissible strategy profiles. It does this either by the constraint that all strategies be the same, $s_1^* = \cdots = s_n^*$ (in simple Kantian equilibrium), or by restricting the class of permissible deviations to lie along some Kantian variation φ. The idea is that while, globally, there may be a conflict between the different V^i, we can find some joint constraint on strategies such that interests are in harmony subject to that constraint.

The moral justification for this procedure is clearest in the case of *simple* Kantian equilibrium. If there is one action such that it would be best for each of us if we all took that action, rather than any other common action, it seems plausible that, out of solidarity, we ought all to take that action. However, this solidarity is itself a *moral* notion; it is not purely self-interested. And it implicates other moral notions such as *fairness* and a recognition that the *interests of others* are important as well.

This moral foundation helps to fill the gap left by a justification in terms of self-interest. With pure self-interest—once we set aside far-sighted Nash equilibrium in a repeated-interaction or complex game—there is no justification for sticking with one's Kantian equilibrium strategy rather than deviating to one's best response. In contrast, if one has a moral motive, then one can justify sticking with the Kantian equilibrium by appealing to the considerations that it would be unfair to deviate, that one has an obligation to do one's part, or that deviating would harm others.

There are problems with Kantian equilibrium as a moral ideal. At an abstract level, we saw in sections II.IV and II.II that Kantian equilibrium can be inefficient and that it might not exist. So other non-Kantian schemes might in some circumstances better advance collective interests or the Kantian scheme may simply fail to yield advice. More concretely, Kantian equilibrium straightforwardly enjoins agents to act in solidarity with others *who have power to contribute to the collective good*, but it is not clear whether it promotes solidarity with the powerless.[39] Consider a two-player game in which there is also a bystander with no power, who we will call player 3, and who is affected by the choices of players 1 and 2 but does not herself choose a strategy. The strategy s^* that jointly maximizes $V^i(s,s)$ for $i = 1,2$ may be very bad for player 3 in comparison to other strategy choices. It is not clear how Kantian equilibrium should be extended to such a setting (where one player is merely a bystander), but *if* we still regard s^* as a Kantian equilibrium in this setting, then we see that it ignores the powerless player 3's interests, which would make it problematic as a moral ideal. More generally, Kantian equilibria depend not only on the interests of players but also on their powers—on the relation between their strategic choices and outcomes. The theory of Kantian equilibrium seems to enjoin solidarity among those who can cooperate to benefit one another, but it is at best silent about what should be done to benefit those who are not in a position to assist in cooperation. Relatedly, consider a game that is purely distributive: there are no potential mutual gains but rather strategic choices determine how some resource is to be shared among agents. Assume also that the outcomes of strategic choices are deterministic, so that there is no issue of mutually beneficial risk-sharing. Then, in general, simple, additive, and multiplicative Kantian equilibria will not exist.[40] This means that Kantian equilibrium is silent about such pure distributive questions.[41] In contrast, if

[39] Here, I am describing Kantian equilibrium as a normative ideal, rather than as a description of how people behave. As Roemer points out, people are parochial and have a tendency to help their neighbors or those similarly situated rather than people in general (see, for example, 20). It may be an advantage of Kantian equilibrium as a descriptive theory if it were to exclude those who cannot aid in cooperation, but unless some moral justification is posited for this feature, it is not satisfactory as a *complete all-things-considered* normative prescription in games.

[40] See the discussion in the last paragraph of section II.III, and Proposition 7 in the Appendix, which applies to n-person distributive games.

[41] In section 2.4 of *How We Cooperate*, Roemer deals with the dictator game, which is purely distributive, and the ultimatum game, which is not quite purely distributive in

we consider cooperative utilitarianism along the lines suggested by Regan (1980), which enjoins each player to choose their part in a strategy profile $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ that maximizes the utilitarian sum $\sum_i V^i(\mathbf{s})$,[42] then we can deal adequately with both affected bystanders and distributive questions (assuming diminishing marginal utility in the resource to be distributed).[43] The point here is not to argue for cooperative utilitarianism per se, but rather to emphasize that Kantian equilibrium may give good moral prescriptions for certain kinds of cooperative problems, but it needs to be integrated with other moral principles to deal with more general problems such as those involving harm to bystanders and distributive questions. This would again be aided by a clearer account of the moral foundations of Kantian equilibrium, which might then apply to a more general class of cases.

the sense that I have in mind, because it also allows for the possibility that the resource will disappear if an agreement is not reached—so there is some possibility for mutual loss, which the players need to mutually avoid. Roemer invokes the device of a veil of ignorance to render these games symmetric and then applies Kantian equilibrium to the point before Nature selects the player roles. However, this treatment appears ad hoc. Why apply it only to the ultimatum and dictator games? We could apply this device to any asymmetric game, rendering it symmetric. But if we were to say that, in general, we should apply this transformation to all games, and *only then* apply Kantian equilibrium to the transformed game, this would amount to a different solution concept and it would in general require interpersonal comparisons of cardinal utility. In fact, assuming that all players make interpersonal comparisons in the same way, using an argument similar to Harsanyi's (1953) impartial observer theorem, simple Kantian equilibrium from behind the veil of ignorance would amount to choosing the strategy s^* that maximizes the utilitarian sum $\sum_i V^i(s^*, \ldots, s^*)$. This is similar to the solution concept of cooperative utilitarianism described in the text.

Note finally that the reason that the existence of Kantian equilibrium in the dictator game modeled from behind the veil of ignorance is not in conflict with Proposition 2 about the non-existence of Kantian equilibria in zero-sum games is that the version of the game that incorporates risk attitudes from behind the veil of ignorance is no longer zero-sum. Effectively, strategy profiles induce lotteries over outcomes and agents have a common interest to reduce their joint risks: from behind the veil of ignorance, both players prefer the lottery induced by the strategy of giving half to the other when you are the dictator to the strategy of keeping all for yourself.

[42] This assumes cardinal interpersonally comparable utility.

[43] Regan's cooperative utilitarianism is actually more complex—I am simplifying here. It enjoins one to anticipate who will and who won't cooperate, and to choose the best cooperative scheme among cooperators, treating non-cooperators non-cooperatively. But, crucially, this just means that one ought to be clear-eyed about who will cooperate, *not* that one only cares about cooperators. The objective that is maximized by cooperators is still $\sum_i V^i(\mathbf{s})$, including *everyone*, both cooperators and non-cooperators. So, this more sophisticated version also deals well with distributive problems and bystanders alike. The behavior chosen by the cooperators is viewed as the behavior that a moral person ought to choose.

V.II. Group Agency

This section discusses the relation of Kantian equilibrium to *collective intentions* and *team agency* (Collingwood [1942] 1947; Sellars 1968; Tuomela and Miller 1988; Gilbert 1989; Hurley 1989; Searle 1990; Bratman 1992; Bacharach 2006; List and Pettit 2011).[44] This perspective encompasses the scheme presented in section V.I, but it may place less emphasis on morality. Gold and Sugden characterize these notions as follows: "Collective intentions are those intentions associated with joint actions" (2007, 109). They also say:

> A starting point for such an analysis can be found in a body of decision-theoretic literature on *team agency*. This seeks to extend standard game theory, where each individual asks separately 'What should *I* do?' to allow teams of individuals to count as agents and for players to ask the question 'What should *we* do?' This leads to *team reasoning*, a distinctive mode of reasoning that is used by members of teams, and which may result in cooperative actions. (Gold and Sugden 2007, 110)

As in the scheme presented in section V.I, each agent is enjoined to do *their part* in the arrangement that best furthers the aims of the group. Gold and Sugden present the following scheme for "Simple Team Reasoning (from a group viewpoint)":

> (1) We are the members of S.
>
> (2) Each of us identifies with S.
>
> (3) Each of us wants the value of U to be maximized.
>
> (4) A uniquely maximizes U.
>
> ---
>
> Each of us should choose her component of A. (Gold and Sugden 2007, 125)

Here S is a group, U is some objective adopted by the group, and A is some action profile. Gold and Sugden (2007) also formulate this schema from the point of view of an individual member as opposed to the group as a whole.

One striking difference between the above scheme and Roemer's discussion of Kantian equilibrium is that whereas Roemer writes as though the action choice is *joint* but the objectives V^i remain *individual*, in the

[44] Roemer discusses this literature on 19–21.

above scheme there is a *group objective U*. The above scheme for team reasoning supposes that individuals adopt a collective objective; each individual is not just concerned with their own narrow goals, but rather adopts a collective perspective. On this conception, one might argue that it is the adoption of the collective goal that keeps individuals from deviating to their narrowly self-interested best response.

The team-reasoning perspective may fall short of the more thoroughgoing moral perspective that I advocated in section V.I but it still must go beyond narrow self-interest: the individual must internalize the interests of the group. It is true that people often form an attitude of solidarity only with a specific group with whom they identify or cooperate rather than accepting and internalizing a more universal morality. Notions such as fairness and consideration of others still apply within this more limited scope of concern. Even with this narrower focus, in cooperating, people would still tend to consider it to be unfair to not do their part and so let down their fellow cooperators, and they would still tend to show concern for the members of their own group.

There is also a connection to the problem of trust raised in section IV: why should you do your part only if you expect others to do theirs? Several authors have written about the ideal of cooperating with those who are willing to cooperate. For example, Regan's cooperative utilitarianism says that "what each agent ought to do is to *co-operate with whoever else is co-operating, in the production of the best consequences possible given the behavior of non-cooperators*" (1980, 124). This can be thought of as a kind of hybrid of Kantian and Nash reasoning, where the group of cooperators is determined by willingness to cooperate. Gold and Sugden (2007) also consider variants of the above scheme that involve cooperation only with those who are willing to cooperate and discuss the importance of assurance that others will cooperate.

The need for trust that others will cooperate may be important for several reasons. First, knowing who else will cooperate may be critical to knowing which of your actions will best contribute to the collective goal. Second, knowing who is cooperating may (or may not) affect the collective goal, because the collective goal may (or may not) pertain only to the interests of those who cooperate.[45] Third, knowing who is cooperating may inform what it is *fair* for each individual member to do. Trust matters because what best achieves the *collective* goal depends on who is

[45] In Regan's scheme, the collective objective is not altered by the set willing to cooperate because it is always the goal of maximizing aggregate utility. In another scheme, the goal may be to maximize the aggregate utility of cooperators, and hence would depend on the set of cooperators.

cooperating. Perhaps the theory of Kantian equilibrium can be developed along similar lines to specify how and why cooperation is sensitive to the collection of agents willing to cooperate.

V.III. Roemer on Morality

One potential criticism of the argument presented in this paper is that whereas I have been criticizing Roemer for attempting to found cooperation on self-interest and trust, rather than on morality, he actually does argue that agents' reasons for doing their part in Kantian equilibrium are based on morality. If this is so, then some of my criticisms are misplaced.

Roemer discusses morality in many passages. I mentioned some in the introduction. In criticizing Brekke, Kverndokk, and Nyborg (2003) for putting a moral penalty term in the utility function, Roemer writes:

> Why say that players pay a 'cost' for deviating from the Kantian action, rather than just saying that they play the action they think is the right thing to do? Is not the latter simpler, although heretical from the classical viewpoint? (40)

When discussing strikes, Roemer writes:

> The important question is whether it is the fear of punishment or Kantian morality that motivates participation for most strikers. The language of solidarity [...] is ubiquitous in the labor movement [...]. (56)

When criticizing the 'warm glow' approach to collective action (Andreoni 1990), Roemer writes:

> Do participators get a warm glow from participating? Surely this is often the case. But I conjecture that the warm glow is the *consequence* of having 'done the right thing,' not the *cause* of participation. (57)

And in the concluding chapter, Roemer writes about fairness as a motive for cooperation (218).

All of the above passages assert that people must be motivated by moral considerations if they are to rationally cooperate. These claims are in line with the arguments that I have been making in section V.I and elsewhere. Reading these passages in isolation, I find myself in sympathy with Roemer, and I agree that moral principles beyond altruistic concerns for others are at play in cooperation. However, Roemer also appears to believe that these moral considerations can be founded in self-interest, trust, and also considerations of symmetry, and that is where we part company.

Elaborating on his view of morality in general, Roemer writes:

> My own feeling is that concepts of fairness (and hence morality) have very much to do with symmetry. Our brains have evolved to focus on symmetry, to search for symmetry in situations, and it is not a stretch to believe that our concepts of fairness, likewise, depend upon symmetry. (70)

Explaining the morality of cooperation in symmetric situations, he writes: "What I propose is that the general rule that always finds the cooperative solution in symmetric games is 'Choose the strategy I would like all to choose.' This *defines* the 'right thing to do'" (22).

I would take issue with both of these claims. While symmetry is an essential constraint on moral systems, it is not sufficient in itself to determine a moral system or to determine the content of fairness because it is too weak a principle for that purpose: for example, a system that pursued bad outcomes equally for everyone could be symmetric. Many systems treat people symmetrically, and we would not regard them all as moral. There must be more to morality, fairness, and cooperation than just symmetry, although symmetry is an important ingredient. With regard to the second statement, Roemer claims that the Kantian rule defines the right thing to do. Perhaps Roemer's Kantian principle defines the moral action in the sense that the two are coextensive: an action is moral if and only if it is what is prescribed by Kantian equilibrium (but see section V for problems with this idea). But Kantian optimization does not define the right thing to do in the sense that morality is *by definition* what Kantian optimization prescribes. There must be a more fundamental moral reason why the prescriptions of Kantian optimization are the right thing to do, and these more fundamental reasons are what an agent must appeal to if she is to rationally choose as Kantian optimization prescribes.

The core of my objection can be explained with regard to the following passage:

> This approach to moral thinking has several advantages: first, it does not require that the optimizer know the preferences of others, and second, it does not require her to care about others. (Indeed, the same trick to engender moral behavior is embedded in 'Do unto others as you would have them do unto you.') We often invoke the same mech-

> anism in teaching our children not to litter: we ask the child how *he* would feel if *others* were to litter the way he is doing, rather than relying on his altruism to desist from throwing his candy wrapper on the sidewalk. Our practice with littering children suggests to me that appealing to the categorical imperative is more persuasive than appealing to altruism. (70)

Let us put aside Kant's categorical imperative, since Roemer admits that he does not claim a deep connection to Kant's philosophy. Let us instead consider the Golden Rule: 'do unto others as you would have others do unto you'. The Golden Rule asks an individual to draw on their internal understanding of what is good for them in determining what is the right thing to do but it is emphatically *not* a self-interested principle. A purely self-interested person would not obey the Golden Rule because it would often not be in their interest to do so. The problem with Roemer's argument, as I understand it, is the view that the morality of cooperation can be founded on self-interest, symmetry, and trust. I think that this is not the case. We must appeal to other moral notions, not reducible to these, to do so. Perhaps there is some rich notion of fairness that can ground the morality of cooperation. But, in that case, an individual must recognize that it is important to behave fairly, *not just* that it is in her interest to do so. Separate questions are whether fairness is enough, so that altruism becomes unnecessary, and whether fairness itself implicates concern for others, or whether there can be a notion of fairness completely divorced from such altruistic concern. These are difficult questions. The key point that I would like to make is that the morality of cooperation cannot be founded on self-interest alone.

VI. Conclusion

In this paper, I have raised several objections to Kantian equilibrium. However, the purpose of these objections is not to undermine Kantian equilibrium but rather to explore its foundations. I think that questions such as 'what is it that I would like everyone to do?' and 'what is it that I think everyone should do?' are basic to both cooperation and morality. Kantian equilibrium attempts to formalize the answers to these questions in the context of games. I have been discussing what I view as some technical challenges to the formal implementation of these questions and their answers in *How We Cooperate*, and also a different way of thinking of the theory's foundation. I think the project initiated by the book is important and that the book persuasively makes the case that an approach with a Kantian flavor can be fruitfully incorporated into economic theory. The array of applications it presents is impressive. I look forward to seeing the further development of this project as it is both promising and important.

VII. APPENDIX

Proof of Proposition 1

Suppose that $\mathbf{s}^*$ is a simple Kantian equilibrium. It is immediate that (2) holds. Let $s^* = s_1^* = \cdots = s_n^*$. Then the definition of simple Kantian equilibrium implies that, for all $r \in \mathbb{R}, V^i(s^*, \ldots, s^*) \geq V^i(\varphi(s^*, r), \ldots, \varphi(s^*, r))$. This implies that $(s_1^*, \ldots, s_n^*) = (s^*, \ldots, s^*)$ is a φ-Kantian equilibrium. Going in the other direction, assume conditions (1) and (2), and let $s^* = s_1^* = \cdots = s_n^*$. Choose $s \in S$. By the assumption on the range of $\varphi(s, \cdot)$, there exists an r such that $\varphi(s^*, r) = s$. Since $(s^*, \ldots, s^*)$ is a φ-Kantian equilibrium, $V^i(s^*, \ldots, s^*) \geq V^i(\varphi(s^*, r), \ldots, \varphi(s^*, r)) = V^i(s, \ldots, s)$. So $(s_1^*, \ldots, s_n^*)$ is a simple Kantian equilibrium. □

Proof of Proposition 2

Part (i): Assume, towards contradiction, that under the assumptions of part (i), s^* is a simple Kantian equilibrium. Then, by our assumptions, there must exist $s \in S$, such that $V^1(s^*, s^*) \neq V^1(s, s)$. Since s^* is a simple Kantian equilibrium, it follows that $V^1(s^*, s^*) > V^1(s, s)$. But then, since the game is zero-sum, $V^2(s^*, s^*) < V^2(s, s)$, contradicting the assumption that s^* is a simple Kantian equilibrium.

Part (ii): Assume, towards contradiction, that (s_1^*, s_2^*) is a φ-Kantian equilibrium. Then, by assumption (9), there exists $r \in \mathbb{R}$ such that $V^1(s_1^*, s_2^*) \neq V^1(\varphi(s_1^*, r), \varphi(s_2^*, r))$. Since (s_1^*, s_2^*) is a φ-Kantian equilibrium, it follows that $V^1(s_1^*, s_2^*) > V^1(\varphi(s_1^*, r), \varphi(s_2^*, r))$. But since the game is zero-sum, it follows that $V^2(s_1^*, s_2^*) < V^2(\varphi(s_1^*, r), \varphi(s_2^*, r))$, so (s_1^*, s_2^*) is not a φ-Kantian equilibrium, a contradiction. □

A generalization of the proposition is as follows.

Proposition 7. *Let* $\left([n], S, \left(V^i\right)_{i \in [n]}\right)$ *be a game satisfying:*

$$\forall \mathbf{s}, \mathbf{s}' \in S^n : \quad \left[\exists i : V^i(\mathbf{s}) > V^i(\mathbf{s}')\right] \Rightarrow \left[\exists j : V^j(\mathbf{s}) < V^j(\mathbf{s}')\right] \tag{15}$$

(i) Suppose that there exist $s, s' \in S$ and $i \in [n]$ such that $V^i(s, \ldots, s) \neq V^i(s', \ldots, s')$. Then a simple Kantian equilibrium does not exist in this game.

(ii) Suppose that:

$$\begin{aligned} &\forall (s_1, \ldots, s_n) \in S^n, \exists i \in [n], \exists r \in \mathbb{R} : \\ &V^i(s_1, \ldots, s_n) \neq V^i(\varphi(s_1, r), \ldots, \varphi(s_n, r)) \end{aligned} \tag{16}$$

Then, a φ-Kantian equilibrium does not exist in this game.

Part (i): Assume, towards contradiction, that under the assumptions of part (i), s^* is a simple Kantian equilibrium. Then, by our assumptions, there must exist $s \in S$ and agent i such that $V^i(s^*, \ldots, s^*) \neq V^i(s, \ldots, s)$. Since s^* is a simple Kantian equilibrium, it follows that $V^i(s^*, \ldots, s^*) > V^1(s, \ldots, s)$. But then, since the game satisfies (15), there exists j such that $V^j(s^*, \ldots, s^*) < V^j(s, \ldots, s)$, contradicting the assumption that s^* is a simple Kantian equilibrium.

Part (ii): Assume, towards contradiction, that $(s_1^*, \ldots, s_n^*)$ is a φ-Kantian equilibrium. Then, by assumption (16), there exists $r \in \mathbb{R}$ and agent i such that $V^i(s_1^*, \ldots, s_n^*) \neq V^i(\varphi(s_1^*, r), \ldots, \varphi(s_n^*, r))$. Since $(s_1^*, \ldots, s_n^*)$ is a φ-Kantian equilibrium, $V^i(s_1^*, \ldots, s_n^*) > V^i(\varphi(s_1^*, r), \ldots, \varphi(s_n^*, r))$. But since the game satisfies (15), it follows that there exists an agent j such that $V^j(s_1^*, \ldots, s_n^*) < V^j(\varphi(s_1^*, r), \ldots, \varphi(s_n^*, r))$, so $(s_1^*, \ldots, s_n^*)$ is not a φ-Kantian equilibrium, a contradiction. □

Proof of Proposition 4

Suppose that $\hat{G}$ is a relabeling of G with relabeling profile $\mathbf{f}$. Suppose that $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ is a Nash equilibrium of G. Consider any agent i and any strategy $\hat{s}_i \in \hat{S}$. Since f_i is a bijection, then there exists $s_i' \in S_i$ such that $f^i\left(s_i'\right) = \hat{s}_i$. Then:

$$\begin{aligned}\hat{V}^i(\mathbf{f}(\mathbf{s}^*)) = V^i(\mathbf{s}^*) &\geq V^i\left(s_i', \mathbf{s}_{-i}^*\right) \\ &\geq \hat{V}^i\left(f^1(s_1^*), \ldots, f^{i-1}\left(s_{i-1}^*\right), \hat{s}_i, f^{i+1}\left(s_{i+1}^*\right), \ldots, f^n(s_n^*)\right)\end{aligned}$$

Where the inequality follows from the fact that $\mathbf{s}^*$ is a Nash equilibrium of G. It follows that $\mathbf{f}(\mathbf{s}^*)$ is a Nash equilibrium of $\hat{G}$. The other direction follows from the fact that if $\hat{G}$ is a relabeling of G with a relabeling profile $\mathbf{f}$, then G is also a relabeling of $\hat{G}$ with the inverse relabeling profile $\mathbf{f}^{-1} = \left(\left[f^i\right]^{-1}\right)_{i \in [n]}$. □

Proof of Proposition 5

Part (i): The statement applied to simple, additive, and multiplicative Kantian equilibrium follows from the examples discussed in the text. In particular, the fact that simple and additive Kantian equilibria are not preserved under the relabeling of strategies follows from considering the transformation of the game with utility functions V^i given by (10) to the game with utility functions $\hat{V}^i$ given by (11). See in particular footnote 21 for the details with regard to additive Kantian equilibrium. That multiplicative Kantian equilibrium is not preserved under relabelings follows from the example discussed in footnote 23.

Part (ii): Suppose that $\mathbf{s}^* = (s^*, \ldots, s^*)$ is a simple Kantian equilibrium with $s^* > 0$ and that the positive linear relabeling is such that $f^i \neq f^j$. Then $f^i(s^*) \neq f^j(s^*)$. So $\mathbf{f}(\mathbf{s}^*)$ is not a simple Kantian equilibrium.

Part (iii): Let $\hat{G}$ be a relabeling of G with positive linear relabeling profile $\mathbf{f}$. Let $f^i(s_i) = \alpha^i s_i$. Suppose that $\mathbf{s}^* = (s_1^*, \ldots, s_n^*)$ is a multiplicative Kantian equilibrium of G. Then, observe that, for any $r \geq 0$:

$$\begin{aligned}\hat{V}^i(\mathbf{f}(\mathbf{s}^*)) = V^i(\mathbf{s}^*) \geq V^i(rs_1^*, \ldots, rs_n^*) &= \hat{V}^i\left(\alpha^1 rs_1^*, \ldots, \alpha^n rs_n^*\right) \\ &= \hat{V}^i\left(rf^1(s_1^*), \ldots, rf^n(s_n^*)\right)\end{aligned}$$

Where the inequality follows from the fact that $\mathbf{s}^*$ is a multiplicative Kantian equilibrium of G. It follows that $\mathbf{f}(\mathbf{s}^*)$ is a multiplicative Kantian equilibrium of $\hat{G}$. To go in the opposite direction, note that V^i is also derivable via a positive linear relabeling from $\hat{V}^i$. □

Proof of Proposition 6

Part (i) is immediate.

First, I prove part (ii) for multiplicative Kantian equilibrium. Start with the game $\hat{G} = \left([2], \mathbb{R}_{++}, \left(\hat{V}^1, \hat{V}^2\right)\right)$ with utility functions given by (11), and uniform relabeling $\tilde{G} = \left([2], \mathbb{R}_{++}, \left(\tilde{V}^1, \tilde{V}^2\right)\right)$ induced by the relabeling profile $\mathbf{f} = (f^1, f^2)$ with $f^1 = f^2 = \tilde{f}$, where $\tilde{f}$ is a strictly increasing differentiable function from $\mathbb{R}_{++}$ to $\mathbb{R}_{++}$ such that:

$$\frac{1}{2} \cdot \frac{\tilde{f}\left(\frac{2}{3}\right)}{\tilde{f}'\left(\frac{2}{3}\right)} \neq \frac{1}{4} \cdot \frac{\tilde{f}\left(\frac{4}{3}\right)}{\tilde{f}'\left(\frac{4}{3}\right)} \tag{17}$$

Observe that:

$$\tilde{V}^i(s_1, s_2) = \ln\left(\tilde{f}^{-1}(s_1)\right) + \ln\left(\tilde{f}^{-1}(s_2)\right) - \tilde{f}^{-1}(s_1) - \tilde{f}^{-1}(s_2)$$

We have established in the text that $(2/3, 4/3)$ is a multiplicative Kantian equilibrium of $\hat{G}$. I now show that $\left(\tilde{f}(2/3), \tilde{f}(4/3)\right)$ is not a multiplicative Kantian of $\tilde{G}$. In particular, observe that:

$$\begin{aligned}
&\left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=1} \tilde{V}^i\left(r\tilde{f}\left(\frac{2}{3}\right), r\tilde{f}\left(\frac{4}{3}\right)\right) \\
= &\left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=1} \left[\ln\left(\tilde{f}^{-1}\left(r\tilde{f}\left(\frac{2}{3}\right)\right)\right) + \ln\left(\tilde{f}^{-1}\left(r\tilde{f}\left(\frac{4}{3}\right)\right)\right) - \right. \\
&\left. -\tilde{f}^{-1}\left(r\tilde{f}\left(\frac{2}{3}\right)\right) - \tilde{f}^{-1}\left(r\tilde{f}\left(\frac{4}{3}\right)\right)\right] \\
= &\frac{3}{2}\left(f^{-1}\right)'\left(\tilde{f}\left(\frac{2}{3}\right)\right)\tilde{f}\left(\frac{2}{3}\right) + \frac{3}{4}\left(f^{-1}\right)'\left(\tilde{f}\left(\frac{4}{3}\right)\right)\tilde{f}\left(\frac{4}{3}\right) - \\
&-\left(f^{-1}\right)'\left(\tilde{f}\left(\frac{2}{3}\right)\right)\tilde{f}\left(\frac{2}{3}\right) - \left(f^{-1}\right)'\left(\tilde{f}\left(\frac{4}{3}\right)\right)\tilde{f}\left(\frac{4}{3}\right) \\
= &\frac{3}{2} \cdot \frac{1}{\tilde{f}'\left(\frac{2}{3}\right)}\tilde{f}\left(\frac{2}{3}\right) + \frac{3}{4} \cdot \frac{1}{\tilde{f}'\left(\frac{4}{3}\right)}\tilde{f}\left(\frac{4}{3}\right) - \frac{1}{\tilde{f}'\left(\frac{2}{3}\right)}\tilde{f}\left(\frac{2}{3}\right) - \frac{1}{\tilde{f}'\left(\frac{4}{3}\right)}\tilde{f}\left(\frac{4}{3}\right) \\
= &\frac{1}{2} \cdot \frac{\tilde{f}\left(\frac{2}{3}\right)}{\tilde{f}'\left(\frac{2}{3}\right)} - \frac{1}{4} \cdot \frac{\tilde{f}\left(\frac{4}{3}\right)}{\tilde{f}'\left(\frac{4}{3}\right)} \neq 0
\end{aligned}$$

Where the last non-equality follows from (17). It follows that $\left(\tilde{f}\left(2/3\right), \tilde{f}\left(4/3\right)\right)$ is not a multiplicative Kantian equilibrium of $\tilde{G}$. This completes the proof of part (ii) for multiplicative Kantian equilibrium.

I now establish part (ii) for additive Kantian equilibrium. I consider the same games $\hat{G}$ and $\tilde{G}$ as above except I replace condition (17) by:

$$\tilde{f}'\left(\frac{1}{\sqrt{2}}\right) \neq \tilde{f}'\left(1+\frac{1}{\sqrt{2}}\right) \tag{18}$$

Next, observe that $(s_1^*, s_2^*) = (1/\sqrt{2}, 1 + 1/\sqrt{2})$ is an additive Kantian equilibrium of $\hat{G}$. To see this observe that $\hat{V}^i\left(s_1^* + r, s_2^* + r\right)$ is strictly concave in r and:

$$\begin{aligned}
&\left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=0} \hat{V}^i\left(\frac{1}{\sqrt{2}}+r, 1+\frac{1}{\sqrt{2}}+r\right)\\
=&\left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=0}\left[\ln\left(\frac{1}{\sqrt{2}}+r\right)+\ln\left(1+\frac{1}{\sqrt{2}}+r\right)-\left[\frac{1}{\sqrt{2}}+r\right]-\left[1+\frac{1}{\sqrt{2}}+r\right]\right]\\
=&\sqrt{2}+\frac{\sqrt{2}}{1+\sqrt{2}}-2=\frac{\left(\sqrt{2}+2\right)+\sqrt{2}-\left(2+2\sqrt{2}\right)}{1+\sqrt{2}}=0
\end{aligned}$$

Next, observe that:

$$\begin{aligned}
&\left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=0} \tilde{V}^i\left(\tilde{f}\left(\frac{1}{\sqrt{2}}\right)+r, \tilde{f}\left(1+\frac{1}{\sqrt{2}}\right)+r\right)\\
=&\left.\frac{\mathrm{d}}{\mathrm{d}r}\right|_{r=0}\left[\ln\left(\tilde{f}^{-1}\left(\tilde{f}\left(\frac{1}{\sqrt{2}}\right)+r\right)\right)+\ln\left(\tilde{f}^{-1}\left(\tilde{f}\left(1+\frac{1}{\sqrt{2}}\right)+r\right)\right)-\right.\\
&\left.-\tilde{f}^{-1}\left(\tilde{f}\left(\frac{1}{\sqrt{2}}\right)+r\right)-\tilde{f}^{-1}\left(\tilde{f}\left(1+\frac{1}{\sqrt{2}}\right)+r\right)\right]\\
=&\sqrt{2}\left(f^{-1}\right)'\left(\tilde{f}\left(\frac{1}{\sqrt{2}}\right)\right)+\frac{\sqrt{2}}{1+\sqrt{2}}\left(f^{-1}\right)'\left(\tilde{f}\left(1+\frac{1}{\sqrt{2}}\right)\right)-\\
&-\left(f^{-1}\right)'\left(\tilde{f}\left(\frac{1}{\sqrt{2}}\right)\right)-\left(f^{-1}\right)'\left(\tilde{f}\left(1+\frac{1}{\sqrt{2}}\right)\right)\\
=&\left(\sqrt{2}-1\right)\left(f^{-1}\right)'\left(\tilde{f}\left(\frac{1}{\sqrt{2}}\right)\right)-\frac{1}{1+\sqrt{2}}\left(f^{-1}\right)'\left(\tilde{f}\left(1+\frac{1}{\sqrt{2}}\right)\right)\neq 0.\\
=&\left(\sqrt{2}-1\right)\frac{1}{\tilde{f}'\left(\frac{1}{\sqrt{2}}\right)}-\frac{1}{1+\sqrt{2}}\cdot\frac{1}{\tilde{f}'\left(1+\frac{1}{\sqrt{2}}\right)}\\
=&\left(\sqrt{2}-1\right)\left[\frac{1}{\tilde{f}'\left(\frac{1}{\sqrt{2}}\right)}-\frac{1}{\tilde{f}'\left(1+\frac{1}{\sqrt{2}}\right)}\right]\neq 0
\end{aligned}$$

Where the last non-equality follows from (18). So $\left(\tilde{f}\left(1/\sqrt{2}\right), \tilde{f}\left(1+1/\sqrt{2}\right)\right)$ is not an additive Kantian equilibrium of $\tilde{G}$. This establishes part (ii) for additive Kantian equilibrium. □

References

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *The Economic Journal* 100 (401): 464–477.

Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory.* Edited by Natalie Gold and Robert Sugden. Princeton, NJ: Princeton University Press.

Bernheim, B. Douglas. 1984. "Rationalizable Strategic Behavior." *Econometrica* 52 (4): 1007–1028.

Bratman, Michael E. 1992. "Shared Cooperative Activity." *The Philosophical Review* 101 (2): 327–341.

Brekke, Kjell Arne, Snorre Kverndokk, and Karine Nyborg. 2003. "An Economic Model of Moral Motivation." *Journal of Public Economics* 87 (9–10): 1967–1983.

Collingwood, Robin George. (1942) 1947. *The New Leviathan; or, Man, Society, Civilization and Barbarism.* Reprinted with corrections. Oxford: Clarendon Press.

Debreu, Gerard. 1952. "A Social Equilibrium Existence Theorem." *PNAS* 38 (10): 886–893.

Dekel, Eddie, and Faruk Gul. 1997. "Rationality and Knowledge in Game Theory." In *Advances in Economics and Econometrics: Theory and Applications. Seventh World Congres: Volume I*, edited by David M. Kreps and Kenneth F. Wallis, 87–172. Cambridge: Cambridge University Press.

Elster, Jon. 2017. "On Seeing and Being Seen." *Social Choice and Welfare* 49 (3–4): 721–734.

Fan, Ky. 1952. "Fixed-Point and Minimax Theorems in Locally Convex Topological Linear Spaces." *PNAS* 38 (2): 121–126.

Gabarró, Joaquim, Alina García, and Maria Serna. 2011. "The Complexity of Game Isomorphism." *Theoretical Computer Science* 412 (48): 6675–6695.

Gilbert, Margaret. 1989. *On Social Facts.* London: Routledge.

Glicksberg, Irving L. 1952. "A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points." *Proceedings of the American Mathematical Society* 3 (1): 170–174.

Gold, Natalie, and Robert Sugden. 2007. "Collective Intentions and Team Agency." *The Journal of Philosophy* 104 (3): 109–137.

Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61 (5): 434–435.

Hurley, Susan L. 1989. *Natural Reasons: Personality and Polity.* New York, NY: Oxford University Press.

Kraut, Richard. 2020. "Altruism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Article published August 25, 2016; last modified August 31, 2020. https://plato.stanford.edu/archives/spr2020/entries/altruism/.

List, Christian, and Philip Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford: Oxford University Press.

Pearce, David G. 1984. "Rationalizable Strategic Behavior and the Problem of Perfection." *Econometrica* 52 (4): 1029–1050.

Regan, Donald H. 1980. *Utilitarianism and Co-operation.* New York, NY: Oxford University Press.

Roemer, John E. 2019. *How We Cooperate: A Theory of Kantian Optimization.* New Haven, CT: Yale University Press.

Rousseau, Jean-Jacques. (1755) 1923. "Discourse on the Origin and Basis of Inequality among Men." In *The Social Contract and Discourses by Jean-Jacques Rousseau*, translated by George D. H. Cole, 155–246. London: J. M. Dent & Sons.

Searle, John R. 1990. "Collective Intentions and Actions." In *Intentions in Communication*, edited by Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, 401–415. Cambridge, MA: The MIT Press.

Sellars, Wilfrid. 1968. *Science and Metaphysics: Variations on Kantian Themes*. New York, NY: Humanities Press.

Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs* 6 (4): 317–344.

Tuomela, Raimo, and Kaarlo Miller. 1988. "We-Intentions." *Philosophical Studies* 53 (3): 367–389.

Itai Sher is an Associate Professor of Economics at the University of Massachusetts Amherst. His research is at the boundary of ethics and economics and focuses on topics such as freedom of choice, voting institutions, and value pluralism in normative economics. He is a Co-Editor at the journal *Economics & Philosophy* and of the Oxford University Press Philosophy, Politics and Economics book series. He is a founder and co-organizer of the interdisciplinary conference series *Normative Ethics and Welfare Economics*.
Contact e-mail: <itaisher@gmail.com>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 85–97.
https://doi.org/10.23941/ejpe.v13i2.502

Roemer on the Rationality of Cooperation

PETER VALLENTYNE
University of Missouri

In *How We Cooperate: A Theory of Kantian Optimization*, John Roemer (2019), a philosophically informed and highly influential normative economist, builds upon some previous work and defends a Kantian, and hence unorthodox, theory of the rationality of cooperation. More exactly, he addresses the question of when it is rational to cooperate with other agents in *non-cooperative games*. These are games with no possibility of first negotiating an agreement that is externally binding (for example, with externally imposed penalties that are large enough to make it irrational to fail to comply with the agreement).

I shall first summarize the philosophical core of Roemer's project. Then I shall raise some concerns about it.

I. AN OVERVIEW OF KANTIAN OPTIMIZATION

Throughout, like Roemer (I believe), I focus on game theory as a normative theory of rational choice in the context of interacting agents, rather than as a descriptive (predictive) theory of such choice.

I follow Roemer and make the standard assumption of *idealized* game theory that the following are common knowledge (that is, they are true, each player knows it, each player knows that each player knows it, etc.): (1) that each agent is perfectly rational, (2) what choices each player confronts, and (3) what the payoffs are for each player given the choices for every player. These conditions do not hold in real life: agents are not perfectly rational. They suffer from deductive incompleteness, inconsistency, confusion, and weakness of the will. Moreover, what one agent knows is typically not known by all agents, let alone common knowledge to all. *Realistic* game theory is based on more realistic assumptions. The problem of rational choice in non-cooperative games is difficult, even in the idealized case, and so focusing on it makes sense.

AUTHOR'S NOTE: For helpful comments, I am grateful to Akshath Jitendranath, Luc Lauwers, John Roemer, Bertil Tungodden, Alex Voorhoeve, and Paul Weirich.

		Column	
		Silent	Confess
Row	Silent	(8, 8)	(1, 9)
	Confess	(9, 1)	(3, 3)

Table 1: The Prisoner's Dilemma.

Following Roemer, we shall assume that the outcome payoffs in a game represent, for each agent, the *prudential value* (what makes her life go well for her), on some arbitrary scale, of the outcome for that agent.

The standard view of non-cooperative games is that rationality requires, when possible, agents to *Nash optimize,* which means making a choice (adopting a strategy) that is part of a *Nash equilibrium*—a choice *n*-tuple (one choice for each player) such that each player's choice is a best response to the choices of the other players.

To illustrate this, consider, for example, the Prisoner's Dilemma game in Table 1, where (1, 9) designates the payoffs of 1 to the row player (*Row*) and of 9 to the column player (*Column*), respectively.

Here, suppose the police caught two criminals and placed them in isolated cells. Each has two options: to remain silent or to confess. If one confesses and the other does not, then the confessor will go to prison for one year and have a life worth 9 units of value. In this case, the silent agent will be convicted and sent to prison for nine years, and have a life worth 1 unit. If they both remain silent, they will each go to prison for 2 years and have lives worth 8 units of value each. If they both confess, they will each go to prison for 7 years and have lives worth 3 units of value each.

In this Prisoner's Dilemma, the only Nash equilibrium is for each to confess. (*Confess*, *Confess*) is a Nash equilibrium because neither player (on their own) can unilaterally do better by remaining silent. (*Silent*, *Silent*) and (*Silent*, *Confess*) are not Nash equilibria since *Row* would do better by switching to confessing. (*Confess*, *Silent*) (as well as (*Silent*, *Silent*)) is not a Nash equilibrium, since *Column* would do better by switching to confessing. Nash optimizing here requires performing one's choice in the only Nash equilibrium.

Above we focused on a single play of the Prisoner's Dilemma. If the game will be played infinitely many times, or with no known upper bound on the number of times, then rational agents will take into account that failing to cooperate in one round (by confessing) may have the result that, in the future, other agents won't trust them to cooperate and will thus exclude them from the possibility of cooperation. In such cases, it may be Nash optimal to cooperate, since the long-term benefits of cooperating may exceed the one-round benefits of defecting. If, however, the number of rounds is finite and also common knowledge to the players, then Nash optimization will require defecting on the first round. This is because all know that Nash optimizers will not cooperate on the final round, and so all know that Nash optimizers will not cooperate on the previous round (since there are no last-round benefits), and, by backward induction, all know that Nash optimizers will not cooperate in the first round. So, although indefinite repeated play can open up the possibility of rational cooperation among Nash optimizers, it does not do so when there is a finite upper bound on the number of plays and this is common knowledge. In what follows, we (like Roemer) shall therefore focus on single-play games.

Roemer denies that, in single-play games, rationality precludes cooperation. In the above Prisoner's Dilemma, if each player chooses cooperatively and remains silent, each would have a life worth 8 units, rather than 3 units. How can it be rational for each to confess (as required by Nash optimization) when this leads to an outcome that is worse for each of them compared to each remaining silent?

Roemer's core claim, which he strengthens in various ways, is that, in simple symmetric games, with players one trusts sufficiently to cooperate, rationality requires players to choose their part of a simple Kantian equilibrium when it produces a Pareto optimal outcome.[1] Very crudely, this requires that each agent make a choice that, *if all agents chose in a like manner* (a kind of Kantian universalization), it would have the best consequences for the agent. I will explain this in terms of two-person games.

Symmetric games are games in which players are 'identically situated' in the sense that (1) each player has the same possible choices (where choice a_i for player 1 is the 'same choice' as choice b_i for player 2), and (2) the ordinal rank, relative to the feasible outcome, for agent 1 for choice-pair (a_i, b_k) is the same as the ordinal rank to agent 2 for the choice-pair (a_k, b_i) (for example, each agent gets their second best feasible outcome).

[1] Initially, I thought that Roemer claimed that rationality requires Kantian optimization (with those one trusts sufficiently to cooperate), whenever the game has a common diagonal (that is, is symmetric when players 'make the same choice'), even if so optimizing is not Pareto optimal. In e-mail correspondence, however, Roemer clarified that his claim is the weaker one made in the text.

The Prisoner's Dilemma displayed above, with staying silent as each agent's first choice, and confessing as each agent's second, is symmetric. For example, (*Confess*, *Silent*) gives payoffs to the two players of (9, 1), and (*Silent*, *Confess*) gives them a payoff of (1, 9).

In a symmetric game, a *simple Kantian equilibrium* is a choice-pair (a_i, b_i) for which no player gets a greater benefit from any alternative pair (a_j, b_j). This considers only choice-pairs in which everyone 'does the same thing', understood as playing the *n*-th choice, for some *n*. A simple Kantian equilibrium is a choice-pair in which everyone 'does the same thing' and doing so is at least as good, for each agent, as any other choice-pair in which everyone 'does the same thing'. In the above Prisoner's Dilemma, each player confessing is a choice-pair in which everyone 'does the same thing', as is each player remaining silent. Each player gets a greater benefit from the latter, and hence only that outcome (that is, each remaining silent) is a simple Kantian equilibrium. Kantian optimization thus requires that each remain silent, which gives each 8 units of value, rather than the 3 units that each would receive, if each confessed (as required by Nash optimization).

The definition of symmetric games I gave above assumes that player 1's *n*-th choice can be identified in a non-arbitrary way with player 2's *n*-th choice. This is so, because symmetry requires that, when both players make their *n*-th choice (for any *n*), the outcome has the same ordinal rank for each player relative to the feasible outcomes. In a later section, I will question whether it is plausible to claim that the rationality of a choice depends on such, as I will call it, 'interpersonal identification' of choices. In the present section, however, I will make Roemer's assumption that the specification of a game includes a privileged enumeration of choices for each agent (such that the *n*-th choice of one agent is the same choice as the *n*-th choice of the other agent).

A Nash optimizer (in the two-person case) asks themself "Given the strategy chosen by my opponent, what is the best strategy for me?", whereas the Kantian optimizer asks "[If I trust my opponent] [w]hat is the strategy I would like both of us to play?" (12). That is, the Kantian optimizer, unlike the Nash optimizer, does not treat the choices of the others as given. Roemer makes clear (9, 20) that Kantian optimization requires cooperation only when the agent sufficiently trusts the other(s) to cooperate. Such optimizers are willing to cooperate with those they believe to be cooperators, but not with those they believe to be non-cooperators.

Kantian optimization is not based on an altruistic concern for others (although it is compatible with that). Nor is it based on team reasoning about some collective goal that each player seeks to promote. Instead, it is based on reasoning cooperatively with those disposed to reason cooperatively (39) about how to promote *shared interests* (distinct interests, but ones that can sometimes be jointly promoted).

Roemer establishes that simple Kantian equilibria always exist for symmetric games, and, more generally, for games with a "common diagonal" (23)—a left-to-right downward diagonal (the 'main diagonal') in the respective strategy matrix (that is, choice *n*-tuples where agents make the same choices) such that the ordering of the diagonal entries is the same for all agents. Moreover, if the payoffs for the game are strictly monotone (that is, strictly increasing, or strictly decreasing, in the strategies of others), then simple Kantian equilibria are Pareto optimal (such that no alternative feasible choice *n*-tuple makes some agent better off and no agent worse off).

Where individuals have different preferences over individual payoffs/outcomes, there may be no common diagonal, and simple Kantian equilibria may not exist. Roemer, however, extends Kantian optimization to require, where the payoffs are strictly monotone, either a multiplicative Kantian equilibrium, an additive Kantian equilibrium, or a mixture of the two. (Simple Kantian equilibria are a special case of each.) My critical comments below won't address these; so I won't explain these notions here.

Kantian optimization among people who trust each other to cooperate solves two major problems that confront Nash optimization: (1) the inefficiency of Nash equilibria in the presence of negative externalities (for example, the tragedy of the commons problem), and (2) the inefficiency of Nash equilibria in the presence of positive externalities (for example, the free-rider problem for public goods). In strictly monotonic, symmetric games, Kantian optimization will select the level of negative (or positive) externalities that, if imposed on each player, will leave each as well off as any other universalized level.

In what follows, I will focus on the philosophical foundations of simple Kantian optimization. I shall thus not address the more sophisticated forms of Kantian optimization, nor the many interesting theorems and applications. For simplicity, I focus on two-person games without the possibility of randomization among pure options.

II. PRACTICAL RATIONALITY (AND ALTRUISM, FAIRNESS, AND MORALITY)

Roemer defends his theory primarily as a normative theory of rational choice (69, 215), although he thinks it has some descriptive/predictive value as well. He sometimes describes his theory as a moral theory (69–70), but it seems clear that he means it to be a theory of practical rationality, where this is the theory of rational choice relative to the values that the agent has (which need not be moral values).

Practical rationality is often thought of as rational choice relative to the agent's *prudential values* (their well-being, narrowly understood), but there is no reason to so limit it. Agents clearly care about their own well-being, but most agents are also partially *altruistic* (care about the well-being of at least some others), care about *cooperative fairness* (for example, the extent to which outcomes approximate the outcome of the negotiation of an externally enforceable agreement), or care about conforming to *moral values* (for example, choosing in morally permissible ways or in morally better ways). To the extent that agents care about such matters (and this may not be rationally required), practical rationality is sensitive to how well a choice promotes them (in addition to the agent's well-being).

Roemer is well aware of this, but, in most of his book, Roemer assumes that agents care only about their own well-being. This is because he believes that *cooperative reasoning*, rather than altruism (or, presumably, valuing cooperative fairness or conformance to moral values), is key to the rationality of cooperation (4). I'm skeptical about this, but I fully agree that any realistic appeal to such valuings will not eliminate the possibility and desirability of cooperative reasoning (since there will still be conflicts between the overall values of different agents).

In most of his book, Roemer assumes that agents care only about their own well-being, but, in chapter 5, Roemer considers the implications of adding some altruism to the values of agents. He there establishes that Kantian equilibria for economies with partially altruistic preferences are observationally equivalent to those in the same economy without altruism.

I agree with Roemer that appealing to altruism, concern for cooperative fairness, or concern for moral values does not eliminate the possibility or desirability of cooperative reasoning.

		Column	
		Choice 1 (Confess)	Choice 2 (Silent)
Row	Choice 1 (Silent)	(1, 9)	(8, 8)
	Choice 2 (Confess)	(3, 3)	(9, 1)

Table 2: The Prisoner's Dilemma Relabeled.

III. Some Criticism

I have two criticisms of Kantian optimization. One concerns its presupposition that there is some privileged manner of making *interpersonal identifications of choices* of different agents (different agents performing the same action). The other concerns its general requirement to cooperate with those one trusts sufficiently to cooperate with one.

III.I. The Irrelevance of Interpersonal Identifications of Choices

Consider the reformulation of the Prisoner's Dilemma in Table 2. This is the same as the Prisoner's Dilemma presented earlier (in Table 1), except that (1) the two columns have been permuted, and (2) the labels of the rows and columns have been changed. For standard game theory, these changes in presentation are irrelevant, and it remains true that each confessing is the only Nash equilibrium. For Roemer's theory, however, this relabeling makes a big difference. First, the game is no longer symmetric (because, for example, the payoffs are not the same for both players, when both make choice 1). Second, there is now no simple Kantian equilibrium, since there is no choice such that all weakly prefer everyone making that choice to everyone making some alternative feasible choice (for example, player 1 most prefers all making choice 2, whereas player 2 prefers all making choice 1). Roemer makes no claim about what rationality requires under such conditions.

My objection here is *not* that Roemer's theory is silent in this case. Rather, it is that his theory gives different answers to what rationality requires in the two different ways of presenting the game. In the original presentation, Roemer's theory holds that rationality requires that both agents remain silent (not confess), but, in the second presentation, his theory is silent about what rationality requires. My claim is that they are the same game, and the different presentations change nothing. More exactly, my claim is that Kantian optimization presupposes that there is a privileged way of identifying the choices of different players (for example, player 1's choice to confess 'is the same choice' as player 2's choice to confess), but such interpersonal identifications of choice are arbitrary and irrelevant to rational choice. Let me explain.

Kantian optimization is well-defined only if each choice of one player can be uniquely identified with a distinct choice of the other player (for example, confess for *Row* is the same choice as confess for *Column*) in a way that is relevant to rational choice. This permits the choices of the two agents to be listed in the same order in the matrix of the game. Thus, the downward common diagonal has the players 'making the same choice' (their *n*-th choice for each *n*). This is why, in the above Prisoner's Dilemma Relabeled, *Row* and *Column* each choosing silent is not a Kantian equilibrium. They are making 'different choices'. *Row* is choosing her first enumerated choice, whereas *Column* is choosing his second enumerated choice. By assumption, a choice by *Row* is 'the same' as a choice by *Column* if and only if they are both the agent's *n*-th choice, for some *n*.

Roemer is quite aware of this issue. He explicitly points out (26–28) how a common diagonal, and hence a simple Kantian equilibrium, may exist for one interpersonal identification of the choices of different agents (one 'common choice space'), but not for another.

This interpersonal identification of choices seems completely arbitrary. Of course, we can identify one person's confessing with another person's confessing, but we can also identify it with the other person remaining silent (for example, if each person is doing what they most want to do, or if each person is doing what their mother told them to do). The question is whether such identifications have any normative significance for rational choice. Nash optimization, like most approaches to rational choice, does not require any specification of which choice by one player 'is the same choice' as that of the other. Roemer's framework, however, assumes that there is a privileged way of doing this (as given by the 'common choice space' from which agents choose).

My point is not that, although there are facts about which choice of one agent is *the same choice* as another agent's, rational choice is insensitive to these facts. It is rather that there are no such (interest-invariant) facts. My pushing the red button and your pushing the red button may be the same choice *relative* to our choices of what colored button to push, but they may be different choices with respect to what shape the button we push is (mine is square and yours is round), with respect to our intentions (for example, I push the button to save my mother's life, whereas you push it in order to kill your father), with respect to the anticipated consequences, etc. I am not, that is, claiming that facts of whether agents make the same choices are substantively irrelevant (as a permutation invariance condition might require). I am claiming that there are no such facts. As far as I know, no other theory of rational choice appeals to such facts.

For non-simple games, which I have not introduced, Roemer further assumes that the strategies have a 'natural' order, based on the 'effort' involved (for example, labor time). One can indeed hold a *moral* view according to which equal effort should receive equal reward, but I see little reason to think that rationality in non-cooperative games requires this, or that it requires the interpersonal identification of choices with the same effort levels. Indeed, in simple normal-form games (such as those discussed above), effort plays no role, and I see no reason to think that it does in non-simple games in which efforts (of some specified sort) are commonly known. This, however, is a complex issue, and I here set it aside.[2]

Let me propose, without endorsing, a friendly but radical revision to the formulation of Kantian optimization that avoids this problem. Let us note that simple Kantian optimization does not depend on any interpersonal comparisons of value, nor on any cardinal (that is, interval) measurability of the values of choices. It merely requires that, in symmetric games, each agent make a choice that, if all other agents make the same choice, maximizes the payoff for the agent and is Pareto optimal. Thus, Roemer's idea of 'same payoff for the same choice' is that of 'same intrapersonal ordinal rank for same choice'. For example, in the Prisoner's Dilemma (the first version in Table 1, where choice 1 for each agent is staying silent), the four outcomes are (8, 8), (1, 9), (9, 1), and (3, 3), and this is ordinally and non-comparably equivalent to respective outcomes of (3, 3), (1, 4), (4, 1), and (2, 2). For each agent, the ordering of the four outcomes is the same for both sets of numbers. Kantian optimization requires a choice that gives each player their best outcome on the common diagonal (that is, (3, 3) rather than (2, 2)), and that requires each to choose to stay silent. So, Roemer's symmetry condition does not (as Roemer knows) require that players get the same cardinal interpersonally comparable payoff when they make the same choice. It only requires that players

[2] Even for simple games, Roemer assumes that the choices of each agent have a 'natural' order. Although his definitions of simple Kantian equilibrium, symmetry, and Pareto optimality do not depend on this order, his definition of monotonicity does. Given, however, that monotonicity is relevant only as a sufficient condition for Pareto optimality, no natural order is required for simple Kantian optimization.

get the same (non-comparable, intrapersonal) ordinal rank for the same choice (for example, that when all make the same choice, they each get their second best payoffs).

Once, however, we drop the idea of a privileged interpersonal identification of choices, the appeal to the diagonal becomes arbitrary (since the diagonal depends on arbitrary identifications of choices). Instead, we should look at *all possible outcomes*, in terms of each individual's ordinal ranking, focus on those outcomes that, for example, give players the same intrapersonal ordinality (that is, all get their *n*-th best outcome), and require that each do their part in producing the outcome that makes that ordinality as good as possible. (For simplicity, I here ignore how ties in ordinality are handled.) For example, with respect to the intrapersonal ordinality of each outcome, the Prisoner's Dilemma Relabeled game (Table 2) is identical to the original Prisoner's Dilemma game (Table 1). In both cases, the only two outcomes with equal ordinality for the two agents are (*Silent*, *Silent*) with payoffs (3, 3), the second-best outcome for each, and (*Confess*, *Confess*) with payoffs (2, 2), the third-best outcome for each. The revised approach would thus require that each agent be silent. This agrees with Roemer's version where confessing for *Row* is the same action as confessing for *Column*, but, unlike Roemer's version, this result holds no matter how choices are interpersonally identified. Roemer's version, by contrast is silent when confessing for *Row* is the same action as staying silent for *Column* (since there is no common diagonal).

The approach just sketched requires *equality* of intrapersonal ordinal payoff, but sometimes equality will not be possible. Moreover, even when possible, it may be possible to make both players better off than they are under the most ordinally equal outcome. So, a more plausible approach is probably to require *leximinning the intrapersonal ordinal payoff* of outcomes (making the lowest ordinal payoff as high as possible, and then doing the same for the second lowest ordinal ranking, etc.). So, if the outcomes are (1, 1), (2, 2), (2, 4), and (4, 3), this revised version of cooperative optimizing would require each to do their part in bringing about (4, 3). I leave open how things are assessed when there is more than one outcome that leximins intrapersonal ordinal payoff (as in (1, 1), (2, 2), (3, 4), and (4, 3)).[3]

[3] If one allows, as Roemer does not, intrapersonal *cardinality* to be relevant to cooperation, cooperation might require leximinning *intrapersonal relative benefit*, defined, for a given agent, as $(X - Min)/(Max - Min)$, where X is the payoff for the agent for a given outcome, and Min and Max are, respectively, the smallest and the greatest payoffs for the agent in the choice situation. For example, if the five possible outcomes are (0, 100),

Call this (leximin) approach *ordinal cooperation*. Like Roemer's simple Kantian optimization, cooperation is to be understood as conditional on sufficiently trusting the other to cooperate appropriately. Unlike Roemer's simple Kantian optimization, it is insensitive to how choices are interpersonally identified.

Call a joint strategy *an ordinally cooperative equilibrium* just in case it leximins the ordinal rank of the outcome. We can now note the following trivial results, which correspond to Roemer's main results for simple Kantian optimization (23):

(1) Ordinally cooperative equilibria exist for all games (whereas Kantian equilibria do not).
(2) Ordinally cooperative equilibria are always Pareto efficient (whereas Kantian equilibria are not).

Of course, I here leave many key issues unresolved. My key claims are simply that (1) Roemer's appeal to the notion of 'same choice' is arbitrary and irrelevant to rational choice, and (2) one can preserve some of the ideas of Kantian optimization (although perhaps not much) by focusing on something like leximinning ordinal benefit. I don't claim that this is a plausible approach. I merely claim that it is more plausible than Roemer's version of Kantian optimization. Indeed, I shall now suggest that the form of cooperative reasoning in ordinal cooperation (and in Kantian optimization) is too strong. The role of cooperative reasoning, I shall suggest, is limited to selecting among Nash equilibria.

III.II. Does Rationality Require (or Even Allow) Cooperation with Cooperators?

Roemer claims that, for symmetric, single-stage games, where cooperation is Pareto optimal, and one's trust that the other (or others) will cooperate is sufficiently high, rationality requires that one cooperate. I have no new argument against this view. I will simply state a fairly standard objection. It applies not just against Kantian optimization, but also to all theories of rationality that require cooperation with trusted cooperators (such as ordinal cooperation above).

(1, 95), (2, 90), (3, 80), and (4, 0), instead of requiring each to do their part to produce the third-best outcome for each (namely, (2, 90)), this approach would require each to do their part to choose (3, 80), since this leximins intrapersonal relative benefits, (0.75, 0.8)—where the five relative-benefit pairs are (0, 1), (0.25, 0.95), (0.5, 0.9), (0.75, 0.8), and (1, 0).

		Column		
		C1	C2	C3
	R1	(7, 5)	(3, 2)	(2, 2)
Row	R2	(4, 3)	(6, 6)	(2, 3)
	R3	(4, 4)	(3, 4)	(5, 7)

Table 3: A game with three equilibria.

Suppose that we have a symmetric, single-stage, two-person Prisoner's Dilemma (and thus cooperation is Pareto optimal). Suppose that it is common knowledge that in this situation each agent is highly disposed to cooperate with the other. For example, suppose that it is common knowledge that there is only an arbitrarily small (infinitesimal, if you like) probability that a player will not cooperate with the other under the given conditions. Each agent is thus virtually (but not perfectly) certain that the other will cooperate. What does rational choice require? Given that, for each player, non-cooperation has higher payoffs, *no matter what the other player chooses*, it seems clear to me that rational choice requires non-cooperation, even though each agent is virtually (or even absolutely) certain that the other will cooperate.

This is not to say that cooperative reasoning is irrelevant to single-play games. Where there is more than one Nash equilibrium, I find it quite plausible that some form of cooperative reasoning is rationally required to choose among the Nash equilibria. The point is rather that rationality, I claim, limits the role of cooperative reasoning, if any, to the selection of Nash equilibria.

To illustrate this, consider, for example, the game in Table 3.

Here, the three (pure) Nash equilibria are (*R1*, *C1*), (*R2*, *C2*), and (*R3*, *C3*). From these three possibilities, only (*R2*, *C2*) leximins intrapersonal ordinal rank (second choice for each agent), and thus a restricted version of ordinal cooperation selects only this choice-pair as rational.[4] This seems reasonably plausible to me.[5] Of course, it's controversial that cooperative reasoning plays even this weak role in rational choice in non-

[4] Ordinal cooperation applied to select a Nash equilibrium will always select a joint strategy that is Pareto optimal *relative to the set of Nash equilibria*, but the selected joint strategy need not be Pareto optimal relative to the entire feasible set. For example, in the Prisoner's Dilemma, there is only one Nash equilibrium, and it is not Pareto optimal.

[5] In the example, none of the three Nash equilibria risk dominates any other. A more restrictive account of the role of cooperation might restrict it to risk undominated Nash

cooperative games. My claim is only that, even if it does, it does not play the strong role defended by Roemer.

In sum, although I agree with Roemer that cooperative reasoning is sometimes relevant to rational choice in non-cooperative games, I claim that: (1) the relevant cooperative reasoning does not depend on any privileged interpersonal identification of choices, and (2) the role of cooperative reasoning, if any, is limited to selecting among Nash equilibria. Obviously, these are big issues that warrant further analysis.

IV. CONCLUSION

How We Cooperate is well-written, philosophically informed, and informative. My discussion has addressed only the core idea of the book, which is the focus of only a small part of the book (roughly 40 of the 218 pages). There are also many important extensions (with production functions, etc.), and applications. Although many of these technical aspects won't be of interest to many philosophers, the presentation typically includes useful accompanying discussion. Those with strong interests in rational choice theory will definitely profit from reading it.

REFERENCES

Harsanyi, John C., and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: The MIT Press.

Roemer, John E. 2019. *How We Cooperate: A Theory of Kantian Optimization*. New Haven, CT: Yale University Press.

Peter Vallentyne, Florence G. Kline Professor of Philosophy at the University of Missouri, has held visiting professorships at Xiamen University, China, and at Central European University, Hungary. He has published over 100 articles in ethics and political philosophy, and edited or co-edited 16 volumes in those areas. He has been associate editor of the *Journal of the American Philosophical Association*, *Economics & Philosophy*, *Social Choice and Welfare*, *Ethics*, and *Politics, Philosophy, and Economics.*

Contact e-mail: <vallentynep@missouri.edu>

equilibria. For a definition of risk dominance, see, for example, Harsanyi and Selten (1988, 82).

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 98–108.
https://doi.org/10.23941/ejpe.v13i2.501

Do Kantians Drive Others to Extinction?

Jean-François Laslier
CNRS, Paris School of Economics

I. Introduction

From Thomas Malthus and Pierre Verhulst to Alfred Lotka and Vito Volterra, theoretical biology has studied the dynamics of living species (see Berryman 1992 for an account of this history). The interaction between theoretical biology and game theory (Smith 1982) has also been fruitful, and—as a result of this interaction—a whole discipline of evolutionary game theory has emerged (Weibull 1995).

Of particular interest in evolutionary game theory is the explanation of cooperative behavior and altruism based on evolutionary arguments. This is because the existence of cooperation may at first sight seem to be in contradiction with the 'individual selection' paradigm in biology. But, as John Roemer recalls in his book, *How We Cooperate: A Theory of Kantian Optimization* (2019), men (and animals too) routinely behave in a cooperative manner, sometimes even at their own expense. This explains why the question of cooperation has been a non-trivial puzzle in evolutionary biology (see the work of Hamilton 1963, 1964, and, more recently, Nowak and Sigmund 2005, or Alger and Weibull 2013). It is also a central question in economic theory, all the more after issues of incentives and selfish behavior became prevalent in mainstream economics. In the eighth chapter of *How We Cooperate*, Roemer applies the solution concept of Kantian optimization to coordination games in order to offer an evolutionary view of this concept. This kind of Kantian optimization is to be contrasted with what Roemer calls 'Nash behavior'.

Coordination requires giving some attention to what others do, and this is of course one element of cooperation. Games where individuals may settle on a low-quality outcome while coordination on a better one is also possible are therefore interesting case studies for the study of co-

Author's Note: I wish to warmly thank Vaios Koliofotis and Marina Uzunova who spotted and fixed an important mistake in the first version of this paper. All remaining errors are mine.

operative behavior, even if cooperation should not be reduced to efficient coordination. The point was introduced by Jean-Jacques Rousseau in his *Discourse on the Origin and Basis of Inequality among Men* ([1755] 1923). The *Discourse* presents an evolutionary perspective on cooperative behavior, albeit a pre-Darwinian and non-modern one. The well-known Stag Hunt game is discussed by Rousseau as an illustration of the fact that behavioral coordination requires a signaling device: some kind of 'language', but a 'language' that can be restricted to specific goals.

Roemer's analysis of the Stag Hunt game incorporates one important idea of modern evolutionary theory, namely the concept of evolutionary stable equilibrium (ESE). It also alludes to possible dynamics that sustain the stability of these equilibria. There are no explicit dynamics in Roemer's construction, but they are a key feature of evolutionary theories, as the word 'evolutionary' itself suggests. Such dynamics make no reference to Rousseau's idea of conceiving communication as a means to coordination but instead model some Darwinian selection process of the fittest.

In the modeling exercise Roemer performs in chapter eight of *How We Cooperate*, the non-Kantian agents are called "Nashers" (117). The word refers to an equilibrium concept, the Nash equilibrium, which, unlike ESE, has no associated dynamic process that would ensure some form of stability. The analysis in chapter eight, which compares the fate of Nashers and Kantians along their evolutionary dynamics, therefore needs to be spelled out more precisely. Additional insights can be provided by a double dynamics model that takes into account both the dynamics of selfish optimization, which might sustain Nash behavior, and the dynamics of selection or survival of Kantians. Below I explore in greater detail the differences between Roemer's model and the double dynamics model.

II. ROEMER'S MODEL

Although Roemer discusses symmetric coordination games in some generality, I will concentrate on one example: the so-called Stag Hunt game.

II.I. The Stag Hunt Game

The Stag Hunt game is commonly traced back to Rousseau who tells a story about how men gradually came to acquire the concept of mutual commitment:

> In this manner, men may have insensibly acquired some gross ideas of mutual undertakings, and of the advantages of fulfilling them: that is, just so far as their present and apparent interest was concerned: for they were perfect strangers to foresight, and were so far from trou-

> bling themselves about the distant future, that they hardly thought of the morrow. If a deer was to be taken, every one saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs.
>
> It is easy to understand that such intercourse would not require a language much more refined than that of rooks or monkeys, who associate together for much the same purpose. (Rousseau [1755] 1923, 209–210)[1]

A standard version of this game is the one described by Roemer on pages 29 and 128 of his book. It goes as follows. There are two hunters who belong to the same population. If they hunt separately, they will grab hares, and this is worth the normalized payoff 0. If they hunt the stag together, they will each earn the highest payoff: say, 1. Now, if only one hunts the stag, he will catch neither the stag, nor a hare, and will therefore earn a negative payoff: say, -1. Meanwhile, the other hunter, who is chasing hares alone, will catch more hares than he would if both players were hunting hares, and this yields a payoff 0.5. This is Roemer's (a, b) game (29) with $a = -1$ and $b = 0.5$ (see Table 1). This game is a classic of game theory; for instance, Brian Skyrms (2004) sees it as a fable that describes the key feature of the social contract, and Ken Binmore uses it to defend his claim that "fairness evolved as Nature's answer to the equilibrium selection problem in the human game of life" (2006, 11).

II.II. Kantians

It is clear that hunting the stag is the thing to do for efficiency reasons. That is to say, the outcome that obtains when both players hunt the stag is the unique Pareto optimal outcome. In the Stag Hunt game, under the

[1] In the original French, the exact quote reads:

> Voilà comment les hommes purent insensiblement acquérir quelque idée grossière des engagements mutuels, et de l'avantage de les remplir, mais seulement autant que pouvait l'exiger l'intérêt présent et sensible; car la prévoyance n'était rien pour eux, et loin de s'occuper d'un avenir éloigné, ils ne songeaient pas même au lendemain. S'agissait-il de prendre un cerf, chacun sentait bien qu'il devait pour cela garder fidèlement son poste; mais si un lièvre venait à passer à la portée de l'un d'eux, il ne faut pas douter qu'il ne le poursuivît sans scrupule, et qu'ayant atteint sa proie il ne se souciât fort peu de faire manquer la leur à ses compagnons.
>
> Il est aisé de comprendre qu'un pareil commerce n'exigeait pas un langage beaucoup plus raffiné que celui des corneilles ou des singes, qui s'attroupent à peu près de même.

	STAG	HARE
STAG	$(1,1)$	$(-1,0.5)$
HARE	$(0.5,-1)$	$(0,0)$

Table 1: The Stag Hunt game in normal form.

unique Kantian equilibrium—the strategy profile where strategies answer the question “what is the strategy I would like both of us to play?” (12)—both players hunt the stag because each player is better off when both hunt the stag than when both hunt hares. The key concept of Roemer’s analysis gives a very clear verdict in this game: any Kantian player hunts the stag.

II.III. Nashers

The game has two pure strategy Nash equilibria:

(1) Both players hunting the stag is a strict Nash equilibrium because, in that case, hunting the stag yields a payoff of 1 while chasing hares alone yields a payoff of only 0.5.

(2) Both players hunting hares is a strict Nash equilibrium because, in that case, hunting hares yields a payoff of 0 while chasing a stag alone yields a payoff of only -1.

Following Harsanyi and Selten (1988), the stag-hunting equilibrium is called the *payoff dominant* equilibrium while the hare-hunting equilibrium is called the *risk dominant* equilibrium. The game also has a mixed strategy Nash equilibrium:

(3) Both players deciding at random and independently to hunt the stag with a probability of 2/3 and to hunt hares with a probability of 1/3 is a Nash equilibrium because if one player uses this mixed strategy, then the other player’s average payoff remains the same whatever he does. (The second player thus has no strict incentive to choose one strategy rather than the other, so, under the usual hypothesis about choice under risk, he can as well randomize in any way, so they might as well randomize in the same way.)

Roemer’s definition of a ‘Nasher’ is not perfectly clear as a general definition. The word is used as a short-hand for the expression “Nash optimizer” (117), but what it means to be a ‘Nash optimizer’ depends upon a given Nash equilibrium—it is not determined by the game itself or the players’ strategies. Roemer writes that “If there are several Nash equilibria, a Nasher randomizes among them” (118). This is difficult to follow:

it is unclear whether this means that different Nashers end up playing different strategies, or that they manage to correlate their randomization so that everyone plays the same (randomly chosen) pure strategy. Moreover, in some instances, as in the game above, each one of the two pure strategies is played in a Nash equilibrium, but choosing at random does not, in general, result in an equilibrium. In fact, except for a very specific randomization scheme, mixed strategies are almost never best responses and are therefore almost never chosen by an optimizer.

Therefore, in order to understand what a 'Nasher' is, one has to look closely at how the concept is used.

II.IV. Roemer's Evolutionary Argument

The argument that leads to the conclusion that "Kantians drive Nashers to extinction" (125) is provided in the proof of Proposition 8.4 (121–122). The proof first fixes the Nash equilibrium under consideration and denotes the associated strategy by q^*. (Thus Nashers do not randomize among different equilibria.) A Nasher hunts the stag with probability q^* and hares with the complement probability $1 - q^*$. Roemer considers two cases: (1) $q^* = 0$ (hunting hares), and (2) the mixed equilibrium $q^* = q_1^*$ that has, in this game, the player hunting the stag with probability $q_1^* = 2/3$ and hares with probability $1/3$, yielding an average payoff of $1/3$. According to Roemer, the third case (Nashers hunting the stag) is not to be considered because in that case Nashers and Kantians cannot be distinguished.

Following the standard evolutionary model, one imagines that individuals are randomly matched in pairs (with no 'assortative matching'). Let ν be the proportion of Kantians in the whole population. Then, the average payoff of a Kantian ($V^K(\nu)$) and of a Nasher ($V^N(\nu)$) in the two cases above can be computed as follows:

(1) Nashers hunt hares ($q^* = 0$):

$$\begin{array}{rcccccl} V^K(\nu) & = & \nu \cdot 1 & + & (1-\nu)\cdot(-1) & = & 2\nu - 1 \\ V^N(\nu) & = & \nu \cdot \frac{1}{2} & + & (1-\nu)\cdot 0 & = & \frac{\nu}{2} \end{array}$$

(2) Nashers use the mixed strategy $q^* = q_1^* = 2/3$:

$$\begin{array}{rcccccl} V^K(\nu) & = & \nu \cdot 1 & + & (1-\nu)\cdot\left(\frac{2}{3}-\frac{1}{3}\right) & = & \frac{2+4\nu}{6} \\ V^N(\nu) & = & \nu \cdot (\frac{2}{3}+\frac{1}{3}\cdot\frac{1}{2}) & + & (1-\nu)\cdot\frac{1}{3} & = & \frac{2+3\nu}{6} \end{array}$$

Now, it is clear that, in the case where Nashers use the mixed strategy q_1^*, Kantians have an evolutionary advantage because their payoff is larger than the payoff of Nashers: $V^K(v) = {}^{(2+4v)}/_6 \geq {}^{(2+3v)}/_6 = V^N(v)$.

When Nashers hunt hares, the advantage will be on the side of Nashers or Kantians depending on the value of v: Kantians have an advantage if and only if v is large enough ($v \geq {}^2/_3$). In other words, if Nashers hunt hares but most players hunt the stag, Nashers earn less, on average, than the other agents.

This is Roemer's argument for the claim that Kantians drive Nashers to extinction.

III. A Double Dynamics Model

In the argument above, there is only a sketch of the evolutionary analysis that is necessary for a convincing evolutionary argument. In particular, 'Nashers' are neither optimizing agents (as they should be in an economic model of rational behavior), nor adapting agents (as they should be in a behavioral model of learning), nor evolving agents (as they should be in a biological model of Darwinian selection)—they are just stubborn hare hunters in one case, and (quite strangely) stubborn users of a specific mixed strategy in the other case. I now propose a standard model of a replicator-dynamics type in order to study, for this game, the evolution of a population that consists of Kantian individuals (in Roemer's sense) and of adaptive individuals. I will simply call the non-Kantian agents 'selfish', although one could think of many names for them.

III.I. Replicator Dynamics

The following explication uses the most standard mathematical model of evolution called the replicator dynamics. So I begin with a very brief presentation of this model.

First, the idea of *fitness* defines the number of offspring of some replicating unit as a function of its environment, so that a population of size n characterized by a fitness per individual f grows at the rate f. In discrete time, the population of size n will be of size $f \cdot n$ at the next generation, and, in continuous time, the time derivative of n is $\dot{n} = {}^{\mathrm{d}n}/_{\mathrm{d}t} = f \cdot n$ with f being now a replication rate by unit of time.

With two groups $i = 1, 2$ of size n_i and fitness f_i each, writing $n = n_1 + n_2$ and $x_i = {}^{n_i}/_n$, one gets in full generality:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{n_i}{n}\right) = \frac{\dot{n}_i n_j - \dot{n}_j n_i}{n^2}$$

That is:

$$\dot{x}_i = x_i x_j \cdot (f_i - f_j)$$

Interestingly, this differential equation, which generates replicator dynamics, appears in like form in different models that describe (i) population genetics, (ii) social imitation, or (iii) individual adaptive learning (see Laslier, Umbhauer, and Walliser 2006). We will now apply this idea to obtain an evolutionary model for Roemer's argument.[2]

In the absence of Kantians, the standard evolutionary analysis of the family of Stag Hunt games indicates that a pure strategy equilibrium is reached in the long run: if the initial composition of the population contains a sufficient number of individuals of one type (be it stag hunters or hare hunters), coordination on this type will occur in the long run. The mixed strategy equilibrium is, on the contrary, unstable and is not reached. To introduce Kantian players, I propose the following model.

Let $\nu(t)$ be the proportion of Kantians in the population at time t. This proportion will vary with time. Following the evolutionary paradigm, non-Kantians will not be assumed to jump directly to some optimal or 'Nash' strategy, but they will adjust their strategies gradually with time. So let $x(t)$ denote the proportion of stag hunters among the non-Kantians. In the whole population the proportion of stag hunters is therefore:

$$y = \nu + (1 - \nu)x$$

Two processes of evolution are coupled: within non-Kantians for their choice of strategy, and between Kantians and non-Kantians. The two processes may occur at different speeds.[3] For instance, one may wish to study the case where selfish individuals can change strategy relatively quickly while it is only at a slow pace that selfish individuals become Kantians or Kantians become selfish. This is rather natural: it means that selfish individuals adjust their behavior by choosing a best response to the circumstances, if not instantly, at least relatively quickly. Roemer's definition of a Nasher does not presuppose a dynamic adjustment process—the underlying assumption is that Nashers find best responses instantly. Therefore, in order to relax this assumption, I will consider the case where this pro-

[2] Note that, in evolutionary game theory, an important literature exists, which deals with the stag-hunt problem and with extensions of the basic game (Kandori, Mailath, and Rob 1993; Samuelson 1997). The main focus in this literature is on the question of communication: does language help to coordinate on the payoff dominant equilibrium? Can one even explain the emergence of language as a coordinating device that allows forward induction in stag-hunt situations? See chapter eight in Samuelson (1997).

[3] The same idea of a two-level dynamic process is used in Laslier and Öztürk Göktuna (2016), and in Öztürk Göktuna (2019).

cess is not instantaneous but relatively faster than the transition from selfish to non-selfish behavior. In my model, individuals can be viewed from the perspective of two time scales. In the long-run or evolutionary time, it takes many generations to converge on a long-run equilibrium in accord with the dynamic process of natural selection (a slow process of change). Being a Kantian or a selfish optimizer is determined by this long-term evolution. In the short-run or decision-making time, rational individuals make choices or acquire new strategies via social learning (a fast process of change). A selfish optimizer chooses her strategy in this shorter term. Hence, the overall evolutionary process has the complex structure of a slow evolutionary and a fast decision-making time horizon.

Among selfish individuals, the difference in payoff between those who hunt the stag and those who hunt hares is:

$$\delta^1 = [y - (1-y)] - \left[\frac{y}{2}\right] = \frac{3y}{2} - 1$$

Let s be the adaptive speed of the selfish individuals. The replicator dynamics within this group is described by the following differential equation (where $\dot{x}$ denotes the time derivative ${}^{\mathrm{d}x}/_{\mathrm{d}t}$):

$$\dot{x} = s \cdot x(1-x) \cdot \delta^1 \tag{1}$$

At the level of the whole population, the difference in payoff between Kantians and selfish individuals is obtained as follows: for the Kantians, since they all play the same strategy (stag), the average payoff is simply the average payoff of the stag strategy, that is: $2y - 1$. For the non-Kantian group, one should think of them as carriers of a ‘selfish’ gene whose fitness is the average fitness of the individuals who carry it. Therefore, the relevant payoff for the evolution of the selfish population is the average payoff in this population. That is:

$$x \cdot (2y-1) + (1-x) \cdot \frac{y}{2}$$

Hence:

$$\delta^2 = [2y-1] - \left[x(2y-1) + (1-x)\frac{y}{2}\right] = (1-x)\left(\frac{3y}{2} - 1\right)$$

If the adaptive speed of Kantianism is normalized to 1, then the associated replicator dynamics is described by the following differential equation:

$$\dot{v} = v(1-v) \cdot \delta^2 \tag{2}$$

The quantities y, δ^1, and δ^2 depend solely on the variables v and x, so equations (1) and (2) define a system of differential equations in the square $(v, x) \in [0,1] \times [0,1]$.

III.II. Results

Figure 1 collects all results. The left panel of the figure is drawn for a speed ($s = 5$) such that selfish individuals adapt their strategy relatively quickly. This is the most natural assumption. On the horizontal axis is the proportion of Kantians, v, and on the vertical axis is the proportion of stag hunters among the population of selfish individuals, x. The graph represents the flow of the differential system.

The lower left corner, $(0, 0)$, corresponds to the situation where there are no Kantians and everyone is hunting hares. The upper right corner, $(1, 1)$, corresponds to the situation where the whole population consists of Kantians and everyone hunts stags. The upper segment, where $x = 1$, also describes situations of full cooperation, where everyone hunts stags but some do it because they are Kantians and others do it for selfish reasons.

Following the arrows, one can observe the fate of the system. The flow is divided in two: a lower left region that points to $(0, 0)$, and an upper right region that always reaches the upper segment (where $x = 1$). Changing the speed, as depicted in the right panel of Figure 1, confirms this point.

The two basins of attraction are separated by the curve of the equation $\dot{v} = 0$. That is, ${}^{3y}/_{2} = 1$ or, in terms of v and x, $3v + 3(1 - v)x = 2$. This is part of the hyperbola $x = {}^{(2-3v)}/_{3(1-v)}$ and is independent of s.

IV. Conclusion

In section 8.2 of *How We Cooperate*, after studying the Stag Hunt game in isolation, Roemer considers several games. He concludes that if Nature chooses at random what kind of a game is played—either a coordination game or a Prisoner's Dilemma—then Nashers and Kantians can co-exist. Instead, this note focused on a single coordination game.

Faced with the claim that "In games of pure coordination, Kantians drive Nashers to extinction" (125), readers of Roemer's book might be tempted to over-interpret the expression 'Nasher'. They may thus conclude that, in coordination games, Kantian optimizers (in the sense of Laffont 1975, and Roemer 2019) have some efficiency advantage that makes them fitter, from an evolutionary point of view, than selfish optimizers. This is not true. In the Stag Hunt game, either Kantians are wiped away by

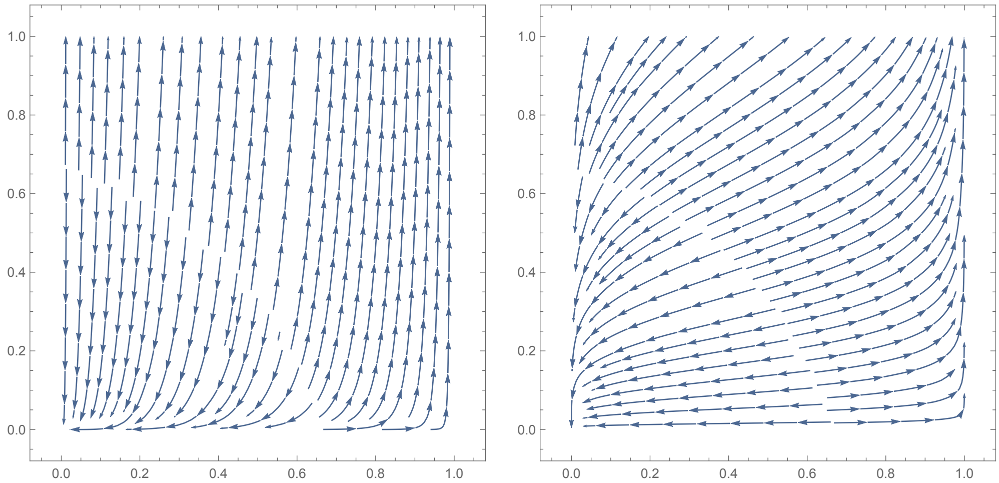

Figure 1: Stag Hunt game. Coupled dynamics for the proportion of Kantians (horizontal axis) and the proportion of cooperators among non-Kantians (vertical axis). Left panel: $s = 5$; right panel: $s = 0.2$.

selfish individuals who do not cooperate, and thus Kantians are 'driven to extinction' by the selfish optimizers, or both remain as some fraction of the population.

Focusing exclusively on 'equilibria' to describe and analyze collective outcomes may be misleading. Following his analysis, Roemer writes that "According to Proposition 8.4, there are no (a, b) games where both Kantian and Nash players exist with positive frequencies in a stable equilibrium" (123). It is not clear what is meant here by 'stable equilibrium' but, as I showed above, the natural process that sustains evolutionary stability leads to, depending on the initial conditions, two possible outcomes. One possibility is that hare hunters (who can be called Nashers) drive Kantian stag hunters to extinction. The other possibility is that Kantians and selfish optimizers (who can also be called Nashers) co-exit, all hunting stags but for different reasons.

REFERENCES

Alger, Ingela, and Jörgen W. Weibull. 2013. "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching." *Econometrica* 81 (6): 2269–2302.

Berryman, Alan A. 1992. "The Origins and Evolution of Predator-Prey Theory." *Ecology* 73 (5): 1530–1535.

Binmore, Ken. 2006. "The Origins of Fair Play." Papers on Economics and Evolution Working Paper No. 0614. Max Planck Institute of Economics, Jena.

Hamilton, William D. 1963. "The Evolution of Altruistic Behavior." *The American Naturalist* 97 (896): 354–356.

Hamilton, William D. 1964. "The Genetical Evolution of Social Behaviour. I and II." *Journal of Theoretical Biology* 7 (1): 1–52.

Harsanyi, John C., and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games.* Cambridge, MA: The MIT Press.

Kandori, Michihiro, George J. Mailath, and Rafael Rob. 1993. "Learning, Mutation, and Long Run Equilibria in Games." *Econometrica* 61 (1): 29–56.

Laffont, Jean-Jacques. 1975. "Macroeconomic Constraints, Economic Efficiency and Ethics: An Introduction to Kantian Economics." *Economica* 42 (168): 430–437.

Laslier, Jean-François, and Bilge Öztürk Göktuna. 2016. "Opportunist Politicians and the Evolution of Electoral Competition." *Journal of Evolutionary Economics* 26 (2): 381–406.

Laslier, Jean-François, Gisèle Umbhauer, and Bernard Walliser. 2006. "Game Situations." In *Evolutionary Microeconomics*, edited by Jacques Lesourne, André Orléan, and Bernard Walliser, 67–112. Heidelberg: Springer.

Nowak, Martin A., and Karl Sigmund. 2005. "Evolution of Indirect Reciprocity." *Nature* 437 (7063): 1291–1298.

Öztürk Göktuna, Bilge. 2019. "A Dynamic Model of Party Membership and Ideologies." *Journal of Theoretical Politics* 31 (2): 209–243.

Roemer, John E. 2019. *How We Cooperate: A Theory of Kantian Optimization.* New Haven, CT: Yale University Press.

Rousseau, Jean-Jacques. (1755) 1923. "Discourse on the Origin and Basis of Inequality among Men." In *The Social Contract and Discourses by Jean-Jacques Rousseau*, translated by George D. H. Cole, 155–246. London: J. M. Dent & Sons.

Samuelson, Larry. 1997. *Evolutionary Games and Equilibrium Selection.* Cambridge, MA: The MIT Press.

Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure.* Cambridge: Cambridge University Press.

Smith, John Maynard. 1982. *Evolution and the Theory of Games.* New York, NY: Cambridge University Press.

Weibull, Jörgen. 1995. *Evolutionary Game Theory.* Cambridge, MA: The MIT Press.

Jean-François Laslier is a member of the CNRS (French National Centre for Scientific Research) and a professor at the Paris School of Economics. His interests include mathematical economics, games, social choice, and political science. He conducts research on democracy, and, in particular, on electoral systems and voting behavior, from the formal and the experimental points of view. He publishes in the two fields of economics and political science.
Contact e-mail: <jean-francois.laslier@ens.fr>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 109–126.
https://doi.org/10.23941/ejpe.v13i2.529

Response to Braham and van Hees, Sher, Vallentyne, and Laslier

JOHN E. ROEMER
Yale University

I am most grateful to the five commentators for the time they spent reading and thinking about *How We Cooperate: A Theory of Kantian Optimization* (*HwC*) (Roemer 2019). They have forced me to think once more about a number of my claims. In particular, I have been ambiguous about whether Kantian optimization is a rational approach, in some situations, or whether it is a moral one. I hope I clarify my present view below. Despite what I say here, I certainly do not believe I have had the last word on this topic.

COMMENT ON MATTHEW BRAHAM AND MARTIN VAN HEES

The summary of my theory of simple Kantian optimization by Braham and van Hees in section I of their contribution is admirable. They note that the theory prescribes which action to take in a game, while Kant's categorical imperative is an instruction of which *maxim* to apply to the choice of one's actions. I presume this is correct.

In section II, they propose to limit their discussion to games with a common diagonal. These are games in which a simple Kantian equilibrium exists: that is, all players will agree on the common-action strategy profile that is best from *each* player's point of view. This is the most persuasive example of Kantian optimization.

The authors describe a Prisoner's Dilemma in which the moral act may be to '*not* cooperate'—the farmer whose family is starving grazes his cow on the overused commons in order to provide food for his family. This is to be contrasted with two prisoners who are gang members and have committed a crime together, as in the usual story told to explain the payoff matrix of the Prisoner's Dilemma. In this case, 'cooperating' is immoral because the crime in question was an immoral act. So, in the first example, the farmer plays the Nash strategy (individualistic), which is morally correct, and in the second example, the cooperative strategy profile (the

Kantian equilibrium) of the game is morally bad. The examples show that one cannot judge the morality of actions without knowing the context in which the payoff functions are defined. By context, I mean the 'extenuating circumstances', which would be, I think, the 'circumstances' that the authors refer to in the "tripartite relation" they introduce in section III (37).

Of course, I agree. In *HwC*, I gave the price-fixing behavior by a cartel of oligopolistic firms as an example of a multiplicative Kantian equilibrium, which was ethically bad because it hurts consumers (53–54). This is the reason (perhaps among others) that I explained that my use of 'Kantian' was not supposed to convey a claim that 'Kantian' equilibria are Kantian in Immanuel Kant's sense of deeply moral.

Braham and van Hees' example, later in their section II (35–36), of the public good that requires different kinds of labor to produce illustrates a case where there is no simple Kantian equilibrium. I agree that there are such games and I will have more to say about the importance of this point in my comment on Itai Sher's objection about existence (115–116). But let me add a few clarifications here. When I presented simple Kantian equilibrium in *HwC*, I restricted my discussion to games in which every player has the same strategy space, an interval on the real line; thus, it is assumed that each player contributes 'effort' which can be measured in a common unit. This does not require the unit be labor time; it could be *efficiency* units of labor, which does permit players to measure their contribution in the same unit. However, if one player contributes carpentry labor and another contributes plumbing labor, there is in general no common unit in which we can measure both contributions.[1] In some situations, we can still discuss Kantian optimization, but this is a generalization away from simple Kantian equilibrium.

Braham and van Hees' two tango games (the 'Tango game' in its first formulation in their Table 1, 35; and the 'Modified Tango game' in their Table 2, 36) illustrate the fact that for simple Kantian equilibrium 'correct' labelling of strategies matters. Notice that the first Tango game is not a monotone increasing game. It follows that it does not have the good features of Kantian equilibrium—that equilibria are Pareto efficient—which apply *only* to strictly monotone games. Now, in the authors' second formulation of the Tango game, where the payoff matrix is as in Table 1 below, the game *is* strictly monotone increasing; there is a common diagonal,

[1] If the environment is a market economy, then the wages of the carpenter and the plumber provide a common unit, and if we are in a competitive equilibrium, then the wages reflect marginal products, a real common unit.

	S	N
S	(3, 3)	(0, 0)
N	(0, 0)	(–1, –1)

Table 1: Braham and van Hees’ Modified Tango game.

	C	D
C	(2, 2)	(0, 3)
D	(3, 0)	**(1, 1)**

Table 2: Both player Row and player Column act according to maxim m^i (the Prisoner’s Dilemma).

and the simple Kantian equilibrium is Pareto efficient. This formulation requires identifying a person’s strategy *not* as being ‘lead’ or ‘follow’ but as being ‘specialize in one’s expertise (*S*)’ or ‘do not specialize in one’s expertise (*N*)’. As the authors acknowledge, I made exactly the same point in *HwC* (26–28) in discussing the ‘Battle of the Sexes’ where the game becomes one with a common diagonal only when we re-label the original strategies of ‘boxing match’ and ‘dance recital’ as ‘one’s favorite event’ and ‘one’s disfavored event’.

It is not surprising that ‘correct’ labelling of strategies matters, because the notion of ‘playing the same action’ requires knowing what ‘same’ means. Nash equilibrium does not require this: the labelling of actions does not matter. In this sense, the payoff matrix of a game for Nash players requires no notion of the ‘sameness’ of strategies, whereas for Kantian players, additional information is required to correctly write down the payoff matrix. I will have more to say about this in my comment on Itai Sher’s objection about strategic non-equivalence (118–120).

Let me turn to Braham and van Hees’ interesting proposal of *Kantian* Kantian equilibrium. I will study the example that they provide. There are two possible maxims: the authors call them “individual” and “collective” (39). I find this confusing, because I use ‘individualistic’ and ‘cooperative’ to refer to optimization protocols (Nash versus Kant) and the terms ‘individual’ and ‘collective’ risk conflating maxims with protocols. So, let’s call the two maxims ‘self-regarding’ and ‘sociotropic’. Each maxim induces a preference order over the strategy profiles, where a profile is an ordered pair whose components are taken from the set $\{C, D\}$. There are five relevant games (one of which is really a ‘game’) to consider. The first is the standard Prisoner’s Dilemma, where the maxim of both prisoners is self-regarding (denoted by m^i), and the preferences are those described by the familiar payoff matrix of the Prisoner’s Dilemma (Table 2).

The numbers have only ordinal meaning. Thus, the *preference order* induced by the self-regarding maxim is, for the Row player: $(D, C) \succ$

	C	D
C	(0, 0)	(1, 2)
D	(2, 1)	**(3, 3)**

Table 3: Both player Row and player Column act according to maxim m^c (the Prisoner's Harmony).

	C	D
C	(0, 2)	(1, 3)
D	(2, 0)	**(3, 1)**

Table 4: Player Row acts according to maxim m^c; player Column acts according to maxim m^i.

$(C,C) \succ (D,D) \succ (C,D)$. The Column player's preference order is: $(C,D) \succ (C,C) \succ (D,D) \succ (D,C)$. The unique Nash equilibrium is (D,D), indicated in boldface. We understand these preferences to follow from the self-regarding maxim which, here, leads to desiring to minimize the time one serves in prison.

If the two players act according to the sociotropic maxim (denoted by m^c), then they adopt Prisoner's Harmony preferences, as defined by the payoff matrix in Table 3.

The preference order over strategy profiles of the Row player is now: $(D,D) \succ (D,C) \succ (C,D) \succ (C,C)$. These are the preferences induced by desiring to build a law-abiding society. My reasoning is as follows. The best action, from the social viewpoint, is that both prisoners confess to the crime. This is (D,D). Second-best for the Row player is that he confesses even if Column does not confess (this is (D,C)). The third-best result for Row is that even if he does not confess, the other prisoner does, (C,D). The worst result from the social viewpoint is that neither confess, (C,C). The unique Nash equilibrium of this game (in pure strategies) is (D,D), again indicated in boldface.

Now, Braham and van Hees limit their analysis to the "universal adoption" of maxims (39)—that is, games where both players follow the same maxim. This is motivated by the appeal to Kantian ethics because Kant's Formula of Universal Law, which Braham and van Hees are formalising, is about the possibility of a maxim becoming a universal law (39). The two relevant games in Braham and van Hees' analysis are thus those in Tables 2 and 3.

I would now like to move away from this analysis by asking the following—non-Kantian but still interesting—question: what about games where players are following *different* maxims? To answer this question, we should consider the other two possible maxim 'profiles', which are (m^c, m^i) and (m^i, m^c) where I keep the authors' notation—but with my nomenclature m^i is the self-regarding maxim and m^c is the sociotropic

	C	D
C	(2, 0)	(0, 2)
D	(3, 1)	**(1, 3)**

Table 5: Player Row acts according to maxim m^i; player Column acts according to maxim m^c.

	m^c	m^i
m^c	(D, D)	(D, D)
m^i	(D, D)	(D, D)

Table 6: The equilibrium outcomes.

maxim. Suppose the Row (Column) player acts according to m^c (m^i). Then, the payoff matrix is given in Table 4.

The unique pure strategy Nash equilibrium is again (D, D). It happens to be Pareto efficient, although I see no reason this will be the case in general. Finally, if the Row player uses m^i and the Column player uses m^c, then we have Table 5, and the unique Nash equilibrium is, of course, again (D, D).

Next, I will write down the 'outcome' matrix for the four games just analyzed (Table 6). It is important to say that Table 6 does not describe a game, because players do not have preferences over *maxims*. This table simply records the Nash equilibria when each of the two players can adopt either the maxim m^c or m^i.

From Table 6, we see that the analysis with maxims *provides no way of distinguishing* which maxims obey Kant's categorical imperative—they all give rise to the 'morally good' Nash equilibrium that both players confess to the crime (D).

It seems to me the lesson is that more examples must be studied in order to see under what conditions this kind of analysis yields interesting results.

Clearly, the authors are interested in situations that are more complex than games. The data needed to implement their algorithm include more than preference orders over strategy profiles. I appreciate their attempt to think about how to model Kant's categorical imperative that is more faithful to Kant's ideas than my approach. What I've offered is a quite simple answer to the query: 'How would players who wish to cooperate optimize in a game as classically defined?' Braham and van Hees are trying to answer the much more complex question of what general maxim should guide those who face many games in life, *seriatim*. I do not object to the idea of having two stages, in which the 'game' where the strategies are maxims induces standard games with preference orders; but the procedure must be well-defined, a non-trivial requirement.

So, what do I think is the morality represented by Kantian optimization? I vacillate between two interpretations. The first is that Kantian optimization is the formalization of the instruction contained in the proverb 'if we do not hang together, we will, most assuredly, each hang separately'. This is an instruction to cooperate, not for altruistic reasons, but because it is the best way to defeat our common enemy, or succeed in our common struggle, which is of value to *each* of us. As such, it is a special case of the claim that cooperation is rational, in the sense of advancing the self-interest of each. This is based upon the premise that, viewed correctly, we are all in the same boat, and therefore it is likely that we are best served by all taking the same action. It turns out that this conclusion is true, however, only in *monotone games*; it does not follow for non-monotone games even if they have symmetric payoff matrices.

This interpretation, that we should act as if we are all in the same boat, is beautifully explained in the writing of Martin Niemöller quoted in *HwC* (6). Niemöller's point is that superficial differences in our situations may obscure the fact that, properly viewed, we are all in the same boat, and our unified behavior is therefore justified. Niemöller's example is well-illustrated by a different but related maxim, 'all for one and one for all', which was the maxim recommending solidarity in the American labor movement.[2]

The second interpretation is that *fairness* requires symmetric behavior if we are all similarly placed. If I am contemplating increasing my grazing on the commons by 10%, I must say it is fair for everyone to do likewise; I can justify my mooted action if and only if I would *prefer* the situation in which everybody increases his grazing by 10%. This interpretation is moral by virtue of symmetry: likes should be treated alike.[3] Remember, we are only here discussing situations in which there exists a simple Kantian equilibrium. In games with heterogeneous preferences, I define more complicated forms of Kantian optimization, which also embody symmetric behavior, even though preference orders in those games are not the same. I consider it a gift that multiplicative and additive Kantian optimization produce Pareto efficiency in strictly monotone games, even when

[2] See John Ahlquist and Margaret Levi (2013) for a history of the International Longshoreman's and Warehousemen's Union, and the centrality of this maxim in their behavior.

[3] Recall Braham and van Hees' example of the poor farmer who (morally) grazes his cow more than others, because his children are hungry. This poor, moral farmer would *not* advocate that other farmers also increase their grazing by 10%, because the others do not have hungry children. However, the *multiplicative Kantian equilibrium* may well be Pareto efficient in this situation! In that equilibrium, poor farmers may well be allowed to graze more than rich farmers.

preferences differ. It's a gift because I can see no a priori reason that when preference orderings differ, symmetric behavior as defined by these optimization protocols should 'work'. One can perhaps glimpse the mechanism in that the tragedy of the commons and the free-rider problem come about because the Nash optimizer ignores the externalities, positive or negative, produced by virtue of her behavior. The Kantian optimization protocols internalize these externalities, to use economics' jargon. But that they should internalize these externalities *to just the right degree* to achieve Pareto efficiency still amazes me.

COMMENT ON ITAI SHER

Itai Sher challenges my analysis on three "technical issues" and advances one "non-technical" challenge (44). I will respond *seriatim*.

Sher writes:

> The three technical issues concern existence, efficiency, and strategic equivalence. First, Kantian equilibrium may not exist. This leads to the question: what is an integrated normative approach to interactions modeled as games that leads to prescriptions both when Kantian equilibrium exists and when it fails to exist? Second, while Roemer documents important cases in which Kantian equilibria are efficient and Nash equilibria are not, it is also easy to construct examples of inefficient Kantian equilibria. This matters insofar as, in the book, efficiency plays an important role in justifying Kantian equilibrium. Third, by relabeling strategies, it is possible to construct strategically equivalent games whose Kantian equilibria differ, whereas it is not possible to do this for Nash equilibrium. [… This] does imply that the informational requirements for Kantian equilibrium are stronger than the informational requirements for Nash equilibrium […]. (44)

My responses:

1. Existence

It is indeed the case that simple Kantian equilibria rarely exist in games. The most convenient sufficient condition for existence of a simple Kantian equilibrium is that the players order the strategy profiles $\{(s, s, \ldots, s) | s \in S\}$ in the same way. I call this the *common diagonal property* (see 23, Proposition 2.1, in *HwC*). Existence is a rare occurrence. It is true for symmetric games, but hardly ever true otherwise. The reason I introduce simple Kantian equilibrium, despite its generic non-existence, is that it implements most literally Kant's instruction that 'one should take that action that one would will be universalized'.

Simple Kantian equilibrium begs to be generalized, and most of the book studies three generalizations: multiplicative and additive Kantian equilibrium, and φ-Kantian equilibrium. These equilibria exist very generally in games where the common strategy space is a real interval—the mathematical condition for existence is the same as for Nash equilibrium. To see this, one has to define the 'best-reply correspondence' for a Kantian optimizer. This is done in the proof of Proposition 7.3 (110 in *HwC*). The condition that guarantees existence of φ-Kantian equilibrium is that the best-reply correspondences be upper-hemi-continuous and convex-valued, as the proof of Proposition 7.3 shows. This is also the essential condition for existence of Nash equilibrium.

I do agree with Sher that the essential requirement for defining Kantian equilibrium is that the strategy space be a real interval (uni-dimensional); Nash equilibrium has no such requirement. But I do not take this to be a problem of *existence*: it is rather a result of conceptualizing *what cooperation means* if players have multi-dimensional strategies or draw strategies from different spaces. Suppose a carpenter and an architect wish to cooperate in building a house. Here the strategy spaces are different for the players, although each may be unidimensional. How can one conceptualize what it means for 'each to take the action she would will be universalized'? The conceptual problem is even worse if the strategy spaces are multi-dimensional. Conceptualizing cooperation in such problems is admittedly something I have not done, except for chapter 10, entitled "A Generalization to More Complex Production Economies". There, I provided a definition of Kantian equilibrium where different players have different occupations (149ff.); but the main point of that chapter is that it's difficult to extend the Kantian approach to such multi-dimensional problems. I repeat: this is not a non-existence problem, it is a deep problem of conceptualizing cooperation when those who contribute to the project have very different roles. Clearly, what we think of as cooperation in such situations occurs in reality, and I invite others to think how to model it.

2. Efficiency

Sher's claim that it is easy to construct examples of inefficient Kantian equilibrium is bizarre. What I prove is that if the game is strictly monotone, then Kantian equilibria, if they exist, are Pareto efficient.[4] There is no claim that Kantian equilibria in non-monotone games are efficient. Strictly monotone *increasing* games are public-good games—each player's contribution to the common project has positive externalities for other players. While Kantian equilibria in such games are efficient, Nash equilibria are generically inefficient: see Proposition 3.3 in *HwC* (44). This is such an important fact that it has a name: the free-rider problem. The free-rider problem occurs because Nash optimizers do not take into account the positive externality that their contribution provides to other players, so the private and social benefits of contributions are not the same. Strictly monotone *decreasing* games are ones with congestion effects: here, Nash equilibrium is also generically inefficient, while Kantian equilibria are efficient. The Nash inefficiency here is so important that it, too, has a popular name: the tragedy of the commons.[5]

That is, Kantian optimization 'resolves' what I think are the two greatest social pathologies of Nash optimization—its failure to deal successfully with positive and negative externalities. Indeed, it's for this reason that I ask the reader to ponder carefully Kantian optimization, not to quickly dismiss the idea as utopian. I give a number of examples where I believe Kantian behavior is prevalent, and I think we should search for other examples (see 14–16, section 1.5, in *HwC*).

Kantian optimization requires, of the player, that she *internalize* the externalities, positive or negative, associated with her contribution (strategy). It does this not by modeling players as altruistic, but by requiring the player to consider how she would feel if others took the action similar to the one she is contemplating taking. The simple example I gave in the book is of the parent and child walking along the beach. The child throws her candy wrapper on the sand. The parent might say: 'Child, don't do that. It spoils this pristine beach for other children.' This response

[4] The only exception is that multiplicative Kantian equilibria must be strictly positive to be efficient.

[5] There is an important point about the meaning of Pareto efficiency. When I am discussing a game, Pareto efficiency means efficiency in the game. In a monotone game, all types of Kantian equilibria (additive, multiplicative, etc.) are Pareto efficient in the *game*. This claim, for instance, applies to all strictly monotone 2 × 2 symmetric games. However, in games that are part of an economy where marginal products can be defined, there is a more demanding concept of efficiency: namely, Pareto efficiency in the *economy*. Here, efficiency will only hold for specific kinds of Kantian equilibrium. For example, in the fishing economy, the *multiplicative* Kantian equilibrium is Pareto efficient in the *economy,* and, in the hunting economy, *additive* Kantian equilibrium is efficient in the economy. However, additive Kantian equilibrium fails to be efficient in the fishing economy, as does multiplicative Kantian equilibrium in the hunting economy. See chapter 3 of *HwC*.

attempts to evoke altruism in the child. On the other hand, the parent might say: 'Child, how would you feel if all the other children threw their candy wrappers on the beach?' This invokes the Kantian categorical imperative: it *internalizes the externality* of the child's action, forcing the child to contemplate how she would be affected were others to do the action she is doing. I advocate the second response. Most people (excepting psychopaths) will feel moral qualms when confronted with the second response. My conjecture is that the moral reaction is more pervasive than the altruistic reaction upon which the first parental utterance depends.

The nice result is that modeling the Kantian protocol in monotone games produces *just the right amount* of internalization of the externality, in that the equilibrium associated with this reasoning among all players is Pareto efficient—it doesn't overshoot or undershoot in rectifying the Nash pathology. God is in the mathematics.[6]

3. Strategic Non-Equivalence

Sher and I agree that the way we name the strategies matters for Kantian optimization, but not Nash optimization. He sees this as a weakness of the Kantian approach; I see it as fundamental to it. I gave the example of the Battle of the Sexes in *HwC* (26–28). Here, the conventional way of naming the strategies is 'go to the boxing match' and 'go to the dance recital'. It's postulated that the man prefers the boxing match and the woman prefers the dance recital. With the strategies labelled 'Box' and 'Dance' the 2 × 2 payoff matrix is asymmetric (28, Table 2.4, in *HwC*). I suggest relabelling the strategies as 'choose one's favorite event' and 'choose one's disfavored event'. This renders the payoff matrix *symmetric* (26, Table 2.3, in *HwC*). The Kantian equilibria for these two variants *differ.* Or, to say it more precisely, the simple Kantian equilibrium in the game of Table 2.3 is 'he goes to the boxing match and she goes to the dance recital', whereas there is no simple Kantian equilibrium in the game of Table 2.4.[7]

In most economic games (which is to say, the main topic of *HwC*) there is also an issue of how to name strategies, although Sher does not point this out. If we are all fishers on a lake, doing essentially the same kind of fishing activity, we can take a fisher's contribution to be the efficiency units of fishing labor she supplies, or the hours of labor he supplies. In the former case, we then say all fishers take the same deviation from a

[6] Sher's Proposition 3 (54) is irrelevant. The game he proposes in his equation (10) is a non-monotone game.

[7] Sher's Proposition 5 (57) beats a dead horse. Nothing is learned from it that is not visible in the discussion of the Battle of the Sexes game.

given contribution profile if they each increase or decrease their efficiency units of labor by the same fraction. One way of saying this is that each fisher contemplates not increasing her fishing time by, say, 4 hours, but rather by fishing long enough to bring in another 100 pounds of fish. The Kantian equilibria will differ in these two variants. As I show, we must measure contributions in efficiency units of labor to demonstrate that the Kantian equilibrium is efficient. Measuring contributions in labor *time* will not work.

This is what Sher means by strategic inequivalence. For Nash optimization, it doesn't matter how we measure the fishers' contributions. Sher sees this as a defect of the Kantian protocol—as he says, the Kantian protocol requires some *additional information* compared to the Nash protocol, namely, how to label or measure the strategies. I see this, however, as coming with the territory, because, I believe cooperation requires that we find the *right kind of symmetry* in describing the game.

Let me return to the tragic example given by Martin Niemöller, who wrote of the Nazi strategy for picking off separate groups, while he was in a concentration camp:

> First they came for the Socialists, and I did not speak out—Because I was not a Socialist…. Then they came for the Jews, and I did not speak out—Because I was not a Jew. Then they came for me—and there was no one left to speak for me. (6)

In terms of my theory of cooperation, the failure Niemöller points to is that those persecuted by the Nazis did not *find the symmetry* in their plight: they were misled by superficial differences—being Socialists, or Jews, or Roma, or homosexual. Evil actors, like the Nazis, elevate this strategy to a principle, called 'Divide and Conquer'. Look for the superficial differences among the people you wish to oppress, emphasize those, for they inhibit the realization among those who are your target that they are 'all in the same boat'. For social movements to succeed in ending the oppression of the many by the enemy, they must emphasize the symmetry in their situations, and not be misled by superficial differences among them.

I believe that cooperation is easier to achieve than altruism. Finding the symmetry in our situations is easier than learning to care about others whom we may not even know. Not being a biologist, I cannot claim there is an evolutionary basis for cooperation among humans, while altruism has a more limited ambit. But I would not be surprised if this were so (see, for example, the work of the evolutionary psychologist Michael Tomasello, whom I discuss extensively in *HwC*).

I come finally to Sher's 'non-technical challenge'. This is that:

> Roemer argues that Kantian equilibrium is founded in self-interest and trust. […] I [Sher] argue that Kantian equilibrium cannot have a foundation on the basis of trust and self-interest *alone*. It must be founded on some moral idea that goes beyond self-interest. (45, my italics)

But this is a mis-reading. I say clearly in *HwC* that I take preferences to be the conventional self-interested ones that are typically assumed in neoclassical game theory. The morality for me comes in *the optimization protocol* (69–70 in *HwC*). A Kantian player internalizes the externalities of his action by asking how he would feel if others changed their actions in like manner. This is where morality comes in. 'Doing the right thing' means taking the action I would like everyone to take. (Of course, as we have been discussing, some care in defining what this means is required.) Sher writes: "I view my most important point as being that a player attempting to justify Kantian equilibrium play must appeal to moral—and not just self-interested—considerations" (47). I agree, with the addendum that these considerations are not represented in preferences, but in how one optimizes—that is, in the set of counterfactuals to the status quo that one envisages.

I emphasize the difference between engaging in moral *behavior* and having moral *preferences.* My objection to behavioral economics, generally speaking, is that its practitioners represent morality as altruism in preferences—caring about the welfare of others. But behavioral economists typically use the optimization protocol of Nash. My approach is the *dual* of this one: I let preferences be conventional and self-interested, and represent morality in how players optimize.

My argument, *inter alia*, is that the Kantian approach allows a much more general theory of cooperation than the altruism approach. If we alter the optimization behavior, we get Pareto efficiency right away at equilibrium, without having to insert exotic arguments into preferences.[8]

[8] Implementation theory takes another route—by having a Center propose a game with new strategies whose Nash equilibrium will induce, according to a stated rule, an efficient allocation of fishing times. I take the Kantian approach to be more decentralized than Maskin-type implementation theory. See my discussion of Vallentyne below (123–125).

Moreover, I argue in chapter 6 of *HwC* that there is no satisfactory general rule for how we should insert the altruism into preferences to guarantee that Nash equilibria of the altered game are efficient. It's not a coincidence that the games that are studied in the experiments of behavioral economists are very simple ones, where the good (efficient or equitable) equilibrium is almost visible to the naked eye (such as public-good games, dictator games, and ultimatum games). These games often do have simple altruistic variants—for example, each player maximizes the sum of player payoffs—that deliver efficient Nash equilibria. But the method does not generalize to more complex games with heterogeneous preferences.

Near the end of his paper, Sher writes:

> One potential criticism of the argument presented in this paper is that whereas I have been criticizing Roemer for attempting to found cooperation on self-interest and trust, rather than on morality, he actually does argue that agents' reasons for doing their part in Kantian equilibrium are based on morality. If this is so, then some of my criticisms are misplaced. (76)

Indeed!

COMMENT ON PETER VALLENTYNE

Vallentyne restricts himself to the consideration of simple Kantian equilibrium, as he believes the central philosophical issues appear clearly in this concept. He characterizes my view as being that, in a situation where each player trusts sufficiently that other players will cooperate, rationality requires players to choose their strategy of the simple Kantian equilibrium. I admit that I waffle on this point. "Method Two", which I propose as the reasoning process players use in a game where they trust others to cooperate (19 in *HwC*), derives simple Kantian optimization as a rational procedure when trust exists. On the other hand, in games where more complex forms of Kantian optimization are being discussed (principally multiplicative and additive Kantian optimization), I say the morality appears in the optimization protocol. Morality, so I propose, requires a player to deviate from her strategy if and only if she would prefer a symmetric deviation (defined in a particular way) by all players.[9]

[9] That is, 'Method Two' purports to derive simple Kantian optimization as rational in certain circumstances, while I emphasize the moral character of optimization in cases with heterogeneous players.

To end the waffling, my present view is that defining an equilibrium of a game requires both a specification of preferences of each player over the set of strategy profiles, and a specification of the optimization protocol that players employ. I am emphatic that cooperation is best explained as conceptualizing optimization as some version of ‘acting in common’, while parsimony recommends using self-interested preferences. ‘Acting in common’ may be justified, as Benjamin Franklin did so, by his instruction to those potential signers of the Declaration of Independence that ‘if we do not hang together, most assuredly, we will each hang separately’. This instruction purports to argue from rationality for Kantian behavior; on the other hand, I also say that morality justifies Kantian behavior, because fairness commands us to be impartial, which I interpret as requiring us to consider symmetric deviations in a game. Another way of putting this is to say it is only fair or moral that we consider the externalities, positive or negative, associated with our strategic choices, and multiplicative and additive Kantian equilibrium provide neat ways of doing so. Thus, in particular, the tragedy of the commons and the free-rider problem dissolve in monotone games when players face the externalities imposed by their actions by asking how they would feel if others took the same actions they are contemplating.

The central mathematical difference between Nash and Kantian optimization in these more complex games with heterogeneous players is that, in Kantian optimization, all players choose a strategy profile from a *common set* of counterfactual profiles, while in Nash optimization, each player chooses a strategy from *different sets* of counterfactual profiles—namely each considers the set of counterfactuals in which *only he* deviates from the status quo profile. The Nash protocol models ‘going it alone’, while the Kantian protocols model ‘acting together’ or cooperating.

In more recent work, I have argued that Nash optimization models the behavioral ethos of capitalism, which is individualism (going it alone), while Kantian optimization is the behavioral ethos of socialism, or cooperating.[10] I propose that each economic system is characterized by three pillars: a set of *institutions*, including property relations, markets, and central planning, to name several; a *behavioral ethos* that specifies how agents make decisions in economic problems; and a *distributive ethic* that specifies a theory of distributive justice that justifies the system. Economic models of socialism heretofore, although they have paid lip service

[10] See Roemer (2020). See also the interview with John E. Roemer in this issue of the journal, particularly, section IV (163–168).

to cooperation, have failed to model it.[11] The behavioral ethos of capitalism is well-modeled by Nash optimization, which captures nicely the protocol of 'going it alone'. Socialism, however, is well-modeled by Kantian optimization: it is a form of behavior that models precisely cooperation.

Like Itai Sher, Peter Vallentyne also criticizes Kantian equilibrium as being dependent upon how we name the strategies players have. I discussed my disagreement with this criticism in my comment on Sher, and have nothing worthwhile to add here.

Finally, I will discuss Vallentyne's interesting proposal in section III.I of his paper that we can get Pareto efficiency and a high degree of equality by what he calls "ordinal cooperation" (95). This means the following. Let there be n, a finite number, of players. Consider all the strategy profiles in a game (the game must have a finite number of these for this proposal to be defined, so the strategy space is finite). Each player, of course, can order all the strategy profiles according to his preferences (payoff function). Thus, each player can rank the set of profiles, since there are a finite number of them (don't worry about indifference). Associated with a particular strategy profile p is therefore an n-vector of ranks $r(p) = (r_1(p), r_2(p), \dots, r_n(p))$, where $r_i(p)$ is the rank of profile p in person i's ranking of all profiles. Now, Vallentyne proposes to say that one strategy profile p is 'at least as good as' another profile q if the rank vector $r(p)$ leximin-dominates the rank vector $r(q)$. Let's write in this case $p \gtrsim_{lex} q$.

This is a social ordering of the set of profiles. It is, indeed, an ordering (unlike the majority non-order). And the maximal elements according to this order are Pareto efficient, as Vallentyne correctly points out (95). Furthermore, as he also points out (94–95), this rule side-steps the 'problem' of Kantian equilibrium, that the equilibrium depends on how we name the strategies. Indeed, this ordering requires ordinal information only on individual preferences: it makes no mention at all of utility functions! That's a nice property.

(Indeed, is not this social ordering a counterexample to Arrow's Impossibility Theorem? The answer is no, because it fails to satisfy Arrow's

[11] I include my own book, *A Future for Socialism* (1994), as an instance of ignoring the cooperative behavioral ethos. At that time, I believed that cooperation under socialism was represented by public ownership of firms (property relations) and a distributive ethic of equality of opportunity, a view I now consider incomplete, because it fails to mention the third pillar. I now say cooperation must be defined as an explicit kind of *behavior,* one that is different from economic behavior under capitalism. The idea of requiring a behavioral ethos as part of the definition of an economic system is due to G. A. Cohen (2009).

axiom ‘Independence of Irrelevant Alternatives’. I do not object to this, however, which is why this paragraph is parenthetical.)

What’s the problem with this proposal? It’s that it is a *central planner’s* proposal. The central planner examines the set of profiles and chooses one that is maximal according to the social ordering $\gtrsim_{lex}$. Vallentyne proposes no decentralization procedure (game) to implement the social choice.

Well, we can perhaps rectify this problem. Let’s consider an altered game. Players have their standard self-interested payoff functions, but we now endow them with ‘meta-preferences’: each player’s meta-preference order is the order $\gtrsim_{lex}$ over strategy profiles. We can consider these to be altruistic preferences, or perhaps more aptly, *fairness preferences*, or even an ordinal version of *Rawlsian preferences*. Now consider the game where each player’s preferences are $\gtrsim_{lex}$, and consider the Nash equilibrium of the game in which each player proposes a strategy from the strategy space of the original game. Take the original game, for example, to be the Prisoner’s Dilemma. What is the best reply of Player 1 to a given strategy profile according to her meta-preferences? She chooses (in the set of pure strategies) a deviation (from *Confess* to *Silence*, or from *Silence* to *Confess*) if and only if that choice gives a new strategy profile that dominates the original one according to the preference order $\gtrsim_{lex}$.

One Nash equilibrium of this game is (*Silence*, *Silence*), which dominates the other three strategy profiles according to $\gtrsim_{lex}$. Thus, neither player deviates from (*Silence*, *Silence*), which is therefore a Nash equilibrium of the new game.

This procedure decentralizes the implementation of the ordinal cooperation. This is actually an example of my criticism of what I say is the main move of behavioral economists to get nice outcomes in games. It is to consider ‘exotic arguments’ in preferences, but to maintain Nash optimization as the behavioral protocol. The exotic arguments, in this case, are the ranks that other players assign to the strategy profile.

Indeed, we can ‘implement’ *any social welfare function* with this procedure. Simply endow each player with the social preference order and have him play his strategy to achieve the highest ranked strategy profile, according to this order, that he can induce by his behavior alone. Trivially, any strategy profile that maximizes the social ordering is a Nash equilibrium of this game! The converse, however, is false. In particular, (*Confess*, *Confess*) is also a Nash equilibrium of this game, because if either player changes his strategy to *Silence*, the new profile (*Silence*, *Confess*) is dominated by (*Confess*, *Confess*) according to $\gtrsim_{lex}$. So, there is a *bad* Nash equilibrium as well as a *good* one of the altered game. If we pursued this discussion further, and tried to construct a game that would implement $\gtrsim_{lex}$ in the more demanding sense that *every* Nash equilibrium of the game is a maximal element of $\gtrsim_{lex}$, we would be led down the path to Maskin-type implementation theory.

As I said earlier, I do not see a natural way of extending $\gtrsim_{lex}$ to games with a continuum of strategy spaces. We encounter such games as soon as we introduce mixed strategies in 2 × 2 matrix games, or more generally when we consider economies. My principal point, however, is that Vallentyne's proposal abandons the concern for decentralization.

COMMENT ON JEAN-FRANÇOIS LASLIER

Laslier points out my lack of sophistication as an evolutionary game theorist. I only wish I had discussed this chapter with him before I published it. I am grateful for his presenting a better model of the problem. Unfortunately, his exposition is a bit too condensed for me, and I cannot comment on it.

I will, however, comment more generally on the evolutionary approach to cooperation. There is no doubt that over time, our species has increased its degree of cooperation immensely. In pre-historic times, the extent of cooperation was limited to one's small band of at most several hundred souls. Today, we have cooperation within nations with over a billion souls. A most significant form of that cooperation is taxation, which now, in the most advanced countries, collects approximately one-half of the national product, and re-allocates it for the public good. This is a twentieth-century innovation. Karl Marx's theory was that cooperation increases as history progresses because of technological development: economic structures develop only so long as they encourage the further development of the productive forces, and are then 'burst asunder' when they become fetters on that development. To link this to the evolution of cooperation one would have to theorize why more advanced productive forces necessarily require more cooperation to operate. Marx gave some examples (feudalism gives way to capitalism which gives way to socialism), but the necessary link to cooperation is, I think, not satisfactorily explained.

Although I agree with Tomasello and others that our species evolved, through selective adaptation, as a cooperative species—in contrast even to other great apes—that cannot explain the rather short time span (say, 10,000 years) in which the extent of human cooperation has increased so dramatically. Something like the kind of mechanism that Marx offers is necessary to explain such rapid social evolution.

REFERENCES

Ahlquist, John S., and Margaret Levi. 2013. *In the Interest of Others: Organizations and Social Activism*. Princeton, NJ: Princeton University Press.

Cohen, Gerald A. 2009. *Why Not Socialism?* Princeton, NJ: Princeton University Press.

Roemer, John E. 1994. *A Future for Socialism*. London: Verso.

Roemer, John E. 2019. *How We Cooperate: A Theory of Kantian Optimization*. New Haven, CT: Yale University Press.

Roemer, John E. 2020. "What is Socialism Today? Conceptions of a Cooperative Economy." Cowles Foundation Discussion Paper No. 2220. Yale University, New Haven, CT.

John E. Roemer is the Elizabeth S. and A. Varick Stout Professor of Political Science and Economics at Yale University. He is a Fellow of the Econometric Society, and has been a Fellow of the Guggenheim Foundation, and the Russell Sage Foundation. His research concerns political economy, and distributive justice. He is the author of numerous books, including, among others, *How We Cooperate* (2019), *Sustainability for a Warming Planet* (2015), *Democracy, Education, and Equality* (2006), *Political Competition: Theory and Applications* (2006), and *Equality of Opportunity* (1998).
Contact e-mail: <john.roemer@yale.edu>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 127–176.
https://doi.org/10.23941/ejpe.v13i2.530

What Egalitarianism Requires: An Interview with John E. Roemer

JOHN E. ROEMER (Washington, 1945) is the Elizabeth S. and A. Varick Stout Professor of Political Science and Economics at Yale University, where he has taught since 2000. Before joining Yale, he had taught at the University of California, Davis, since 1974. He is also a Fellow of the Econometric Society, and has been a Fellow of the Guggenheim Foundation, and the Russell Sage Foundation. Roemer completed his undergraduate studies in mathematics at Harvard in 1966, and his graduate studies in economics at the University of California, Berkeley, in 1974.

Roemer's work spans the domains of economics, philosophy, and political science, and, most often, applies the tools of general equilibrium and game theory to problems of political economy and distributive justice—problems often stemming from the discussions among political philosophers in the second half of the twentieth century. Roemer is one of the founders of the Anglo-Saxon tradition of analytical Marxism, particularly in economics, and a member of the September Group—together with Gerald A. Cohen and Jon Elster, among others—since its beginnings in the early 1980s. Roemer is most known for his pioneering work on various types of exploitation, including capitalist and Marxian exploitation (for example, *A General Theory of Exploitation and Class*, 1982a), for his extensive writings on Marxian economics and philosophy (for example, *Analytical Foundations of Marxian Economic Theory*, 1981, and *Free to Lose: An Introduction to Marxist Economic Philosophy*, 1988a), for his numerous writings on socialism (for example, *A Future for Socialism*, 1994b), and for his work on the concept and measurement of equality of opportunity (for example, *Equality of Opportunity*, 1998a).

The *Erasmus Journal for Philosophy and Economics* (*EJPE*) interviewed Roemer on the occasion of his latest book, *How We Cooperate: A Theory of Kantian Optimization* (2019a), to which the *EJPE* is devoting the present special issue. The interview covers Roemer's intellectual biography

EJPE'S NOTE: This interview was conducted by Akshath Jitendranath and Marina Uzunova, editors at the *EJPE*. They thank Nicholas Vrousalis for generously contributing interview questions, and Ben Ferguson and Roberto Veneziani for advice while preparing the interview.

Most footnotes throughout the interview are editorial annotations directing readers to the relevant literature discussed in the text. Footnotes by the author, John E. Roemer, are indicated by the phrase 'Author's Note'.

(section I); his extensive writings on exploitation, egalitarianism (section II), socialism, bargaining, and justice (section III); his latest work on Kantian optimization, his vision for the future of socialism (section IV); and, finally, his methodological commitments and the value of interdisciplinarity (section V).

I. INTELLECTUAL BIOGRAPHY

EJPE: Professor Roemer, during your childhood years, your family lived in Switzerland and Canada, before your parents—Ruth and Milton Roemer—returned to Cornell and then UCLA. Can you tell us a bit about the people and events that were formative for you during the time before you entered university?

JOHN E. ROEMER: In October 1948, my father received a letter from the Board of Inquiry on Employee Loyalty of the US Federal Security Agency (FSA), interrogating him about his association with people and organizations associated with the US Communist Party. He was at the time still a member of the US Public Health Service, a unit of the Army, that he had joined, as a physician, during World War II. This letter was the first in an intensive correspondence between the FSA and my father about his professional and political activities, in which my father took the position that, since he had joined the Public Health Service, he had had no contact with the Communist Party. In April 1949, he received a letter from the FSA clearing him of suspicion of disloyalty to the United States. However, the Agency re-opened his case in 1950, at which time he was an assistant professor at Yale University, on leave or loan from the Public Health Service. At this point, he was advised by his lawyer that he would probably be found to be disloyal, and would be discharged from the Public Health Service and fired from Yale, and it would be prudent for him to leave the country. He received an offer from the World Health Organization (WHO), and our family moved to Geneva later that year.

About a year later, the US State Department retracted my parents' passports, because they were considered to be disloyal citizens, and the WHO was obliged to fire him, although they gave him a year's grace in order to find another job. At this time, there was a social-democratic government in Saskatchewan, Canada, led by a Scottish socialist named Tommy Douglas. The provincial government offered him a job to work on designing a provincial health insurance system, which eventually became the first single-payer health insurance system in North America. We lived in Regina, Saskatchewan for three years, until the peak of McCarthyism in the United States had passed. At that point my father accepted an offer from Cornell University, and the family moved to Ithaca, New York.

I relate this history because it was similar to the political persecution that many left-wing Americans were subjected to in the early 1950s. Of course, my parents' troubles were at the center of the discussions in the household. Certainly, the most important influence on me until I left home for university was my parents. The culture of the household was deeply political. Even though my parents were not members of the US Communist Party since some time in the 1940s, they remained staunch socialists, and supporters of the Soviet Union, even after the revelations by Khrushchev about Stalin's crimes at the twentieth Communist Party of the Soviet Union (CPSU) party congress in 1956. My father would defend Stalin until sometime in the 1990s. My mother was somewhat less political, but she was also pro-Soviet until late in her life, as were the two grandparents whom I knew.

So, since my earliest memories, I have been a socialist. I remember as a child thinking the good guys were the workers, the Democrats, and the Brooklyn Dodgers, and the bad guys were the bosses, the Republicans, and the New York Yankees. (There was pretty strong class allegiance to the Dodgers and Yankees as I have described.)

I also had good friends in high school, both girls and boys. Only one of these friendships was largely based on a political bond—her parents were also close to the Communist Party. The bond in the other friendships was based on love of mathematics. One of these high school buddies was Roger Howe, who remains a friend until today, and a collaborator and teacher in my professional work. Roger is a superb mathematician, who has retired after many years teaching at Yale University where, by some coincidence, I have also ended up.

You mention your friendship and collaboration with the mathematician Roger Howe. The two of you wrote a 1981 paper together, "Rawlsian Justice as the Core of a Game", which was one of the first attempts at applying game theory so rigorously to questions of justice.[1] This was

[1] Howe and Roemer (1981) model the original position as a game with specified withdrawal payoffs and argue that the difference principle is in the core of a game in which no coalition will withdraw after the veil of ignorance is lifted, unless it can guarantee every one of its members a better payoff in a new lottery. Apart from being one of the first attempts at applying game theory to questions of justice, this result is of particular interest for two reasons. First, it shows that an assumption about individuals being

before Ken Binmore's two-volume* Game Theory and the Social Contract *and your own* Theories of Distributive Justice*, both published in the 1990s.*[2] *How did the joint work on this paper come about? And why have you not published any other manuscripts together?

I don't recall how Roger and I came to collaborate on the Rawls paper: I must have initiated discussing it with him. Although that is the only paper on which we collaborated officially, Roger has made key contributions to the mathematics in a number of my papers. See, for instance, the lead footnote in my article "Equality of Resources Implies Equality of Welfare" (1986a). Roger was responsible for a theorem in convex analysis that he stated and proved at my request, which was the key to the main result of that paper. You will find an acknowledgment to Roger in quite a few of my papers and books.

As you have already related to us so vividly, you come from a socialist household, and your parents were active health services researchers. What kind of conversations did the family have when you were all together?

My parents were avid hosts: my mother organized several large dinner parties a month, while she also raised her children and was a full-time faculty member at UCLA, from 1962 on. There were always academic friends and former students passing through Los Angeles, and each visitor provided an excuse for a dinner party. At the cocktail hour before dinner, my father would invariably begin a political discussion, which often continued through dinner. These events exposed me not only to a political worldview, but introduced me to left-wing public-health professionals from around the world. When later I began travelling abroad as a young adult, I would eagerly look up my parents' friends in the cities that I visited, and would invariably be shown a good time, with lessons about the political history of the country since the war.

What kind of books and authors were you reading as a child? Were you also reading philosophical works at that time?

moved by a "special psychology" that makes them "peculiarly averse to uncertainty" (Rawls 2001, 107)—rather than rational self-interest—was implicit in the argument for the difference principle. This is because Howe and Roemer show that a risk-neutral game has no core, but the difference principle is in the core of an extremely risk-averse game. Second, the result is also of import for the question of stability, which was a central concern for Rawls. For Rawls' own discussion of the implications of Howe and Roemer's results for his theory, see Rawls (2001, 109–110).

[2] See Binmore (1994, 1998), and Roemer (1996), respectively.

I'm afraid I did not read much as a child: the great literature that I read was only that which was assigned in high school. My extra-curricular activity involved either math or music.

You were interested in socialism from an early age and yet you have said in the past that up to your first job at UC Davis, you had never read Marxian economics.[3] When did you start reading Marxian economics—and theory more broadly—and which books and authors were formative for you in that respect?

Although my identity was socialist from early on, I did not become politically active until graduate school. I graduated from Harvard in 1966, having taken as many math courses as was permitted. I took one freshman philosophy course, one history course, one economics course, and one music theory course. I think that was the sum of my general education. I also took only one physics course: unlike most good mathematicians, I did not have an aptitude for physics. I have since regretted the narrow focus on mathematics that I had during those years.

I enrolled in the PhD program in mathematics at the University of California, Berkeley, with Roger Howe. We had also gone together to Harvard, where we both majored in math. I chose Berkeley not only for its great math department, but also because the left-wing student movement was so active there. Arriving in Berkeley, I finally became deeply involved in left-wing politics. At this point, I started reading Marx, although largely his political pamphlets, as well as those of Lenin and Mao. I do not think I read *Capital* until 1974. There were no courses in Marxian economics at Berkeley at that time, and if there was one at Harvard, which is possible, I was not interested in it when I was an undergraduate.

You have mentioned your interest in music a couple of times and that is interesting because literature is normally much more dominant in the discussions taking place at the intersection of philosophy and economics. Can you tell us a bit more about the kind of music that you were and are listening to? Also, did you ever think about the relation between mathematics and music, and was that a source of inspiration in some way?

I took clarinet lessons as a child, and participation in the concert band was an important activity for me in high school. There was a piano in our house, and I started picking out jazz as a pre-teen. My musical heroes

[3] See Roemer's interview with Maya Adereth and Jerome Hodges (2019).

were Duke Ellington, Oscar Peterson, and Erroll Garner. By the end of high school, I played jazz and blues piano, by ear, quite well. I never took piano lessons; my style was heavily influenced by Erroll Garner. He is much easier to imitate than Oscar Peterson—Garner had no classical education in piano, whereas Peterson did, and so had much more accomplished technique, which I could not hope to copy. I still play occasionally, although my musical ideas have not developed much since the age of twenty. It is said there is a link between math and music, although I don't see it in my own case. My musical intuition seems quite different from my mathematical intuition. The only attribute both intuitions share, it seems, is their requiring thousands of hours of doodling around to develop.

As you said, you obtained your undergraduate degree in mathematics from Harvard in 1966. What did you write your thesis on, and did you use it in your later studies and work?
My senior thesis at Harvard was on abelian groups. I never worked in that area again. The mathematics that I have used is applied analysis.

You then moved to Berkeley for your graduate studies in mathematics but quickly changed your major to economics. You have explained this change with your political activism around the anti-Vietnam War movement at the time. This was also the time, particularly in the tumult of 1968, when you got arrested together with a group of students who occupied the university administration building.[4] Have your views on political activism changed? Did you, and do you still, remain politically active after 1968, and if yes—how?
When I was suspended from Berkeley in 1968, I lost my draft deferment. I took a job teaching math in a virtually all-black junior high school in San Francisco; with this, I received another deferment from the draft, for teaching in an inner-city school which was considered to be a kind of national service. I was politically active in a left-wing caucus in the teachers' union. In 1973, I was re-admitted to Berkeley, and wrote my PhD dissertation in economics.

My views on activism have not changed, although my left-wing activity has been largely restricted to my writing since 1976, as well as to participating in the occasional mass demonstration.

[4] See Adereth and Hodges (2019) for more details on this episode.

The title of your PhD dissertation at Berkeley was "U.S.-Japanese Competition in International Markets: A Study of the Trade-Investment Cycle in Modern Capitalism". Who was your supervisor, and how did you decide to work on this topic?
I chose the topic of US-Japanese competition because the gauchiste party to which Natasha and I belonged in Berkeley thought that, with the end of the Vietnam War, the major conflict in the world would be inter-imperialist rivalry between the US and Japan. Unfortunately, I knew very little about international trade and finance, and my dissertation was journalistic rather than academic. My adviser was a left-wing economic historian, Richard Roehl. I received my degree, despite the rather unsatisfactory dissertation, due to the support of Benjamin Ward, an iconoclastic professor at Berkeley, who advocated for me because I was not treading the usual professional path. Ward was the author of *The Ideal Worlds of Economics: Liberal, Radical, and Conservative Economic World Views* (1979).

What were the kind of topics that your fellow PhD students at Berkley were working on at the time? Did any of your colleagues have a particular influence on your move away from international trade?
I did not have much contact with my fellow students at Berkeley, because when I was taking classes, in 1966–1968, I was spending all my extra time in campus political work. And then there was a five-year hiatus before I returned to Berkeley to write the dissertation. I took a job in 1974 as an assistant professor at the University of California at Davis. In the summer of 1975, I read *Marx's Economics* by Michio Morishima (1973), a Japanese mathematical Marxist economist. I was excited by this book, for Morishima was using the tools I had learned in mathematical economics to study Marxist questions: exploitation, the labor theory of value, the transformation problem. Two micro-economic theorists on the Davis faculty, Ross Starr and Richard Cornwall, suggested I teach a course on Morishima's book. This began my work in mathematical Marxian economics, the culmination of which was my book *A General Theory of Exploitation and Class* (1982a). I am grateful to my Davis colleagues who set me on the path of Marxian economics.

We have reached the end of the 1970s and the start of the 1980s—an important point in time that saw the formation of the 'September Group'. The Group was formed by Jon Elster and Gerald A. Cohen in 1979 and you joined it the following year. Can you tell us more about

the organisation of the meetings and the general environment? Did you follow a typical seminar format, with a presentation followed by a discussion, or did you pre-circulate the relevant texts and devote the meetings only to discussions?

While I was working on the book I just referred to, I read G. A. Cohen's *Karl Marx's Theory of History: A Defence* ([1978] 2001) and Jon Elster's *Logic and Society: Contradictions and Possible Worlds* (1978). I learned I was not alone: here were two young academic Marxists, using the latest tools in analytical philosophy and social science to study Marxian questions. I sent a few chapters of the draft of my *General Theory* to Cohen, who replied with a lengthy letter. He invited me to the next meeting of a group that he and Elster had convened in Paris the year before, of similarly inclined young Marxist academics (all male). The first year I attended the September Group was 1980 or 1981. The current name of the group was adopted later: in the early years, we referred to ourselves as the NBSMG (No-BullShit Marxist Group).

The annual meetings lasted two or three days, with ten to fifteen in attendance. We followed the usual format of paper presentations, all read before the meeting, with discussants.

Was the name 'No-Bullshit Marxist Group' proposed by Cohen? He has a colourful section in the 2000 introduction to his **Karl Marx's Theory of History** ***where he clarifies what he calls his practice of "non-bullshit Marxism" (Cohen [1978] 2001, xxv–xxviii).***

I don't recall who came up with the handle. Both Jerry and I were pretty profane, and it could have been either of us. I don't think it was Jon Elster—it was not his style. There was, however, a slight difference: Jerry always said *non*-bullshit Marxism and I said *no*-bullshit—the latter must be more of an Americanism.

You mentioned the importance of Michio Morishima's **Marx's Economics** ***to your initiation in Marxian economics. At the time when the September Group was formed, Morishima was teaching at the London School of Economics as the Sir John Hicks Professor—a position he held from 1970 until 1989. Did you ever meet him? And why was he never a member of the Group?***

None of us knew Morishima. Furthermore, he did not have any obvious leftist sympathies. Much of the work that Morishima made famous was developed by other Japanese Marxist economists such as Nobuo Okishio

(1927–2003). Unfortunately, I never met Okishio, who, I believe, was more the pioneer of mathematical Marxian economics in Japan than Morishima.

Can you tell us more about the kind of topics—and papers—that were discussed at the beginning? For example, Cohen's own account of the pre-history of the Group says that the first two meetings (in 1979 and 1980) were on exploitation.[5] What kind of work on exploitation was being discussed at the time—normative, conceptual? How did the topics change over the years?

In 1986, I edited the book *Analytical Marxism,* which published a dozen or so papers from the September Group.[6] The topics were quite broad-ranging. Members included philosophers, economists, sociologists, historians, and political scientists. The common task was to re-state Marxian questions in a modern way, and to study them using the tools of analytical social science and philosophy. The school of 'analytical Marxism' was quite influential in the 1980s: it was attacked from the left by traditional Marxists, who believed that using these 'bourgeois' tools of analysis would surely infect our conclusions. In reply, we called these critics bible-thumpers. The preface of Cohen's 1978 book on historical materialism contains a lovely comment about bible-thumping (though not using that terminology).[7]

In a recent interview, you said that the "group continues to meet every year, though most of us no longer identify as Marxists" (Adereth and Hodges 2019). Why is that? Why did you cease to identify as a Marxist?

Some people left the group in the early 1990s because they felt we had accomplished what we had set out to do—to find what part of Marxism stood the stress test of analysis with modern tools. Others, like myself, still valued the meetings, although the topics tended to diverge quite a bit, as the members became older. I tend not to call myself a Marxist anymore because I do not credit many of the ideas that Marx believed were at the center of his view: the labor theory of value, the falling rate of profit, and the claim that dialectical materialism is a special kind of logic.

In this period, from 1980 on, G. A. Cohen and Jon Elster were my closest intellectual comrades outside of economics. Cohen died suddenly at

[5] See Cohen ([1978] 2001, xviii–xix).

[6] See Roemer (1986c).

[7] See Cohen ([1978] 2001, ix).

age 68 in 2009. To this day, I remain close to Jon Elster, both intellectually and personally.

In 2000, you became the Elizabeth S. and A. Varick Stout Professor of Political Science and Economics at Yale—a position you hold to this day. Can you tell us something more about the namesakes of the professorship, and what made you decide to move to, and stay at, Yale?
I'm afraid I don't know anything about the Stouts, who endowed the chair I hold. I have never been asked to pursue an intellectual agenda associated with the chair—it comes with no strings attached. My wife Natasha and I decided to try to move to New York from California, after 26 years at UC Davis, because we loved the city after spending a year here in 1998–1999. Luckily, I was offered the Yale position, so that this became a reality. Although I had a wonderful academic environment at Davis, Yale is something special, and we've had no thought of moving again.

II. Exploitation and Egalitarianism

Your views on exploitation have changed considerably over the years. Let us start by asking: when and why did you become interested in exploitation?
As I said, I considered myself a Marxist from early adolescence. However, I never took any left-wing, let alone Marxist, courses as an undergraduate. I was certainly familiar with Marx's theory that exploitation of labor was the key to understanding capitalism. As I related, in the summer of 1975, after my first year of teaching at Davis, I read Morishima's book *Marx's Economics*, published in 1973, in order to prepare a seminar I was planning to teach on the topic. This was my first exposure to mathematical Marxism, and I was enthusiastic. Morishima (and again, I should say, the school of Japanese mathematical Marxists) provided rigorous definitions of embodied labor time and exploitation, and proved theorems. The main theorem Morishima called the Fundamental Marxian Theorem, which states that, in a market economy, profits of firms are positive if and only if workers are exploited.[8] This marked the beginning of my professional interest in exploitation.

[8] The Fundamental Marxian Theorem is credited as the Morishima–Seton–Okishio theorem after the contributions of Michio Morishima and Frances Seton (1961), and Nobuo Okishio (1963). See Morishima's discussion of the theorem in Part II of his book (1973, 53ff.).

You are known for developing two conceptions of exploitation—one based on the 'surplus value' approach, and the other based on the 'property relations' approach. Can you explain intuitively what the differences between these two forms of exploitation are and why you decided to abandoned the 'surplus value' conception in favour of the 'property relations' conception?

The surplus-value definition says this: a worker is exploited if the embodied labor time in the goods that he can purchase with his/her wage income is less than the labor he/she expended in production to earn those wages. More generally, a producer's income can come from three sources: wage labor, profits, or work done by a producer on his own capital. A producer is exploited if the amount of consumer goods she can purchase with her income embodies less labor than she expended in production, whether as a wage worker, or a petty-bourgeois, working up her own capital. A producer is an exploiter if his income purchases goods embodying more labor than he expended in production.

In my models, individuals (producers) choose, constrained by their wealth, whether to sell their labor power, to expend their labor on their own capital, or to hire others to work on their capital. The combination of these three activities determines a producer's *class position.* What I proved was that each producer would end up either being exploited, or being an exploiter, or being neither exploited nor exploiting, and one's exploitation status (defined by the surplus-value definition), as determined by preferences and the value of one's capital, corresponded in a clear way to one's class position. Proletarians, who owned no capital, had no choice but to sell their labor power to others. If a person has a lot of capital, she can optimize by not working at all and only hiring others. It turns out there are five *class positions*, which may be associated with being a landlord, a rich peasant, a middle peasant, a semi-agricultural proletarian, or a landless laborer. The *class-exploitation correspondence principle* says that if there are positive profits in a capitalist economy, then any producer who must sell labor to solve his optimization problem is exploited and any producer who must hire labor to optimize is an exploiter. This is a theorem: one proves the relationship between class membership and exploitation status as a consequence of the definitions.[9] We prove from axioms that the classical Marxist relationship between working for others and being exploited must hold—it is not simply a description of

[9] See Roemer (1982a, 78–82, 129–132) for statements, proofs, and explanations of this result in economies with and without capital accumulation.

reality or a definition. This provides *microfoundations of class membership* from the optimization behavior of individuals.

But is this exploitation immoral or unethical? We cannot say, until we know *how it came to be* that some people begin owning capital and others do not. Marx established, in his researches in the British Museum, that the 'primitive accumulation of capital' did not emerge through honest work, but through plunder, enclosure of the peasant commons, regal gifts of land to feudal lords, and so on. Thereby Marx established—assuming his history is correct—that ownership of capital is morally tainted, and *that's* what makes exploitation a bad thing: exploitation of some by others is a manifestation of differential ownership of capital whose genesis is immoral.

Now this history of primitive accumulation suggests that we contemplate an alternative distribution of land and capital, an equal one. We can propose another definition of exploitation that does not mention surplus labor or value at all. We can ask of the equilibrium in a capitalist economy: suppose the workers were to withdraw from the economy, taking with them *their per capita share* of the capital stock. Would they be better off or worse off than in the capitalist equilibrium? More generally, we can define a group or coalition of producers as *exploited* if, were they to withdraw from the present situation with their per capita share of the capital stock, their lot would improve, and a group or coalition is exploiting if, were they to withdraw with their per capita share of the capital stock, they would be worse off. We don't refer to labor embodied in goods at all. This is the *property-relations* definition of exploitation.

It turns out that one can show that (under certain conditions) the property-relations definition and the surplus-value definition are equivalent in the sense that under both definitions, the group of exploited producers is the same and the group of exploiters is the same.[10] The virtue of the property-relations approach is that it builds in the ethical condemnation of capitalism: for conceptualizing the counterfactual to the capitalist equilibrium as an alternative where each coalition gets to keep its per capita share of the capital stock (and of course its own labor power) is salient because, absent the plunder of primitive accumulation (according to Marx), the equal-per-capita distribution of capital is what justice would require. Or at least this is one obvious alternative to capitalism with unequal ownership of the capital stock (means of production).

[10] See Roemer (1982a, 194–237, and particularly, for a summary, 233–237).

These are questions that we will come back to in more detail later, but just to clarify: at least in the model of an economy with capital accumulation, it seems that assuming that the equal-per-capita distribution is the right counterfactual distribution also assumes that justice requires the* complete *elimination of material bequests—though non-material inheritance might not be entirely problematic, as you have argued in relation to intergenerational mobility (Roemer 2004). Is this a view on the injustice of (material) wealth inheritance that you generally subscribe to?

I believe that all young adults should begin their productive years with the same amount of wealth. This implies that the inheritance of wealth, and *in vivos* transfers to the young, must be sharply constrained. If the educational system has succeeded in eliminating inequality of opportunity, and people make different career choices, then differential wealth will emerge during adult lifetimes, and I believe those differences are consistent with justice, as long as there is sufficient income and wealth taxation to prevent income differences from becoming too extreme—so extreme as to threaten solidarity. As I said, Marx's condemnation of the distribution of capital was based on the history of 'primitive accumulation' that he presented. If wealth accumulation is a result of freely chosen labor with equal-opportunity background conditions, I do not believe modest wealth differences are unjust.

Allow us to briefly go back to the 'surplus value' conception of exploitation. Your formal definition of this form of exploitation is based on transferable-utility (TU) cooperative games and in the TU framework, as you say, "there are no considerations of incentives and strategy* within *the [exploiting and exploited] coalition" (Roemer 1994a, 19), where coalitions here stand for the relevant (exploiting and exploited) classes. Was this an intentional or a pragmatic choice? Did you consider defining this form of exploitation in a non-transferable utility (NTU) framework, which would have allowed modeling not just* inter*-class but also* intra*-class conflict? More generally as well, should socialists be interested in* intra*-class conflict?

You are getting technical here, talking about TU and NTU games. In fact, one can show there is intra-class conflict with my approach. It can be that if the coalition of all the workers *W* (in the present capitalist equilibrium) were to withdraw with its per capita share of the capital, its members would be better off, but if the coalition of highly skilled workers, call it *S*, which is a proper subset of *W*, were to withdraw it would be even better off, and the remaining workers (in *W* but not in *S*) would be *worse* off if they were to withdraw. This shows there may be a conflict between skilled workers and unskilled workers—the latter may need the former to be better off than under capitalism.

Of course, we should be interested in intra-class conflict, to the extent that it is a real phenomenon.

We will come back to intra-class conflict towards the end of this section, but for now let us return to exploitation. There have been roughly two strands of thinking about exploitation: (1) the non-moralised approach, which understands exploitation **positively** ***or*** **descriptively*****—here, for example, we have Allen Wood and his interpretation of Marxian exploitation—and (2) the moralised approach, which understands exploitation*** **normatively*****—here, we have, Hillel Steiner, Jon Elster, and Robert Goodin, among others. Your first conception of exploitation based on the 'surplus value' approach has a domination condition and thus seems to fall in the moralised camp. However, your second conception based on the 'property relations' approach is purely descriptive. Have your views on the moralised versus non-moralised nature of exploitation changed? And if yes, why?***

There are several problems with what you call the non-moralized approach. The first is that it turns out *any* input (say, coal or energy) can be shown to be 'exploited' in a capitalist economy with positive profits. That is, we can define the energy value of a commodity as the amount of energy embodied in producing it and all the inputs needed for its production. We must be able to define the energy embodied in a unit of labor power as well: this is the amount of energy the producer has to consume in order to reproduce her labor power—heating, gasoline in one's car to get to work, and so on. Then one can show that profits are positive if and only if energy is exploited, in the sense that a unit of energy has embodied in it less than one unit of energy. (Just as: labor is exploited if the production of one unit of labor power requires consuming goods that embody less than one unit of labor.)

Well, if this is the case, then what's special about labor power? I claim it's because there is *no moral opprobrium* associated with the exploitation of energy, or of steel, or seed corn. The moral opprobrium associated with the exploitation of labor is that its source is the vastly unequal distribution of capital that came about through robbery, plunder, enclosure, etc.

So the surplus-value definition, I claim, has its appeal because we intuitively feel that the vastly unequal distribution of wealth (capital) is morally indefensible. And it's *not* indefensible because unequal wealth produces labor exploitation—that would be circular—but because *the source of unequal wealth is immoral takings.*

This raises the important question: what if unequal capital ownership comes about *morally*? An important question, which we will address later.

In "Should Marxists Be Interested in Exploitation?", you concluded that "exploitation theory is a domicile that we need no longer maintain: it has provided a home for raising a vigorous family who now must move on" (Roemer 1985, 33). And, indeed, since then, your work has noticeably strayed away from issues of exploitation. Can you explain why you reached this conclusion? In your view, is there still a place for the concept of exploitation in Marxism?

I think I have explained this. The central question that Marxists should be interested in is the ethical status of the distribution of wealth. Marx believed that socialism would expunge the immorality of capitalism by prohibiting the privatization of capital. Under socialism, capital would be owned collectively by the entire coalition of producers. As we know, Marx said hardly anything about the details of how such an economy would function; his concern was to diagnose how wealth could emerge in such a concentrated form in a mode of production in which the coercion of workers no longer existed—in the sense, that is, that serfs and slaves were coerced to work. The sleight-of-hand of capitalism was to produce a highly skewed distribution of wealth and income *even though* the direct producers were free and not induced to work by the bosses' whip. Exploitation of labor, in the surplus-value sense, is a symptom of the immorality of capitalism, but it's not the source of that immorality. The source is the set of practices that leads to a highly skewed distribution of wealth in the first place.

In a series of papers following "Should Marxists Be Interested in Exploitation?" (1985), you amend a part of your 'property relations' definition of exploitation. More precisely, in your 1985 paper you argue that gain from the labour of others, including unequal exchange of labour, is irrelevant to a charge of exploitation. But in a later response to an example by Erik Wright, you reintroduce "gain by virtue of the labour of

others" as part of the definition of exploitation.*[11] *What accounts for this vacillation and what is your present take on it?

I stopped thinking about these puzzles years ago, because I came to believe, as I've explained, that exploitation is an irrelevant tangent. There isn't much point in worrying about exactly which conception of exploitation is the best one, if the genesis of unequal wealth is what's key to understanding why capitalism is unjust and socialism might be just, if we can figure out how it should be designed.

There are plenty of issues in deciding when the distribution of wealth/income is just that can be addressed more directly without going through the detour of exploitation. The twentieth-century contribution to this inquiry begins with the political philosophy of John Rawls.

Let's turn to this inquiry, then, and the debates inaugurated by the work of John Rawls. In* Egalitarian Perspectives*, you said that when you met Gerald A. Cohen for the first time in the spring of 1981, you "began to learn from him the range of questions addressed by modern political philosophy" (Roemer 1994a, 1). You had also been reading Cohen's* Karl Marx's Theory of History: A Defence *while writing* A General Theory of Exploitation and Class *in 1979–1980. Is it fair to say that your initiation in contemporary political philosophy was through Cohen? Did the 'Rawlsian storm' of the 70's not reach you till the 80's?

That's correct. I was led to see the importance of my ignorance of philosophy as I struggled to understand why Marxian exploitation, according to

[11] Suppose that a society is divided into two coalitions, *S* (the exploited) and its complement *S'* (the exploiting). Wright's example is the following:

> Consider the case of two agents, Rich and Poor, who are initially endowed with 3 and 1 units of capital, respectively. This distribution is unfair: suppose that the fair distribution is egalitarian. Rich wants to consume prodigiously, while Poor only wants to subsist and write poetry (a good for which there is no market). Rich works up all his capital stock, but wants to consume even more than what is thereby produced, and and [*sic*] so Poor hires Rich to work up Poor's capital stock, paying Rich a wage and keeping enough of the product to enable him to subsist. According to the PR ['property relations'] definition of exploitation Rich is an exploiter and Poor is exploited. But this seems intuitively wrong because although Rich gains by virtue of being unfairly rich, he does not gain by virtue of the labor of Poor. I previously wrote that Rich did exploit Poor in this example, but I now do not think so. Therefore, I would substitute, for clause (3) [of the 'property relations' definition of exploitation: *S'* would be worse off if *S* withdrew from society with its own assets], the following: *S' gains by virtue of the labor of S*. (Roemer 1994a, 106; emphasis in the original)

See Vrousalis (forthcoming, 6–7) for a discussion of these revisions in Roemer's account of exploitation.

the surplus-value account, was unjust. I did not take Rawls' course as an undergraduate at Harvard—this was before he published *Theory of Justice,* but I am sure his course had a reputation that I did not learn about, because of my narrow focus on mathematics. Jerry Cohen introduced me to the egalitarian debate in the form of Ronald Dworkin's two 1981 articles in *Philosophy & Public Affairs.*[12]

Let us now turn to Robert Nozick. Cohen famously said that Nozick was the author who shook him from his "dogmatic socialist slumber" (1995, 4). The particular occasion for this was Nozick's Wilt Chamberlain argument and Cohen's subsequent realisation that self-ownership itself is baked into the Marxist condemnation of exploitation. You have credited Nozick as the author who exposed "Marx's false-positive error—that some instances of (Marxist) exploitation are not unjust" (2017, 264; see also 291–292). This answers the question you asked just above: "what if unequal capital ownership comes about **morally*****?" (141). Was this your own 'shaking' moment? If not, have you had such an experience?***
The first article I published giving examples of just Marxian exploitation was "Should Marxists be Interested in Exploitation?" published in 1985. I must have been writing that article in 1983. I had a few years earlier (1982) published my property-relations, game-theoretic model of exploitation in *The Economic Journal* and also in my 1982 book on exploitation and class.[13] I was working on that book in 1979–1980, so I had surely understood the problems with the surplus-value account of exploitation as a theory of injustice by then. I don't remember any 'shaking moment', but I do recall many conversations with Jerry Cohen at that time that were hugely exciting.

On Nozick, more generally, what do you think has been his broader political and cultural influence—beyond this exposure of Marx's error?
Nozick constructed a clear argument for capitalism, based upon the premise of self-ownership. Jerry took Nozick's argument seriously, because he pointed out that self-ownership was *also* assumed by Marx, when he viewed the surplus labor that capitalism transfers from workers to capitalists as an ethically illicit transfer. There was a discussion about whether Marx really argued that the transfer was ethically illicit, but Jerry

[12] See Dworkin (1981a, 1981b).
[13] See Roemer (1982b) for *The Economic Journal* model, and Roemer (1982a) for the book on exploitation and class.

and I believed that, despite Marx's occasional protests to the contrary, one cannot explain the depth of his condemnation of capitalist property relations without supposing that he viewed exploitation as *wrong.* Jerry and I argued that even if self-ownership were granted as a premise, Nozick's defense of capitalism did not work, because it assumed that the physical world (real property and resources) was, ethically speaking, owned by no-one before it was privately appropriated. We argued that, morally speaking, it was justifiable to view the physical world as owned in common by everyone, before pieces of it were privately appropriated. In particular, we argued against Nozick's amendment of the 'Lockean proviso', which postulated that if a piece of the natural world is unowned, it is all right for an individual to stake it out as her property, so long as she leaves others no worse off by doing so.[14] Well, I guess you could say we didn't necessarily disagree with Nozick's proviso, but we said its premise was vacuous (that there are unowned parts of the natural world), because what Nozick called unowned was properly viewed as owned in common by everyone. (I won't address the question of whether we now might want to include other sentient beings as common owners…) We then argued (Jerry, verbally, and I, using mathematical analysis) that the common ownership of the natural world meant that if someone wanted to appropriate part of it to grow crops on it, or mine it, for instance, she had to bargain with the common owners of that property (everyone else). This would alter sharply the distribution of benefits/revenues from the land. Arguably, no individual would become fabulously wealthy by appropriating parts of the 'unowned' natural world.

In the light of engaging with Nozick's arguments, in the 1980s, you concluded that "the political philosophy justifying Marxism's condemnation of capitalism was a kind of resource egalitarianism" (Roemer 1994a, 2). Further, you write, "the Marxist condemnation of the injustice of capitalism is not so different from the conclusion that other apparently less radical contemporary theories of political philosophy reach, albeit in language less flamboyant than Marxism's" (1988a, 5).

We have two questions here. First, can you briefly explain why you reached this conclusion and also tell us whether you still agree with it?

[14] See chapter V, paragraph 33 of Locke's "Second Treatise" ([1690] 1988, 291) for Locke's formulation of the proviso, and Nozick ([1974] 2013, 178–182) for Nozick's formulation of the proviso.

And second, should Marxists be liberal egalitarians?*[15] *That is, how would you respond to the following charge: while the resource egalitarianism you defend makes Marxists "more consistent egalitarians", as Cohen put it (1990, 382), it has nevertheless left behind all that is distinctive about the original Marxist approach in at least two ways. First, the account of exploitation is now derived from a general principle of distribution (of productive assets), and not from the exchange that occurs within the wage relationship. Second, which is a related but distinct point, we have to now abandon what was the original* raison d'etre *of the original Marxist exploitation argument, to wit: there is an inherent injustice in wage labour.*[16]*As you said just above, before starting your own work on this, you were aware that "exploitation of* labor *was the key to understanding capitalism" (136; emphasis added).

Yes, I think that Marxism advocates a kind of resource egalitarianism: we have discussed that above. Rawls is at once more radical and less radical than Marx. He is more radical because he also views the distribution of the natural talents of people as morally arbitrary, meaning that people should not be viewed as self-owners. He is less radical because he does not condemn the accumulation of wealth as such—or, at least, I and others so argue. I stated that argument above, when I said that if conditions of equal opportunity are implemented through the tax and educational systems, then moderate accumulation of wealth is ethically all right. Marx wrote approvingly of James Meade's concept of a property-owning democracy, and I agree.

I do not believe there is an inherent injustice in wage labor. If I did believe there were, I could not advocate the use of markets under socialism. And I think that without markets, we would be—at this point, before we discover some other way of allocating resources—condemned to terrible inefficiency and poverty. In my recent work, which is the focus of this issue of your journal, I argue that markets combined with solidaristic optimization by workers and investors, produces much better results than capitalism—in terms of both efficiency and equity.

[15] The question is motivated by Will Kymlicka's discussion of Marxism in his introduction to contemporary political philosophy (see especially, chapter 5, in Kymlicka 2002). Building on Roemer's work, Kymlicka concludes that liberal egalitarianism has the superior theory of justice because its account of the institutional requirements of justice is superior to that of Marxism which, in his view, is guilty of a kind of 'fetishism' about labor, stemming partly from its theory of exploitation.

[16] We assume here, as analytical Marxists and others also accept, that there is an—at least implicit—normative condemnation of exploitation in Marx's writings, whatever the truth about his more expressed aims is.

Of course, these ideas will have to be tested in real economies. We will probably discuss that below.

The success of luck egalitarianism, of the type you and Cohen have defended, depends on meeting a pragmatic, and possibly even conceptual, challenge: the disaggregation of those outcomes which result from luck—and for which a person cannot be held responsible—and those which result from choice—for which a person can be held responsible. You have proposed an ingenious solution to this problem that follows a sort of 'fixed effects' approach (1993): partition all relevant agents into 'types'—such as occupation, ethnicity, gender, and the like—whose (socially chosen) characteristics can be said to result from luck. Intra-type differences in characteristics, such as effort, say, are then said to result from choice. We have three questions on this. First, do you believe that such a stark partitioning is, in fact, possible? You have proposed correcting for the fact that some choice-based characteristics are partially also luck-determined by comparing people across types that share the same **rank** ***in the type distribution rather than the same (absolute)*** **level.** ***But to the extent that even striving—to be in the top rank, say—is itself partly due to luck, is this binary partition really sustainable—pragmatically, and conceptually? Some might claim that the relation between choice and circumstance is akin to that between the acquisition of a practical skill, such as playing the clarinet or speaking a language, and its performance: one can't really be said to know how to play the clarinet without playing (sufficiently) decently, but one can't play decently without knowing how to play in the first place—the two happen at the same time.***

Second, can you explain how you propose to compensate **across** ***types while preserving differences*** **within** ***types?***[17] ***And, as a follow up to this, you have pointed out that finding a policy that completely equalises opportunity is almost impossible, and hence that the real policy choice consists in choosing a (social) preference ranking on the available policy alternatives.***[18] ***What kind of properties do*** **you** ***think such a ranking should satisfy—properties that are in concordance with your socialist commitments? And if the criterion is multi-dimensional, how should a society avoid the kind of general aggregation impossibilities observed by Kenneth Arrow?***

[17] This second question is asked by Susan Hurley (2002).

[18] See Roemer and Trannoy (2016, 1308–1312).

You have given a succinct description of my approach to equality of opportunity. But your question, I think, illustrates the different tasks of philosophy and social science. Let us look at history. The abolition of slavery comprised a huge equalization of opportunities: making it illegal for one person to own another destroyed one magnum opportunity inhibitor. It took years, even centuries, for so-called civilized society to understand what the descendants of slaves are owed for the effects of their ancestors' slavery on their own income, wealth, and welfare. When the slaves of Haiti overthrew French colonialism and slavery at the beginning of the nineteenth century, France demanded that the new country pay huge tribute to the former slave owners, on pain of a French military invasion that would otherwise be mounted, to restore Haiti to its former enslaved condition. According to Piketty (2020, 217–220), the Haitians were still paying off this tribute well into the twentieth century, which, he says, is a major reason for Haiti's impoverished condition today. The fact that the French required such tribute illustrates that they did not believe that Haitian slavery was immoral. In the twentieth century, massive improvements in opportunities for women have been brought about by the struggles of women to loosen their shackles. In the 1960s, the injustice of racial discrimination was the focus of the Civil Rights Movement in the United States, led by Black Americans, which greatly improved opportunities for African Americans.

I am saying that the history of the last several centuries can be viewed as one of rectifying the terrible truncation of opportunities of certain peoples, due to certain circumstances—morally arbitrary characteristics of persons, that come to inhibit their chances of leading a fulfilling life. In the middle of the twentieth century, John Rawls provided a general argument that race and sex were only special cases of the morally arbitrary distribution of circumstances whose effects on income and welfare would be eliminated in a just society.

Of course, as you say, it will be impossible ever to eliminate completely these effects. Highly talented people will probably always lead lives that are more successful and happier than they deserve. But we proceed incrementally: we do the best we can. The Enlightenment, beginning, let us say, with the French Revolution, is still far from complete.

As for critics like Elizabeth Anderson, my reply is that the kind of democratic equality that she and I desire doesn't stand a chance of developing when income inequality is as huge as it is today—within almost all nations, and of course, internationally. My goal is to focus on building solidaristic societies, and I think that the most important barrier to solidarity is the individualistic ethos of capitalist society where the accumulation of private wealth is the guiding force. We are still very much in the era when inequality of income and wealth is the main problem. I speak not only of poverty, but of the way capitalist society distorts human behavior and politics. For this reason, I think Thomas Piketty is the most profound social scientist in the world today, for he has revolutionized the study of how massive are the degree and effects of material inequality.

Third, endorsing this sharp partition between choice and luck opens up the accusation—made most recently by Katrina Forrester (2019, 221)—that this move concedes too much ground, and gives too much weight, to the concept of* individual *responsibility which is traditionally associated with the politics of the right. How would you respond to this accusation?

As biology and neuroscience develop, we learn precisely how all manner of biological and environmental circumstances affect our accomplishments. In this process, the ambit of personal responsibility is continually diminished. Behavioral problems of children can be precisely understood as reflections of the poverty of the families in which they are raised, beginning with *in utero* nutrition of the foetus. We learn how stress reduces life expectancy in predictable ways.

The concept of responsibility must be deeply encoded in our genes. Although the boundaries of responsibility differ across societies, I believe no society lacks the concept. I am a compatibilist: I believe that our actions all have a physical representation in our brains, and at the same time, that we rightly hold people responsible for some of their actions.

We correctly educate our children about the difference between right and wrong by commenting on their choices. We attempt to imprint upon their minds a conception of responsible behavior. Can we imagine being human without doing so? I believe leftists have a *deeper* understanding of responsibility than rightists: after all, we teach our children that they are to a degree responsible for others, even if those others are not family members. This is far less true of right-wing parents, is it not?

To press the last point, resource egalitarianism, of the luck-egalitarian variety you have defended, has been dubbed "harsh or paternalistic" by so-called relational egalitarians such as Elizabeth Anderson (1999, 302). It also seems to imply that you cannot wrongfully exploit someone

if their exploitable situation—say, their dire vulnerability—is their own fault. How do your respond to these objections?

Since I proposed my approach to modeling equality of opportunity, in 1993, a large empirical literature has developed in which social scientists around the world have measured the degree to which income inequality in their societies is due to inequality of opportunity.[19] Before 1993, almost all measures of unequal opportunity focused upon *one circumstance*: the rank of the individual's father in the income/wealth distribution of his generation. What these studies call *intergenerational immobility* is a special case of opportunity inequality. Societies in which the individual's rank in the income distribution of *his* generation was only weakly related to the father's rank in the income distribution of his generation were ones with relatively equal opportunity. These studies, to be precise, looked at only one circumstance in explaining the child's income: his father's income rank. It turns out, using the algorithm that I propose to measure inequality of opportunity, that circumstance (father's rank) accounts for less than 10% of income inequality in a society.

Today, in the plethora of studies measuring inequality of opportunity (IOp), it is not uncommon to explain 30%, even 50% of income inequality, as due to circumstances. Of course, these studies look at many other circumstances in addition to father's income rank! This shows how the IOp theory has greatly *reduced* the set of actions for which people are implicitly held responsible. The 'harsh and paternalistic' accusation against luck egalitarianism is belied by the results of scholars who apply the theory to real data. I doubt Anderson has looked at these studies, because very few philosophers look at data. If I can show that, in my country, 50% of income inequality is due to factors that anyone would agree individuals should not be held responsible for, whereas the standard conservative view in my country is that everyone should be capable of pulling herself up by her bootstraps, I have a powerful argument to reform tax, educational, and healthcare policy.

Around the time that the analytical Marxists were presenting their response to Rawls (and Nozick) there was also the so-called communitarian critique of liberalism. Were you ever attracted by these communitarian ideas of, for example, Alasdair MacIntyre and Charles Taylor,

[19] See Roemer (1993) for this early proposal; see Roemer (2002), and Roemer and Trannoy (2016) for subsequent 'progress reports'.

among others? Did they ever inform the discussions during the meetings of the September Group?
No, I wasn't; and we didn't.

You have presented a criticism of Amartya Sen's capability approach as being insufficiently specified (Roemer 1996, 191–193). To be more precise, the claim you make is that a partial ordering (of functioning vectors and capability sets) is an insufficient specification of the object of interest in the context of distributive justice.[20] This is interesting for a variety of reasons, not least because it is a very precise—and very prescient—articulation of a point that has come to occupy the minds of those who are, in principle, committed to the capability approach (we are thinking here of the debate, internal to the capability approach, on whether or not the approach should have a list of relevant capabilities). And further, it presents, we think, a very general challenge to Sen's entire oeuvre which assumes without much argument that partial orderings of states of affairs or opportunity sets are a sufficient specification for the analysis of concepts like rationality, justice, poverty, inequality, and freedom.[21]

But to attempt a defense of Sen, why isn't a partial ranking of functioning vectors and capability sets a sufficient specification? Further, one might argue, a partial ranking (of functioning vectors and capability sets, and for that matter, most objects of social interest, like justice) is not just a sufficient specification of objects of interest in social and political thought, but such a ranking is all that we can really hope to get. Indeed, to demand completeness would be to demand a level of

[20] This is closely tied to discussions on the *extent* of measurability we may hope to get in any analytic exercise. To see why, recall that a partial ordering or ranking is a reflexive, transitive, but not necessarily complete binary relation that stands for a ranking of social states of affairs or opportunity sets (or whatever object the relation is defined over). Partial orders can be seen as a very minimal form of measurability (still weaker forms of measurability—that is, weaker than partial rankings—are, for example, so called fuzzy orders; for an introduction, see Barrett and Salles 2011). Stronger forms of measurability will involve stricter restrictions on the binary relation like, for example: (i) complete orderings; or (ii) numerically representable complete orderings (the so-called ordinal utility scale); or (iii) numerically representable complete orderings that are invariant up to positive affine transformations (the so-called interval scale); or (iv) numerically representable complete orderings that are invariant up to positive multiplicative transformations (the so-called ratio scale). In the context of the capability approach, Sen argues that a partial ranking of functioning vectors and of capability sets is all that we can hope to measure. Demanding more than this is a mug's game for Sen (see Sen 1985). But Roemer is asking Sen for a stronger measure (minimally, Roemer is asking for a complete ordering of these objects).

[21] See the collection of papers in Sen (2004).

precision in measurement that the object being measured does not in fact have. How would you respond?

When Sen first proposed his capability approach, the 'functionings' he mentioned were, as I recall, all objectively measurable. Indeed, the Human Development Index (HDI) for countries that is published each year by the United Nations Development Programme (UNDP) is an average of income per capita, the literacy rate, and life expectancy of the country, three objectively measurable statistics. A few years later, he added happiness as a functioning.[22] I do not know why he did this, but I conjecture that he came under attack from neoclassical economists for ignoring the subjective nature of well-being that is at the heart of neoclassical economics. I, for one, preferred his original approach, where functionings were all something that outside observers could agree upon.

At that time, almost thirty years ago, I viewed Sen's defense of partial orderings as a kind of cop-out, of his not being willing to make hard choices. Quite a few people working in social choice in those years were trying to characterize *complete* social orderings axiomatically. Today, I am not so bothered by this, as Sen has surely played a progressive role in social science, and I think the Human Development Index is an important statistic to have.

Questions of power appear in your writings on political competition and democratic theory. And yet, the topic of power, more broadly—and social and structural power, more concretely—is not as well articulated and focused in your other writings. This seems surprising given the importance of power relations—and their relevance to exploitation—in Marxism. Has this been a conscious choice?

This is an interesting question. In part, the answer has to do with the tools I learned as an economist. Let me begin with Marx, who wanted to show that the inequality (or exploitation, although we could just say income inequality) of capitalism comes about even if all economic transactions are 'fair'. Instead of fair, one might better say 'competitive'. In other words, the vast inequality of capitalism can come about when all workers compete with each other and all capitalists compete with each other. In economics lingo, we say that every buyer and every seller is a price-taker. Neither capitalists nor workers have the power to set wages or prices.

[22] See Stiglitz, Sen, and Fitoussi (2009) for the proposal to include a 'happiness' or 'quality of life' measure in addition to those of income, literacy, and life expectancy.

This is akin to tying one hand behind one's back. It's much easier to show that differential wealth could emerge with cheating, price-fixing, monopolistic practices, physical coercion, and so on. I think Marx was right in this methodological choice. In modern terminology, we'd say that we want to show the genesis of vast inequality of income and wealth—and indeed of exploitation of labor—in a competitive model. This is why I worked with the model of competitive equilibrium in my book on exploitation and class. I showed you could deduce the central Marxian facts about class and exploitation in a perfectly competitive model.

In particular, this means that the important aspects of capitalism can be understood *even if* no individual has market power.

However, it must be said that I am a product of my time. My intellectual development as an economist occurred during the heyday of the general competitive equilibrium model. Had I been educated twenty years later, I might well have worked more with non-competitive models, as are often used in game theory.

Here's another consequence of this approach. Many leftists believe the key to understanding capitalism is to understand the extraction of labor from labor power at the point of production. And indeed, I think Marx sometimes erred in thinking this, as well. My view is that the essence of capitalism is the set of institutions which sanctify and enforce private and unequal ownership of capital—that is, vastly unequal wealth.

Now, workers, surely, do face all kinds of oppression at the point of production—bosses who crack the whip, speed up the assembly line, fire workers who organize, etc. There is a constant struggle at work between workers and bosses about the conditions of work. Today, we see this most dramatically in the low-paid service sector.

I think these struggles occur because of the impossibility of writing a complete and costlessly enforceable contract regulating the exchange of labor power for the wage. The labor contract is notoriously incomplete. The worker shows up at the job in the morning, and at the end of the day, collects a wage. But what happens between showing up and collecting the check is contention and struggle. Imagine one could completely specify exactly what the job entails and what the wage is, and if either the worker or the boss tries to deviate from the agreement, an arbitrating robot immediately enforces the contract. Then there would be no struggle at the point of production. But we'd still have capitalism, exploitation, and inequality, because those things occur even in perfectly competitive equilibrium!

It's in this sense that I believed that power—in at least one form—is not of the essence in capitalism. Now at the more macro or systemic level, wealth brings political power, which produces laws favouring the reproduction of capital, and so on. Of course, I am most interested in political power. In the last analysis, power comes in the police force that enforces property relations. This is the key locus of power; oppression of workers at the point of production, though perhaps very important in building class consciousness of workers, is relatively small potatoes. Coercion at the point of production was essential in feudalism and slavery, but capitalism has subtler techniques for accumulating wealth.

Allow us to dwell a little on the claim that the problem at the point of production is really the (practical) impossibility of writing a complete contract. This also relates back to the prior discussion of intra-class conflict. One of the general results in the literature on principal-agent problems—particularly, those problems arising from the asymmetric information about labour productivity that workers and employers have—is that, given such asymmetric information, optimal (second-best) contracts reward more efficient, or productive, agents with positive rents.[23] This means that incomplete contracts, due to asymmetric information, benefit more productive workers and, as we know, productivity itself is significantly tainted by the arbitrariness of the birth lottery. Wouldn't it be fair to say then that the incompleteness of voluntary contracts itself gives rise to intra-class conflict (between more and less productive workers) at the point of production and that, hence, one of capitalism's 'subtler' types of power is not just related to the reproduction of capital but also to the exacerbation or fuelling of intra-class conflict? We are wondering about the importance of intra-class conflict also because it comes up, as you have shown (1998b), in the political arena as well—when voters vote not just over economic (distributive) issues, such as taxation, but also over other salient noneconomic issues, such as race in the US context. Hence, in addition to these more specific queries, our question is also more general: what kind of theoretical role do you see for intra-class conflict, and if there is any, conflict in which class(es) is the most relevant one?

The point you make (I cannot easily check the Laffont-Martimort citation) is interesting, but it does not strike me as more significant than the effect that trade unions have on *reducing* wage differentiation between skilled

[23] See Laffont and Martimort (2002, 41–43) for a concise presentation.

and unskilled workers. New work using big-data methods in trade union history shows this is a pervasive and significant effect. It's a general phenomenon, which is observed most dramatically in the Scandinavian economies, where the 'solidaristic wage' entails raising the wages of unskilled workers and lowering the wages of the skilled. One effect of this solidarity was to build strong unions, high labor productivity, and high labor-force participation rates. If wage differentials are higher than competitive differentials in a world of complete information, then they are surely much lower than competitive differentials in a world with strong unions. The social solidarity that exists in Nordic countries is, I conjecture, both an effect and a cause of their relatively low degree of income inequality.

I believe racism is the Achilles' heel of the working-class movement in the United States. It is, I think, the main reason that a large section of the white working class supports right-wing politicians who advocate economic policies that impoverish those same workers. Absent racism, the US would be much closer to European-style social democracy. Obama received only 10% of the white vote in Alabama in the 2012 presidential election—most of those white voters were working-class. We need hardly mention that Donald Trump's support among white men with low educational levels has hardly suffered from his open racism and misogyny.

III. BARGAINING, JUSTICE, AND SOCIALISM

We will come back to the importance of solidarity in the next section, but for now let us turn to a theme that, to us, seems to connect your earlier work on Marxian exploitation and equality of opportunity, on the one hand, and your later writings on socialism more broadly, on the other. This common thread—or at least one thread among many—seems to be your criticism of bargaining theory as a suitable framework for discussing distributive justice. Over the years, you have argued that the utilitarian model underlying bargaining theory—the fact that, in the final analysis, its objects consist of thin utility pairs and nothing else—makes it "informationally too impoverished to capture the important issues in distributive justice" (Roemer 1986b, 90). This is of course a criticism made famous by Amartya Sen—the so called critique of 'welfarism'—but on the basis of this criticism you have gone beyond Sen and have defended the use of a much richer informational framework, what you have called **an economic environment.** ***This allows incorporating issues of preferences, needs, resources, and rights,***

including property rights.[24] Is this a fair explanation of your motivation to discard the bargaining framework in favour of the economic-environment framework? Was the main reason the informational penury of standard bargaining theory—as well as progressive alternatives like Sen's account—which did not capture the importance of having a framework where* property rights *can be explicitly represented and discussed?

This is an important question, but we should clarify for readers that what Sen, and later I, were criticizing is that many of the models in social choice theory and what is called bargaining theory take the only language to be the language of utility. The building blocks of all theory are the vectors (or lists) of utility numbers that persons realize under different policies, or institutions, or systems. There is no way to formulate private (or public) ownership of firms because property rights are not 'utilities'. Sen showed, with a simple example, that our moral intuitions often require the idea that people have *rights*: but rights do not exist in a model where the only way to describe a person's situation is by his utility.[25] (Sen's example spoke of human rights, whereas I referred above to property rights. Neither can be represented in a welfarist framework.) Think of Locke, or Nozick. Property rights (who owns the external world, who owns a person's labor power) are of the essence. Utilitarianism is a theory that judges the goodness of a situation by the vector of utilities of persons that is associated with it: how that utility vector was generated is of no interest.

Now one might respond: in the final analysis, we are interested in human welfare. So, property or human rights are only important in so far as they generate patterns of welfare or utility across persons. However, one's

[24] Economic environments are used in Roemer's latest book, *How We Cooperate* (2019a), but see Roemer (1986b, 1988b), and Moulin and Roemer (1989) for earlier motivations of the framework. For Sen's critique of welfarism, see Sen (1979).

[25] Sen's 'human rights' example has two parts. Let x and y be two states of affairs, involving two agents, r (rich) and p (poor). In x, there is no redistributive taxation, while in y, some of r's money is taxed away for the benefit of p, but r remains richer than p. Suppose that the utilities of r and p in the two states are: (10, 4) in x, and (8, 7) in y.

Next, let a and b be two states of affairs, and let r ("a romantic dreamer") and p ("a miserable policemen") be two agents. r has a happy disposition, is rich, in good health, etc.; while p is morose, poor, in ill health, etc., and his only pleasure is torturing other people. In a, no torturing takes place, while in b, p tortures r. Suppose that the utilities of r and p in the two states are: (10, 4) in a, and (8, 7) in b.

Sen's point is that whatever one's ranking between x and y, it must be the same as that between a and b. If one believes that y (redistribution) is better than x (no redistribution), but that a (no torture) is better than b (torture), then, to account for this, one must bring in non-utility information. See Sen (1979, 473–474) for a fuller discussion.

language is severely impoverished if one cannot refer to property or human rights. Recall one of Sen's examples.[26] Under standard feudalism, let us say, the welfare levels of the Serf and the Lord are one and ten, respectively. Now suppose we can change this to four and six by reducing the days of labor that the Serf works on the Lord's demesne. Or, alternatively, we can improve standard feudalism by allowing the Serf to whip the Lord every Friday, in which case their welfare levels are also four and six. If you have only utility language to discuss outcomes, you must be indifferent between these two 'policies', for they are equivalent in their utility consequences! Most of us think that achieving (4, 6) by allowing whipping is a very different thing from achieving it by changing the labor contract.[27]

I extended Sen's critique of welfarism in the theory of equality of opportunity that I proposed. The language of that theory includes circumstances, effort, and type. These are fundamentals, along with utility. One cannot judge how just a situation is by knowing only the welfare levels of people in it: one must know *how hard they tried* and what their *circumstances* were. The equal-opportunity theory is *non-welfarist*. It's not only rights talk that is banned by welfarism, but all non-utility talk.

We wonder—from an intellectual-history perspective—whom did you see as the main interlocuter(s) you wanted to convince with this work? Was it economists and game theorists, such as John Harsanyi and Ken Binmore, who at that time were cementing a tradition—among economists and political philosophers—of modeling the question of distributive justice as a utility-allocation problem? Or political philosophers of the contractarian tradition, such as David Gauthier, whom you do mention in your writings, who were picking up on bargaining theory as a framework for discussing distributive justice? Or was your intended audience different?

I came to think about these problems from a Marxist background, where *exploitation* was the key idea, and the grounds for the critique of capitalism. Exploitation, par excellence, is a non-welfarist idea. We don't say exploitation is bad because it gives the worker lower utility than the capitalist: we say it is bad because it violates the freedom of the worker to develop her capacities, or that it is the consequence of unequal ownership of capital that came about through robbery and pillage. There is also a

[26] See note 25 for this example.

[27] Author's note: But not all of us. My friend David Donaldson, a welfarist, responds: 'On the contrary, it would be great if we could have solved the injustice of feudalism by allowing some Lord-whipping, instead of having to go through bourgeois revolutions.'

condemnation of exploitation along grounds of freedom. These are all non-welfarist reasons for attacking exploitation. Of course, I absorbed the Marxist critique long before learning the philosophical concept of welfarism.

My own first attempts to work on distributive justice used welfarist models—the models of axiomatic bargaining theory. I eventually saw how welfarism severely restricted and *over-simplified* the discussion of justice in economic theory. I argued that we should study justice using economic models, where we have a language for ownership, commodities, markets, and so on. After all, one cannot even define what socialism and capitalism mean without such a language.[28]

Coming back to the question of property rights, you have argued for making a distinction between **common ownership** ***and*** **public ownership.** ***Common ownership of a resource refers to "the right of each to free access" to the resource (1988b, 700) and should be distinguished from public ownership. While you have not provided an explicit definition of the latter, in your axiomatic discussion of allocation mechanisms that respect public ownership, you have drawn on the idea of respecting the right of*** **use,** ***as opposed to the right of*** **ownership** ***(1988b, 705). And the distinction between these two kinds of property rights seems to run throughout many of the proposals for market socialism you have advanced over the years. Is it fair to say that, for you, the full right of ownership should be restricted to labour power; while, for all other factors of production, the relevant property right is that of the right of use? Further, is it fair to see this distinction as motivating your proposal for a coupon economy (1994b, 75–84), for example, and, more recently, for a sharing economy (2020b, 27–32)?***

I am unsure how we should define common and public ownership. If a village owns some land, upon which all members of the village can graze their livestock without formal constraint, that is surely common ownership. However, if this practice leads to overgrazing, and the village restricts how much each resident can graze in order to sustain the land, that land becomes publicly owned. If, however, residents learn to choose how much they graze by Kantian optimization instead of Nash optimization, and they thereby sustain the land without need of formal restrictions, I suppose I would say the land is still owned in common. Public ownership, I think, should mean that everyone in the community has

[28] Author's note: One article where I presented this view was Roemer (1986b).

access to the land under rules established by the community, which are constraining, while common ownership is a system with no stated rules. Perhaps common ownership is analogous to common law, where there is no written constitution, but tradition suffices to regulate the commons.

I do not agree that 'the full right of ownership should be restricted to labour power'. We should not have all the rights over our labor that a slave-owner has over a slave—this would be full ownership of our labor power. If the talents we have are in part morally arbitrary, they should in part be owned by the community. For a person not to be a full self-owner does not mean the community is free to harvest one of his kidneys to transplant into another, but it may well mean that he must pay taxes on his earnings to the state. To be a self-owner means (according to G. A. Cohen) that a person has *all* the rights over his bodily powers that a slave-owner has over a slave. To be a *non-self-owner* means a person does not have *all* the rights over his bodily powers that a slave-owner has over a slave. It's a logical error to say that a non-self-owner has *none* of the rights that a slave-owner has over a slave. The libertarian attack on common ownership of talents—that it would expose everyone to possible kidney harnessing—is a non sequitur.

We would now like to turn to two of these practical proposals for implementing the ideal of socialist equality of opportunity. In 1994, writing in* A Future for Socialism *(1994b), you argued for a form of managerial socialism, which would give every citizen an equal and tradeable share in the beneficial ownership of the means of production. In this system, there is a coupon stock market in which coupons are freely tradeable for shares in firms but not monetisable or bequeathable. In what way was this a form of socialism—as opposed to, say, a form of corporatism or, to use Lenin's term, 'state capitalism'?

It isn't state capitalism, because the state does not receive the profits of firms: these profits are distributed to citizens as individuals. Furthermore, individuals can trade their rights to receive the dividends of particular firms on a stock market. But it isn't private ownership of firms by citizens, because an individual cannot capitalize his right to receive firm profits by selling these rights to another individual for *money*. I intended this system to insure that every citizen had a right to a share of the nation's capital income, which he could not relinquish. This is very different from the coupon capitalism that was introduced in some Eastern

European countries in the early 1990s, in which the poor rapidly sold their coupons for cash to the rich.

Suppose I live in a city which has a large park that all city residents can freely visit. I cannot sell my right to a person in another city to use our park. The only way I can sell my right to use the park is to sell my apartment and move out of the city. I then transfer my right to use the park to the person who buys my apartment. The right to capital income from the nation's firms in the coupon economy is this kind of right. I don't think it's appropriate to call this state capitalism or corporatism.

The sharing economy you have defended more recently (2020b) is similar to the coupon economy in at least one respect (although we will come back to the major difference, the idea of a* behavioral ethos, *a little later)—both models of public ownership tend not to include worker control over the firm. Does worker control have a secondary, or derivative, role in your vision of socialism? And, if so, wouldn't that make this vision liable to being overtaken by another ruling class, this time in the form of managers?

There certainly is worker *ownership* in my model of the sharing economy. There is also, so I propose, ownership by investors. In the sharing economy, a person receives a share of the firm's profits by either investing or working in the firm. Under capitalism, one can purchase a right to firm profits by buying another person's right to receive profits, by purchasing her shares. In the sharing economy, there is no stock market—rights to profits only go to investors and workers. Granted, I have not discussed worker *management*.[29] My view is that the board of directors should consist of workers, investors, and other citizens. Probably the closest model today is the German corporate system.

As you acknowledge, one of the more controversial proposals in the sharing-economy model is the idea of ownership by investors. The vision behind this proposal includes a pool of households which can supply capital, or labour, or both, and so which can receive profit shares proportional to their investment, or labour, or both—at least in the non-degenerate variants of the model where residual profits are not allocated entirely to workers. Yet, given the current patterns in the ownership of capital, if such a model were implemented, it is plausible to conjecture that, as you say, "class differences will continue to remain

[29] Author's note: Investors buy bonds issued by the firm, not stock.

between those whose incomes come primarily from labor, and those whose incomes have a significant capital component, and membership in these classes will therefore continue to be closely correlated to social and economic advantage in family background" (Roemer 2020a, 24). Particularly so, if the model allows for the existence of labour and bond markets. Doesn't the implementation of this vision require a substantial redistribution of capital before it is put in place? And, in general, how do you envision implementing the model—and the initial conditions it requires—in practice?

Absolutely it does. Not only must there be high estate taxes, preventing the transmission of large amounts of wealth to descendants, but the distribution of wealth must be far more equal than it is today *inter vivos*. Thomas Piketty (2020, chapter 17) discusses taxation in some detail. He speaks of the "progressive tax triptych: property, inheritance, income" (2020, 981). I have little to add beyond his discussion, except for the motivation for substantial wealth taxation, which is to provide the conditions for sustaining a solidaristic society. That is to say, there is an inconsistency between permitting positive returns on private investment, and maintaining conditions on distribution that will support solidaristic economic behaviour. In my view, what has to give is unconstrained accumulation. I emphasize that this is, I believe, the key problem of socialist finance. If the state is not to own all the wealth in society, then households must be able to invest, and that would lead, without sufficient taxation, to inequality of wealth, lack of solidarity, and political influence by the wealthy. I cannot claim to have the definitive solution to this problem, but I follow the tradition of James Meade and others who thought that a property-owning democracy was a feasible version of socialism. The alternative, of having the state be the sole owner of capital, has its own pathologies, as we know.

Models of market socialism—among others—have been criticised by, for example, feminist economists and philosophers that, while the models pay careful attention to the conditions conducive to the reproduction of capital, they do not pay sufficient attention to the conditions conducive to the (physical) reproduction of labour.[30] The sharing-economy model, which is indeed attentive to investment incentives, seems to be liable to the same criticism. Have you thought about this objection,

[30] See Müller (forthcoming).

and, particularly, about the ways in which incentives for the reproduction of labour can be similarly aligned in practical terms?

I believe that the *sine qua non* for a society that invests in people is solidarity, which, as I say, requires a quite equal distribution of income and wealth. That's why I focus on material distribution. Obviously, the socialist society should invest in education, health, housing, infrastructure, the arts, research, and so on—that is, the basis for the 'production' of successful human beings. I am impatient with critics who claim that those of us who focus on the distribution of income and wealth do not care about these things. If one is a democrat, one understands that the only way to produce good policies is to have a solidaristic polity whose members will choose the right politicians and policies. Capitalism is a system which breeds greed; most successful capitalists are greedy people, and this infects the whole society, as Marx made abundantly clear. We see a glimmer of what solidaristic societies would look like when we examine the Nordic countries. Leftists in these countries are highly critical of their societies, and bemoan the departure from a more solidaristic period after the Second World War. But for global human society, I believe these countries remain a beacon. Preserving their example is of utmost importance to the world.

Let us now turn to your more recent thinking about socialism with a very general question. In 1988, G. A. Cohen outlined three overarching issues that "should command the attention" of those working "within the Marxist tradition" at the time:

> ***They are the questions of design, justification, and strategy, in relation to the project of opposing and overcoming capitalism. The first question is, What do we want? What, in general, and even not so general terms, is the form of the socialist society that we seek? The second question is, Why do we want it? What exactly is wrong with capitalism, and what is right about socialism? And the third question is, How can we achieve it? What are the implications for practice of the fact that the working class in advanced capitalist society is not now what it was, or what it was once thought to be? (Cohen 1988, xii)***

How would you, most broadly, answer these questions today? Is it fair to summarise your answers as follows: we want equality of opportunity, because of the injustice of the unequal capitalist distribution of the means of production, and we can achieve it through market socialism?

We want socialist equality of opportunity (Cohen's term) and we want to build a cooperative ethos, because I conjecture that the only way of achieving sustainable equal opportunity is through cooperation. I'll expand upon this below. I believe a version of market socialism is the path to take.

Related to Cohen's strategy question above, in acknowledging the changing nature of the working class, you said that "[t]he proletariat, those who own nothing but their labor power, no longer constitute a majority of advanced capitalist societies. Nor are the neediest […] clearly members of the productive working class" (1994b, 15–16). How would you define the working class today? And what do you think is the size and scope of the petty bourgeoisie today? We ask this latter question also in relation to the class-exploitation correspondence principle you talked about earlier, because the petty bourgeoisie in those models is the only one for which the principle does not—realistically—hold as a one-to-one correspondence; that is, for which class membership does not necessarily imply exploitation status. Finally, in light of these class changes, if any, what are your current views on the usefulness of a class-based analysis?

The class-exploitation correspondence principle (CECP) is a theorem relating the class position of a person in capitalist society to her exploitation status—whether she is exploited, is an exploiter, or is neither.[31] Because I have come to think that exploitation is a detour around our main concern—to implement socialist equality of opportunity—I now think of the CECP as a contribution to the history of thought. The CECP shows that one can define exploitation and class position independently, and then prove that there is a tight relationship between the two characterizations of an individual (worker, capitalist, rich kulak, landed laborer, etc.). Marx defined exploitation quite abstractly, in terms of labor commanded versus labor expended, which is not evidently the same thing as a person's class position, his relationship to the means of production. But he did not possess the economic theory to show precisely the link between these two central concepts of his theory of capitalism.

I do not have anything special to say about class analysis. One part of Marxism I continue to find enlightening is historical materialism, which has an important role for class struggle. I find it useful to view evolution in the economic structure as a mandated adjustment to technological

[31] For more on the CECP, see the discussion, and references, on page 137.

change. Historical materialism, as explained by Cohen in his magisterial book on the subject,[32] views class struggle as the midwife on the birth of new social systems, although not the fundamental cause of that birth, which lies in 'the development of the productive forces'.

IV. KANTIAN OPTIMIZATION AND THE FUTURE OF SOCIALISM

***Let us turn to your most recent book,* How We Cooperate: A Theory of Kantian Optimization*, to which the current issue of the* EJPE *is devoting a book symposium. The book systematises your work on Kantian optimization across a series of papers in the 2010s.*[33] *In a recent manuscript, you argue that "any socio-economic system has (in my view) three pillars: an* ethos *of economic behavior, an* ethic *of distributive justice, and a set of* property relations *that will implement the ethic if the behavioral ethos is followed" (Roemer 2020a, 3).*[34] *The behavioral ethos of socialism, you continue, is cooperation and you propose to model this "cooperative ethos" with the concept of Kantian optimization. This you contrast with the "individualistic ethos" of capitalism which, you say, is "neatly modeled by Nash optimization" (Roemer 2020a, 5). The manuscript thus places the concept of Kantian optimization in this ambitious project that follows naturally from your work throughout the years. And yet, this more ambitious project is absent from* How We Cooperate *where Kantian optimization is presented more narrowly as an alternative—descriptive and normative—solution concept to the dominance of Nash optimization. We have two questions here. First, which of these two motivations—the broader or the narrower—was what inspired you to work on Kantian optimization in the first place? And, second, why did you omit the broader role of Kantian optimization as a model of the socialist cooperative ethos from* How We Cooperate?**

It was the narrower goal—to conceptualize cooperation as something quite different from individualism, as a project in game theory—that motivated my work on Kantian optimization. Indeed, the 'three pillars' idea that you mention only congealed in my thinking recently. That's why it's not in the book *How We Cooperate*.

[32] See Cohen [1978] 2001.

[33] See Curry and Roemer (2012), and Roemer (2010, 2015). See also the 2019 special issue "Cooperative Behaviour, Kantian Optimisation and Market Socialism" in the *Review of Social Economy* 77 (1).

[34] For a more accessible discussion of these same issues, see Roemer (2020b).

Research is full of luck and serendipity. When I discovered the approach of Kantian optimization, I felt as if I had found a $5 bill lying on the sidewalk. A nice alternative to Nash optimization, which resolved the tragedy of the commons and the free-rider problem. It took me literally years to pick up that $5 bill, and to notice there was a $100 bill lying underneath it. This was a solution to the design problem of socialism that Jerry Cohen expounded. The 2020 'three pillars' paper that you cite even proposed, boldly, that each economic system has its specific form of rationality—individualism (Nash optimization) for capitalism, and cooperation (Kantian optimization) for socialism. I will attract much flak, I think, for this proposal, and I may in the end abandon it. Let's see what people have to say about it.

The design problem, just to be clear, that Jerry Cohen proposed was that although we have many ideas about the goals of socialism, we lack the engineering details to make it work. The design details of capitalism that make it function are the 'greed and fear' induced by huge wealth inequality and markets. My alternative design of cooperation conceived of as Kantian optimization is a specific answer to Cohen's challenge to replace greed and fear. I fully expect others to improve upon it.

In the book, your main motivation of the assumptions behind Kantian optimization is grounded on the concepts of 'solidarity'—"in the sense of our all being in the same boat"—and 'trust'—"trust that if I take the cooperative action, so will enough others to advance our common interest" (2019a, 6). Why did you choose solidarity and trust instead of, for instance, a more Smithian concept such as 'empathy'? In a joint 1991 chapter with Ignacio Ortuño-Ortín, you defended the possibility of interpersonal comparisons of utility precisely on this latter basis—that is, empathy—when you said that "it may be quite reasonable to suppose the existence of an interpersonal ordering of the states of the world, based on a kind of empathy that a person can legitimately feel, because he has, during his life, indeed been a person of various different types" (Ortuño-Ortín and Roemer 1991, 321). Doesn't Kantian optimization also require interpersonal comparisons, not in the traditional cardinal sense, but in this broader sympathy-based sense?

Yes, it does. Kantian optimization works by forcing actors (players in a game) to take into account the externalities, positive or negative, of their actions for others. The trick is to find an appropriate sense in which our joint actions enjoy a kind of symmetry. This can be described as 'taking the action one would will be universalized', which is the link to Kant. I find cooperation, as so described, as easier to achieve than altruism. It may be quite close, however, to empathy.

Could you expand on the relation between your account of Kantian optimization, on the one hand, and your luck egalitarian account of the central injustice of capitalism, on the other? By luck egalitarianism, income differences due to choice, adequately compensated for differences in luck, are just. It follows that a form of capitalism, the 'cleanly generated capitalism' of luck-compensated choice, is just. But such capitalism is likely, if not bound, to conflict with the 'cooperative ethos' warranted by Kantian optimization. Does the 'cooperative ethos', at least from some point on, preclude what is just?

No, it's the opposite. The constraint against a fully luck-egalitarian ethic is the need to restrict income inequality in order to preserve the cooperative ethos. I claim human nature is incompatible with cooperation among individuals whose incomes or consumptions differ by orders of magnitude. So, to preserve cooperation, we must limit income and wealth inequality, and therefore, perhaps, a fully luck-egalitarian system.

Implicit in the view that there are three pillars to a socio-economic system is the claim that it is not sufficient to define capitalism and socialism as modes of organizing activity with very different underlying property relations. There is, of course, an old tradition within socialist thought which holds that behavior matters as well, and it gets a powerful articulation in G. A. Cohen's **If You're An Egalitarian How Come You're So Rich** ***(2000). But how did*** **you** ***come to this view?***

I came to this view because cooperation has long been a characterization of socialist behavior. Until recently, I thought of cooperation under socialism as fully represented by collective ownership of capital. I now think that such ownership is insufficient to characterize cooperation. Cooperation refers to *behavior* in economic behavior, which is insufficiently summarized by property relations of a certain kind. The contrast between individualism represented by Nash optimization—going it alone—and cooperation as represented by Kantian optimization—hanging together—is self-evident. As I've said, whether one should go as far as saying these represent system-specific forms of rationality is an open question for me. I say this third pillar of an economic system—it's behavioral ethos—is as important as the other two.

As a follow-up to the previous question, what in your view is the relation between the notion of historical materialism, which you have acknowledged you are still attracted to (Adereth and Hodges 2019), and the notion of a behavioral ethos?

I think I have covered that. What I'm proposing is that there is a specific behavioral ethos associated with any economic structure. You might ask me, what's the behavioral ethos of feudalism? I invite suggestions from readers.

In* A Future for Socialism *(Roemer 1994b, 28–36), you outlined a short five-stage history of the idea of market socialism. The ideas of incentive compatibility and the principal-agent problem are two important developments you note in this history. These ideas, however, are based on a kind of Nash optimization. Do you believe that they are also relevant for Kantian optimizers? Put differently, do Kantian optimizers face incentive-compatibility constraints?

This is a very good question, which I've danced around, but have not thought about sufficiently. The only place where I've addressed the issue is section 3.3 of *How We Cooperate* (2019a, 51–53). It deserves deeper consideration.

One place where a version of incentive compatibility comes up is in my insistence that trust among the players of a game is a necessary condition of their playing the Kantian equilibrium. In a simple Kantian equilibrium, each player is supposed to take the action she would like everyone to take (say, going out on strike). I say that each must *trust* that if she takes the Kantian action, then so will all (or at least most) others. If others, in contrast, play the Nash-optimal action, she who plays the Kantian action will generally be very badly off—she will be exploited, if you will, by the Nash players. (What would happen if *only one worker* goes out on strike, when the Kantian action is that all should do so?) To say that trust is required for the players to take the Kantian action is therefore admitting that, in the absence of trust, it is reasonable for a player to take the Nash action—to avoid being left out on a limb by the (non-cooperative) others. This is admitting that the player contemplating the Kantian action should not be expected to take it *unconditionally*, but only on the assumption of the cooperative behavior of others. This is a version of incentive compatibility. In contrast, a true Kantian, one who follows the categorical imperative, must be committed to taking the Kantian action regardless of what others do, because morality requires it.

Questions of epistemology have been central to Marxism, in for example, discussions of false consciousness and ideology. Have you worked on or thought about epistemic issues?
I've written a short piece on epistemic questions in the theory of equality of opportunity, but that's it.[35]

As a segway to the next section, you have made the case for Kantian optimization, as a rival of Nash optimization, on the basis of the properties that the equilibria it gives rise to satisfy (at least in the most fraught situations plagued by positive and negative externalities). Have you thought about approaching the problem of adjudicating between these (and other) solution concepts from a more general Arrovian perspective—the kind of approach that is standard in cooperative game theory? What would be an indispensable list of desirable properties that **you** ***would like a solution concept to satisfy, and would these be context-dependent?***
I've come to think that the approach of axiomatic characterization practiced by social choice theorists is only worthwhile if the axioms are few and transparent, and the result is surprising. Arrow's impossibility theorem and Nash's solution to the bargaining problem are examples that pass the test. A less well-known example is the Hart-Mas-Colell axiomatization of the Shapley value, which requires only a single axiom.[36] In the case of Kantian versus Nash optimization, I think the two approaches are so transparently different, and so clearly related to individualist and cooperative behavior, respectively, that little would be gained by the kind of axiomatization you suggest. But I do not want to discourage you from thinking about the problem.

To take one salient property, Pareto efficiency features prominently in your work as a desirable property that Kantian—but not Nash—equilibria satisfy in canonical situations plagued by positive and negative externalities. More generally, from our experience, Pareto efficiency is perhaps the first property that an economist would point out as a desirable property for an equilibrium to satisfy. Why, in **your** ***view, should people care about efficiency?***
Efficiency means not wasting resources. This is obviously of huge importance. Saying so does not imply one would never trade off some

[35] See Roemer (2020c).
[36] See Arrow ([1951] 2012), Nash (1950), and Hart and Mas-Colell (1996), respectively.

efficiency for greater equity. Often, however, one can achieve equity and efficiency at the same time. It is that goal that characterizes the best economic analysis. Excessive carbon emissions, inducing damaging climate change, is perhaps the greatest inefficiency of our time; it is a classical example of the difficulty of achieving efficiency in a global economy with public goods and bads. Kantian optimization enables us to see the precise link between lack of global cooperation and climate change.

You mentioned earlier that, in the final analysis, the theoretical superiority of the Kantian protocol over the Nash protocol, in terms of both efficiency and equity, would need to be tested in real-world economies. Have you seen evidence that this superiority also holds in past or present existing economies?
I think that we should examine the experience of the Nordic economies to see if their success, to a degree, is a result of Kantian optimization. I have some conjectures, but no results to discuss at this point. In almost all capitalist countries, I see trade-union consciousness, or more generally working-class consciousness, as closely linked to Kantian optimization. I am currently studying actual vaccination behavior in a society as perhaps being better explained by Kantian optimization than Nash optimization. Showing this requires some careful analysis. Recycling and other behaviors to reduce environmental degradation are another example where Kantian optimization seems to better explain behavior than Nash optimization. My hope is to show that there are indeed many examples of Kantian optimization today, in many societies: in part, we tend not to see them, because we (economists at least) look at the world through the lens of Nash optimization. Recall the warning ‘equipped with a beautiful hammer, every problem looks like a nail’.

V. METHODOLOGY AND INTERDISCIPLINARITY

Your work, together with that of Jon Elster, has been called ‘rational choice Marxism’ because, unlike ‘analytical Marxism’, it is committed to methodological individualism. You have defended this methodological position extensively in the past,[37] although you have also been explicit in recognising the limitations of “the individualist formulation of the economic problem” (1978, 149) that underlies neoclassical theory.

[37] See the 1982 debate between John Roemer, Jon Elster, Gerald Cohen, Philippe van Parijs, Johannes Berger, Claus Offe, and Anthony Giddens in *Theory and Society* 11 (4).

Have you ever considered straying away from methodological individualism? Just to be clear: we are asking about* methodological *individualism which is to be contrasted with the type of, one might say,* behavioural *individualism you spoke of earlier (Nash optimization vs the cooperative individualism of Kantian optimization).

I agree with Elster that a complete explanation of a social phenomenon requires exhibiting the mechanism whereby it occurs. In one of his recent books, Elster proposes that the study of mechanisms should be the full program of social science.[38] I probably would not go that far: that is, I believe there is a role in social science for observing relationships, even if one cannot prove causation. The search for mechanisms is the search for causation by human decisions. Is it fair to say that Newtonian mechanics, despite its wonderful precision, is not a full mechanism in Elster's sense? For it provides no explanation for gravity: it 'merely' describes how gravity behaves, but does not answer the question of how it comes to be that masses of atoms attract each other.

In my own recent work, the question arises as to what would cause a group of people, engaged in a project that can be modeled as a monotone game, to employ Kantian optimization rather than Nash optimization as their optimization protocol. I have said there are three requirements for such cooperation: desire, understanding, and trust.[39] People must desire to cooperate, based upon their understanding that if such cooperation succeeds, the results will be better for them than they would be if everyone 'goes it alone'. Furthermore, each must trust that if she cooperates, so will others—as I have noted above. This is to a degree a methodologically individualist explanation. But, like the problem of gravity, it does not go deeply enough. I have further argued that our brains seek symmetry, and our concept of morality is deeply linked to symmetry.[40] Skeptics can argue that I am only *describing*, not *explaining*.

I must comment on your citing my 1978 article in which I said that mass action is not individually rational. I wrote mass action is explained by collective rationality, and I attacked 'constrained optimization' as an instance of neoclassical economics that a Marxist would not use. I now blush at those words. Of course, I did not have the concept of Kantian optimization in 1978; I would now explain mass action as an instance of it (see, for example, my model of 'strikes' in *How We Cooperate*, 2019a,

[38] See Elster (2007).
[39] See Roemer (2020b, 44).
[40] See Roemer (2019a, 70).

54–57). Although a generous reader might allow me to interpret Kantian optimization as a more precise formulation of the ‘collective rationality’ referred to in the 1978 article, there is still much in that article that I now disown.

A common objection to rational choice Marxism is that it makes no room for endogenous preferences or for macro-structural constraints. G. A. Cohen, for example, argued that the proletarian in a capitalist economy is individually free to enter bourgeois society (for example, by starting her own shop), but the proletariat as a whole is not collectively free to do so. This is a macro-structural constraint. How does your version of individualism deal with the problems of endogenous preferences and macro-structural constraints?

It seems to me that what you call macro-structural constraints are dealt with by the equilibrium method, in the sense that supply must equal demand at equilibrium. That seems to ‘explain’ Cohen’s example of the collective unfreedom of the proletariat under capitalism. I think the lack of a full theory of endogenous preferences is a major weakness of economics. Progress is being made on this front, however, with many people thinking about culture, as you say.

Finally, allow us to turn to a few questions on economics as a discipline and as a practice, and particularly in its relation to philosophy. In 1996, you wrote that “economics is the handmaiden in this relationship [between economics and philosophy]. The economist’s way of thinking can check the consistency of a philosophical theory or provide a concrete formulation (a model) to make precise some of its still vague assertions” (1996, 3). This statement was made in the context of theories of distributive justice. Have your views on this changed? What do you now think is the value of the philosophical and the economic way of thinking?

Jerry Cohen once said to me that the goal of philosophy is to formulate vague ideas as precise questions. Once an idea is posed as a precise question, philosophers move onto something else—they lose interest in it. Economics, in contrast, attempts to answer precise questions. It does not typically worry about the vague idea that must have led to the precise question. If this is the intellectual division of labor, then obviously both philosophy and economics are important.

We see most of your work, both your past and your most recent writings, as falling unambiguously in the domain of welfare economics. Is this a fair characterization? Indeed, your recent work on Kantian optimization has even replicated and extended the first theorem of welfare economics to the case of socialist economies populated by agents optimizing according to the Kantian protocol.[41] And yet, it is hard to disagree with Anthony Atkinson (2001) that, compared to the heyday of welfare economics in the 1960s and 1970s, that approach has 'strangely disappeared'—at least from the mainstream discussions and standard curricula. When Angus Deaton put the same question to Amartya Sen, Sen noted that the loss associated with the 'strange disappearance' of welfare economics is not just exclusive to economics. Even the discipline of **philosophy** ***has lost something valuable.[42]***

If indeed you agree with Amartya Sen and the late Anthony Atkinson, we have three related questions. First, should this 'strange disappearance' be seen as a loss—both as a loss in economics, and as a failure to bring serious economics into philosophy—and if so, why? Second, what, in **your view,** ***has been responsible for this loss, with respect to both disciplines? And finally, how do you think welfare economics should, if at all, be incorporated in the economics and the philosophy curriculum today?***

Of course, I agree that the disappearance of welfare economics is unfortunate. I'm not sure, however, that I would focus upon rejuvenating it. The most imaginative work among progressive economists today is empirical work: the work on inequality by Piketty and Emmanuel Saez, on inequality of opportunity in the United States by Raj Chetty and his lab at Harvard, the imaginative work on economic history by young scholars like Suresh Naidu at Columbia University and Avidit Acharya at Stanford. This is where the energy appears to be—I am simply naming a small number of scholars as representative of a much larger group. I'd also like to give a plug to the wonderful new introductory economics textbook project CORE, written by a team led by Samuel Bowles and Wendy Carlin.[43]

We ask about welfare economics not only because it is regrettable that it has 'strangely disappeared', but also because tackling most pressing

[41] See chapter 13 in Roemer (2019a); see also Roemer (2019b) and, for a discussion, Maniquet (2019). An accessible, informal, presentation of these results is in Roemer (2020b).

[42] See Sen, Deaton, and Besley (2020, 17–18).

[43] The textbook is freely available online at core-econ.org.

issues today requires the kind of interdisciplinarity that used to inform the work of welfare economists. What in your view are the most pressing issues—in terms of both specific questions as well as broad research agendas—that we need to be tackling today?

Surely it is right to emphasize interdisciplinarity—although this has become somewhat of an empty mantra. I am less familiar with what is happening among young philosophers; I am skeptical that there is the kind of creative explosion based on the careful use of data that I have described in economics. I hope I am wrong. Surely Katrina Forrester's work is important although it is marking the closing of an era in political philosophy, not the beginning of a new one, isn't that right?[44]

The* EJPE *is an interdisciplinary journal, and our readers are scholars who either work at the intersection of philosophy and economics, or are at least open to such an interdisciplinary approach. We would like to ask you to address a couple of questions which might be particularly pressing for young scholars just entering the field. First, do you think that there is an ideal profile, an ideal set of skills, or at least an indispensable set of skills, that someone who follows a philosophy-and-economics approach should have or strive to develop? Second, how should one go about developing these skills? Finally, and somewhat in relation to the second question, what set of specific readings—or courses—would you recommend to junior scholars who are just starting out and starting to adopt an interdisciplinary philosophy-and-economics approach?

You make me feel like the jazz artist in the mid-twentieth century—I forget who it was—who was asked by a journalist: 'Where do you think progressive jazz is going?' He responded: 'Man, if I knew that, I'd already be there.' A young left-wing intellectual who wants to do good work should focus on the aspect of the academic trade that she enjoys. One must love the practice of the trade in order to put in the thousands of hours needed to become proficient. In my case, the trade was mathematics, but I certainly wouldn't say everyone has to learn mathematical modeling. Great contributions are made by people in all fields. Technological change may also be influential (shades of historical materialism): the important empirical work being done now in economics would not have developed absent the computer and the internet. Do what you enjoy doing, and attempt to make a long-range plan of what you want to accomplish. Pick a problem

[44] See Forrester (2019).

and work on it hard. I find it takes about ten years for my work on a problem to become mature, so be patient. In an intellectual life of forty years, count yourself a success if you can develop to fruition three or four good ideas.

Professor Roemer, thank you so much for sharing your time and ideas with us.

And I thank you. It has been my pleasure to ponder the astute questions that you have posed. It is shy-making to see that you have paid such detailed attention to my meandering path.

REFERENCES

Adereth, Maya, and Jerome Hodges. 2019. "Exploitation, Cooperation, and Distributive Justice: An interview with John E. Roemer." *Phenomenal World*. Accessed July 28, 2020. https://phenomenalworld.org/interviews/john-roemer.

Anderson, Elizabeth S. 1999. "What is the Point of Equality?" *Ethics* 109 (2): 287–337.

Arrow, Kenneth J. (1951) 2012. *Social Choice and Individual Values*. 3rd edition. New Haven, CT: Yale University Press.

Atkinson, Anthony B. 2001. "The Strange Disappearance of Welfare Economics." *Kyklos* 54 (2–3): 193–206.

Barrett, Richard, and Maurice Salles. 2011. "Social Choice with Fuzzy Preferences." In *Handbook of Social Choice and Welfare. Volume 2,* edited by Kenneth J. Arrow, Amartya Sen, and Kotaro Suzumura, 367–389. Amsterdam: Elsevier.

Binmore, Ken. 1994. *Game Theory and the Social Contract. Volume 1: Playing Fair*. Cambridge, MA: The MIT Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract. Volume 2: Just Playing*. Cambridge, MA: The MIT Press.

Cohen, Gerald A. 1988. *History, Labour, and Freedom: Themes from Marx*. Oxford: Clarendon Press.

Cohen, Gerald A. 1990. "Marxism and Contemporary Political Philosophy, or: Why Nozick Exercises Some Marxists More Than He Does Any Egalitarian Liberals." *Canadian Journal of Philosophy Supplementary Volume* 16: 363–387.

Cohen, Gerald A. 1995. *Self-Ownership, Freedom, and Equality*. Cambridge: Cambridge University Press.

Cohen, Gerald A. 2000. *If You're an Egalitarian, How Come You're So Rich?* Cambridge, MA: Harvard University Press.

Cohen, Gerald A. (1978) 2001. *Karl Marx's Theory of History: A Defence*. Princeton, NJ: Princeton University Press.

Curry, Philip A., and John E. Roemer. 2012. "Evolutionary Stability of Kantian Optimization." *Review of Public Economics* 200: 131–146.

Dworkin, Ronald. 1981a. "What is Equality? Part 1: Equality of Welfare." *Philosophy & Public Affairs* 10 (3): 185–246.

Dworkin, Ronald. 1981b. "What is Equality? Part 2: Equality of Resources." *Philosophy & Public Affairs* 10 (4): 283–345.

Elster, Jon. 1978. *Logic and Society: Contradictions and Possible Worlds*. Chichester: John Wiley & Sons.

Elster, Jon. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Forrester, Katrina. 2019. *In the Shadow of Justice: Postwar Liberalism and the Remaking of Political Philosophy*. Princeton, NJ: Princeton University Press.

Hart, Sergiu, and Andreu Mas-Colell. 1996. "Bargaining and Value." *Econometrica* 64 (2): 357–380.

Howe, Roger E., and John E. Roemer. 1981. "Rawlsian Justice as the Core of a Game." *The American Economic Review* 71 (5): 880–895.

Hurley, Susan. 2002. "Roemer on Responsibility and Equality." *Law and Philosophy* 21 (1): 39–64.

Kymlicka, Will. 2002. *Contemporary Political Philosophy: An Introduction*. Oxford: Oxford University Press.

Laffont, Jean-Jacques, and David Martimort. 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton, NJ: Princeton University Press.

Locke, John. (1690) 1988. *Two Treatises of Government*. Edited by Peter Laslett. Cambridge: Cambridge University Press.

Maniquet, François. 2019. "Comments on John Roemer's First Welfare Theorem of Market Socialism." *Review of Social Economy* 77 (1): 56–68.

Morishima, Michio. 1973. *Marx's Economics: A Dual Theory of Value and Growth*. Cambridge: Cambridge University Press.

Morishima, Michio, and Francis Seton. 1961. "Aggregation in Leontief Matrices and the Labour Theory of Value." *Econometrica* 29 (2): 203–220.

Moulin, Hervé, and John E. Roemer. 1989. "Public Ownership of the External World and Private Ownership of Self." *Journal of Political Economy* 97 (2): 347–367.

Müller, Mirjam. Forthcoming. "Who Cares? Market Socialism and Social Reproduction." *Review of Social Economy*.

Nash, Jr., John F. 1950. "The Bargaining Problem." *Econometrica* 18 (2): 155–162.

Nozick, Robert. (1974) 2013. *Anarchy, State, and Utopia*. With a foreword by Thomas Nagel. New York, NY: Basic Books.

Okishio, Nobuo. 1963. "A Mathematical Note on Marxian Theorems." *Weltwirtschaftliches Archiv* 91: 287–299.

Ortuño-Ortín, Ignacio, and John E. Roemer. 1991. "Deducing Interpersonal Comparisons from Local Expertise." In *Interpersonal Comparisons of Well-Being*, edited by Jon Elster, and John E. Roemer, 321–336. Cambridge: Cambridge University Press.

Piketty, Thomas. 2020. *Capital and Ideology*. Translated by Arthur Goldhammer. Cambridge, MA: The Belknap Press of Harvard University Press.

Rawls, John. 2001. *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.

Roemer, John E. 1978. "Neoclassicism, Marxism, and Collective Action." *Journal of Economic Issues* 12 (1): 147–161.

Roemer, John E. 1981. *Analytical Foundations of Marxian Economic Theory*. Cambridge: Cambridge University Press.

Roemer, John E. 1982a. *A General Theory of Exploitation and Class*. Cambridge, MA: Harvard University Press.

Roemer, John E. 1982b. “Exploitation, Alternatives and Socialism.” *The Economic Journal* 92 (365): 87–107.

Roemer, John E. 1985. “Should Marxists be Interested in Exploitation?” *Philosophy & Public Affairs* 14 (1): 30–65.

Roemer, John E. 1986a. “Equality of Resources Implies Equality of Welfare.” *The Quarterly Journal of Economics* 101 (4): 751–784.

Roemer, John E. 1986b. “The Mismarriage of Bargaining Theory and Distributive Justice.” *Ethics* 97 (1): 88–110.

Roemer, John E., ed. 1986c. *Analytical Marxism*. Cambridge: Cambridge University Press.

Roemer, John E. 1988a. *Free to Lose: An Introduction to Marxist Economic Philosophy*. Cambridge, MA: Harvard University Press.

Roemer, John E. 1988b. “A Challenge to Neo-Lockeanism.” *Canadian Journal of Philosophy* 18 (4): 697–710.

Roemer, John E. 1993. “A Pragmatic Theory of Responsibility for the Egalitarian Planner.” *Philosophy & Public Affairs* 22 (2): 146–166.

Roemer, John E. 1994a. *Egalitarian Perspectives: Essays in Philosophical Economics*. New York, NY: Cambridge University Press.

Roemer, John E. 1994b. *A Future for Socialism*. London: Verso.

Roemer, John E. 1996. *Theories of Distributive Justice*. Cambridge, MA: Harvard University Press.

Roemer, John E. 1998a. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Roemer, John E. 1998b. “Why the Poor do not Expropriate the Rich: An Old Argument in New Garb.” *Journal of Public Economics* 70 (3): 399–424.

Roemer, John E. 2002. “Equality of Opportunity: A Progress Report.” *Social Choice and Welfare* 19 (2): 455–471.

Roemer, John E. 2004. “Equal Opportunity and Intergenerational Mobility: Going Beyond Intergenerational Income Transition Matrices.” In *Generational Income Mobility in North America and Europe*, edited by Miles Corak, 48–57. Cambridge: Cambridge University Press.

Roemer, John E. 2010. “Kantian Equilibrium.” *The Scandinavian Journal of Economics* 112 (1): 1–24.

Roemer, John E. 2015. “Kantian Optimization: A Microfoundation for Cooperation.” *Journal of Public Economics* 127: 45–57.

Roemer, John E. 2017. “Socialism Revised.” *Philosophy & Public Affairs* 45 (3): 261–315.

Roemer, John E. 2019a. *How We Cooperate: A Theory of Kantian Optimization*. New Haven, CT: Yale University Press.

Roemer, John E. 2019b. “A Theory of Cooperation in Games with an Application to Market Socialism.” *Review of Social Economy* 77 (1): 1–28.

Roemer, John E. 2020a. “What is Socialism Today? Conceptions of a Cooperative Economy.” Cowles Foundation Discussion Paper No. 2220. Yale University, New Haven, CT.

Roemer, John E. 2020b. “Market Socialism Renewed.” *Catalyst* 4 (1): 8–57.

Roemer, John E. 2020c. “Epistemological Issues in the Theory of Equality of Opportunity (EOp).” Unpublished Manuscript.

Roemer, John E., and Alain Trannoy. 2016. “Equality of Opportunity: Theory and Measurement.” *Journal of Economic Literature* 54 (4): 1288–1332.

Sen, Amartya. 1979. "Utilitarianism and Welfarism." *The Journal of Philosophy* 76 (9): 463–489.

Sen, Amartya. 1985. *Commodities and Capabilities.* Amsterdam: Elsevier.

Sen, Amartya. 2004. *Rationality and Freedom.* Cambridge, MA: Harvard University Press.

Sen, Amartya, Angus Deaton, and Tim Besley. 2020. "Economics with a Moral Compass? Welfare Economics: Past, Present, and Future." *Annual Review of Economics* 12: 1–21.

Stiglitz, Joseph E., Amartya K. Sen, and Jean-Paul Fitoussi. 2009. "Report by the Commission on the Measurement of Economic Performance and Social Progress." Commission on the Measurement of Economic Performance and Social Progress, Paris.

Vrousalis, Nicholas. Forthcoming. "How Exploiters Dominate." *Review of Social Economy.*

Ward, Benjamin. 1979. *The Ideal Worlds of Economics: Liberal, Radical, and Conservative Economic World Views.* New York, NY: Basic Books.

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 177–184.
https://doi.org/10.23941/ejpe.v13i2.533

Reflections on the 2020 Nobel Memorial Prize Awarded to Paul Milgrom and Robert Wilson

MAARTEN C. W. JANSSEN
University of Vienna, National Research University Higher School of Economics (Moscow)

The 2020 Nobel Memorial Prize in Economic Sciences has been awarded to Paul Milgrom and Robert Wilson, both at Stanford University, for improvements to auction theory and inventions of new auction formats. As with any award of the Nobel Prize, this year's prize may raise some questions. For example, what is there to learn about auctions that isn't already intuitively known by most economists or practitioners? Or, shouldn't the Prize instead be awarded to research dealing with big and important questions, such as those related to wealth, poverty, inequality, or the environment? Is the importance of auctions on a par with these topics? What important aspects of economic life do we better understand because of the work of the Nobel Prize laureates? The aim of this essay is to try to answer these questions.

Most people, when they think of auctions, probably think of an art auction where rich people come together to determine who will be the new owner of an exclusive piece of art. Bidders bid against each other, the price goes up, and eventually no one wants to compete against the highest bid. At that point, the auction stops, and the winner is announced as the bidder who has submitted the highest bid and who now must pay. This seems to be not so important or complicated that it deserves a Nobel Prize in economics. So, is there more to it or did the Nobel Prize Committee get it terribly wrong this year?

A first part of the answer to this question (and the questions raised above) is that what goes unnoticed in this example is that the auction mechanism is created. Someone has thought that it would be good to organize it and that an auction is probably better than alternatives, such as a lottery, a beauty contest, or a bargaining process with some selected potential buyers. Put into this perspective, the more general problem is

AUTHOR'S NOTE: The author would like to thank Theo A. F. Kuipers for comments on an earlier version of this article.

an allocation problem; namely, how to allocate certain products or resources and what the objectives that the organizer wants to achieve are. Generally, in economics, we think that markets allocate resources; but, in many circumstances, a market simply does not exist, so one must be created. Markets can, however, not be so easily created. There must be buyers (and sellers) who are willing to participate in the market. This is, in a nutshell, the area of market design to which auction theory belongs.

Once the problem is framed in these more general terms, a large set of potential applications suddenly opens up. Who should have the right to use landing or gate slots at busy airports? Who should have the right to produce electricity at a certain time in a certain country? Who should get the right of using certain spectrum frequencies? How should parking places be allocated in the streets of larger cities? Where should gasoline stations be allocated along highways and who should get the right to exploit these locations? Who should get the right to drill for natural resources in a certain designated area and what conditions do we want to impose on these rights? Who should supply a car manufacturer with windshields? Who should get the right to advertise on an internet page? Or how should we allocate emission permits?

Once it is clear that there is a very large set of potential resource-allocation problems to which auction theory may be applied, new issues emerge. First, context is important. In some cases (such as the right to drill for natural resources) uncertainty is important: when I get the right to drill, for example, I still do not know for sure how much of the natural resource I will be able to extract. In other cases (such as rights to landing slots or spectrum usage), a combinatorial element is important: bidders will only have a use for a landing slot at an airport if they have another slot at another airport where they can depart. In still other cases (for example, with electricity and parking slots), timing is an issue. Depending on the context, the allocation problems may have different aspects. Second, objectives may differ from case to case. In some contexts, it may be natural to think that the objective is to generate (or even maximize) revenue; in other contexts, objectives may include fairness, a division of resources over different bidders, or the creation of a competitive market to maximize consumer welfare after the resources are allocated (such as is the case in spectrum auctions). Auctions have the merit of forcing the organizer of the auction to make the objectives of the allocation process transparent and argue why they think that the choice of a particular mechanism (such as an auction) is the best way to reach these objectives.

The possibility of organizing auctions makes it more difficult (for government authorities) to simply hand over valuable resources to friends or to simply go for historical precedent.

I will now explain that, depending on the context, different complexities arise and that the work of Paul Milgrom and Robert Wilson has proved very fruitful in overcoming these complexities. To do so, and given the space limitations, I present two simple examples that make two of these complexities clear. Milgrom and Wilson's work is, of course, richer than what is depicted in these examples.

WINNER'S CURSE

Let us first consider an example where uncertainty plays a role. Suppose that, in a classroom, I auction off a jar with euro coins of 1, 2, or 5 eurocents each. Students can inspect the jar and estimate how much money it contains; but they cannot count the number of coins. I ask the students to write down their bid on a piece of paper, and I announce that the jar will be won by the highest bidder and that the winner has to pay their own bid. This auction format is a so-called first-price sealed-bid auction and the auction environment is one of common values: the value of the object is the same for every bidder, but there is uncertainty about what that value is. This is clearly relevant in auctions for the right to extract natural resources; but spectrum auctions or art auctions may also have a common-value flavour.

Suppose that there really is 10 euros in the jar, but different students reach different estimates about the value of the jar. Some students may think there is really only 8 euros in it, whereas more optimistic estimates may say there is 12 euros in the jar. Bidders place a bid below what they think the jar is worth, where the amount they bid less than their value depends (among other things) on how many bidders participate in the auction. Importantly, as bidders are unaware of the estimates of other bidders, they can make their bid conditional only on their own estimate of the jar's worth. Suppose that each bidder bids one euro less than what they think the jar is worth. The bidder with the most optimistic estimate wins and pays, say, 11 euros, which is more than the value of the coins in the jar.

This phenomenon is called the *winner's curse*, first studied by Robert Wilson (1969). As more optimistic bidders typically bid higher, the winner is typically the bidder with the most optimistic estimate. After the winner is announced, the winner realizes that she was the most optimistic bidder, and, knowing that the others were more pessimistic, now believes (after her bid is committed and cannot be changed anymore) that she paid too much for the jar. In a series of articles, Wilson (1967, 1969) analysed how to optimally take the winner's curse into account. He showed that the rational response to the winner's curse in a sealed-bid auction is to shade (that is to say, lower) your bid in such a way that the auction generates little revenue. Milgrom (1981), and Milgrom and Weber (1982) extended the pure common-values setting to include common and private values, and showed that open auctions, where bidders get more information during the bidding process, perform better relative to sealed-bid auctions in generating a higher expected revenue. The reason is that other bidders dropping out of the auction gives remaining bidders the opportunity to learn about the estimates of their competitors and to adjust their own estimates. In this way, bidders will be more certain of the true value of the object and will shade their bids much less.

The question of how much information bidders should be allowed to have has been an important consideration in the auction design literature ever since. This is important, not only for generating revenue, but also for the efficiency of the allocation process and for making sure that bidders would like to participate in the auction as they do not want to run unnecessary risks.

MULTI-UNIT AUCTIONS

In many allocation problems, (some) buyers want to acquire either multiple units or nothing. This is clear in the example of airport slots above; but as the present second example will show, it also plays a key role in the spectrum allocation for mobile telecommunications. In these cases, it is clear that multiple objects have to be allocated simultaneously as a bidder may value an object only if she is also able to acquire another object. Consider as an example a country that consists of two regions, *a* and *b*. A national regulator wants to allocate a frequency band and believes that there may be either a national operator interested in acquiring the right to use the frequency band in both regions, or two regional players (one in *a* and one in *b*) who want to use the frequency for regional usages. In large countries such as the USA, Canada, or Russia, licenses are almost always regionally defined; but, similarly, in a recent 5G auction in Europe, some countries also chose the regional format.[1] One of the issues Milgrom

[1] A similar issue arises if one allocates three objects and some players want to acquire two and others only one.

has addressed, both in his academic papers and in his advisory work, is how to design such a multi-unit auction. Together with Robert Wilson and Preston McAfee,[2] they provided important input in designing the Simultaneous Multi-Round Auction (SMRA) in 1994 for the Federal Communications Commission (FCC) in the USA to allocate spectrum rights. It was the first practical auction design where multiple units were allocated simultaneously and similar designs have been used in subsequent years in many countries around the globe.

To illustrate the working of the SMRA and to show that it is not satisfactory in all settings, assume for simplicity that the national operator has a value of around 8 for the two licenses together, and a value of 0 for each regional license separately. The value for the regional players is highly uncertain and can be anywhere between 0 and 5 for the license in their region. The regulator decides to allocate regional licenses instead of national licenses to give the regional players a chance to acquire a spectrum that is useful for them. Detailed rules may differ across different SMRAs. One typical set of rules is as follows. In round one, the auctioneer announces a starting price (say, 1) for both licenses, *a* and *b*, and asks the bidders whether they want to acquire the license(s) at that price.[3] If multiple bidders demand a license, the auctioneer announces one of them as a *provisional* winner for that round (assigned, for example, at random, or to the bidder who was quicker in expressing demand). For all licenses with excess demand, the auctioneer announces a price for the next round (say, a price of 2) and all bidders who were not provisional winners in the previous round can express their demand at the higher price—one of them is then designated a provisional winning bidder.

This is illustrated in Table 1 below: in round one, the regional players (*A* and *B*—with capital letters representing bidders) bid on the license they are interested in and the national player (*N*) bids on both licenses. Suppose that for license *a*, bidder *N* is randomly selected as the provisional winning bidder; and for license *b*, bidder *B* is selected. If the bidders that were not provisional winning bidders come back in the next period at higher prices, then demand remains the same for both licenses, but the provisional winners will now be different.

What is going to happen in round 3? For bidder *B*, the answer is simple. If their willingness to pay is 3 or higher, then they will come back and

[2] See the foreword to Milgrom (2004) where some of the early work on FCC auctions is described.

[3] In the initial versions of the SMRA, bidders themselves could bid on the licenses and the auctioneer did not announce clock prices.

ROUND / PRICE	DEMAND FOR LICENSE *a*	DEMAND FOR LICENSE *b*	PROVISIONAL WINNING BIDDER IN REGION *a*	PROVISIONAL WINNING BIDDER IN REGION *b*
1	*A*, *N*	*B*, *N*	*N*	*B*
2	*A*, *N*	*B*, *N*	*A*	*N*
3	. . .	. . .	. . .	. . .

Table 1: An illustration of the SMRA.

bid on license *b*; otherwise, they will drop out. But will bidder *N* bid again on license *a*? Both options (of continuing to bid, or not) are associated with considerable risk. If bidder *N* does not bid on *a*, they would hope that they are overbid on *b* in round 3, so that they do not acquire only one license (whose value for *N* is zero). But this is not known to *N*. On the other hand, if bidder *N* does bid on *a*, it may be that both bidder *B* will come back in round 3 and bidder *A* will come back in round 4. In that case, *N* will only be able to acquire both blocks for a total price of at least 9, which is higher than *N*'s value. This problem cannot be resolved easily and is due to the fact that bidder *N* only values a full bundle, but in an SMRA they may end up with only part of the bundle. This is called the *exposure problem*, and it makes multi-unit auctions much more complicated than single-unit auctions. Also, bidders may prefer not to participate in the auction in the first place if they fear that the exposure problem is a real risk. This potential non-participation may create significant inefficiencies.

The so-called combinatorial clock auction, invented by Paul Milgrom together with Larry Ausubel and Peter Cramton (Ausubel, Cramton, and Milgrom 2006), addresses this problem. In this auction format, at each round price, bidders announce which licenses they would like to get. This demand is interpreted as an all-or-nothing demand so that if bidder *N* says that, at a price of 2, they demand both licenses, they either get both of them or nothing. The combinatorial clock auction has, however, problems of its own,[4] and currently there is no auction design that works efficiently in all possible multi-object contexts. The work by Paul Milgrom has, however, been important in illuminating the different issues that are relevant for auctioning multiple objects and in sketching the circumstances under which a particular auction design is expected to do best.

[4] For a more detailed treatment of the combinatorial clock auction and some of its problems, see Levin and Skrzypacz (2016), and Janssen and Kasberger (2019).

CONCLUDING

In conclusion, let me go back to the questions I started out with. First, it should be clear by now that, because of the vast number of potential applications in very different areas, it is important to have a better understanding of how different possible auction formats may work and when they can be applied. Efficiency gains that can be obtained through auctions are relevant in many different economic and non-economic situations. If governments use auctions to allocate public assets, they can have a variety of objectives. If they aim at maximizing revenue, then the proceeds can be used (and many governments do use them) for many different social goals.

Second, by discussing two important issues in auction design, I have explained that once one goes beyond standard auctions, such as antique auctions, there are important issues that need to be addressed. These issues are non-trivial and deserve careful analysis. To outsiders, these issues may appear to be 'details'; but often, in auction design, the devil is in the details, and if the details are not properly dealt with, the whole auction design may fail and the goals may not be realized.

Finally, we may ask whether because of the work of this year's laureates we now understand better some aspects of economic life? The answer to this question is more subtle, I think. In many auctions, the bidding data is not publicly available, and even if they are available, we typically do not know the valuations of different bidders and, therefore, cannot investigate why they bid the way they did. The assumptions underlying the theory—namely, that bidders have a clear valuation for the objects, that they only care about what they acquire and at what price, and, therefore, that these valuations typically do not depend on who else wins part of the objects and what others have to pay—cannot be verified. Often, also, we do not know the counterfactual, that is, what the (auction) outcome would have been had a different format been chosen. The success of auction theory seems therefore driven not by being able to better predict behaviour or explain what happens in a particular auction.[5] Rather, success here seems to be related to what Alvin Roth said already some time ago in *The Economic Journal*: "the real test of our success will be […] how well we can bring this knowledge to bear on practical questions of microeconomic engineering" (Roth 1991, 113). More and more, it seems that a proper assessment of economics as a science should not only rely

[5] Progress in auction design seems to have common properties with what Kuipers (forthcoming) describes as progress in concept explication.

on whether it is able to make better predictions, but also on how it is used to design new mechanisms to allocate resources. Together with the 2007 and 2012 Nobel Prizes for mechanism design (to Leonid Hurwicz, Eric Maskin, and Roger Myerson) and market design (to Alvin Roth and Lloyd Shapley), respectively, this year's award testifies to this shift in the economics profession.

REFERENCES

Ausubel, Lawrence M., Peter Cramton, and Paul Milgrom. 2006. "The Clock-Proxy Auction: A Practical Combinatorial Auction Design." In *Combinatorial Auctions*, edited by Peter Cramton, Yoav Shoham, and Richard Steinberg, 115–138. Cambridge, MA: The MIT Press.

Janssen, Maarten C. W., and Bernhard Kasberger. 2019. "On the Clock of the Combinatorial Clock Auction." *Theoretical Economics* 14 (4): 1271–1307.

Kuipers, Theo A. F. Forthcoming. "The Logic of Qualitative Progress in Nomic, Design, and Explicative Research." In *Current Trends in Philosophy of Science: A Prospective for the Near Future*, edited by Wenceslao J. Gonzalez. Cham: Springer.

Levin, Jonathan, and Andrzej Skrzypacz. 2016. "Properties of the Combinatorial Clock Auction." *American Economic Review* 106 (9): 2528–2551.

Milgrom, Paul R. 1981. "Rational Expectations, Information Acquisition, and Competitive Bidding." *Econometrica* 49 (4): 921–943.

Milgrom, Paul R. 2004. *Putting Auction Theory to Work.* Cambridge: Cambridge University Press.

Milgrom, Paul R., and Robert J. Weber. 1982. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50 (5): 1089–1122.

Roth, Alvin E. 1991. "Game Theory as a Part of Empirical Economics." *The Economic Journal* 101 (404): 107–114.

Wilson, Robert B. 1967. "Competitive Bidding with Asymmetrical Information." *Management Science* 13 (11): 816–820.

Wilson, Robert B. 1969. "Competitive Bidding with Disparate Information." *Management Science* 15 (7): 446–448.

Maarten C. W. Janssen is Professor of Microeconomics at the University of Vienna. Before joining Vienna he was a professor of Microeconomics at Erasmus University Rotterdam and director of the Tinbergen Institute. His main academic research area is the theory of industrial organization, where he is known for his work on consumer search and auctions. He created the Vienna Graduate School of Economics, is an elected foreign member of The Royal Holland Society of Sciences and Humanities, a fellow of the Centre for Economic Policy Research (London), and holds an honorary doctorate from the Higher School of Economics (Moscow). He has been awarded the Distinguished Visiting Austrian Chair Professor at Stanford University for the year 2022.
Contact e-mail: <maarten.janssen@univie.ac.at>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 185–190.
https://doi.org/10.23941/ejpe.v13i2.516

Review of Alberto Mingardi's *Classical Liberalism and the Industrial Working Class: The Economic Thought of Thomas Hodgskin.* New York, NY: Routledge, 2020, 160 pp.

Daniel Layman
Davidson College

One of the most exciting recent trends in the history of social and political thought is the attention scholars have begun to pay to non-canonical figures. Thomas Hodgskin, however, has, with a few exceptions (Stack 1998 and Layman 2020, for instance), remained on the sidelines of this expanding field of play. Moreover, to the extent that Hodgskin has received scholarly attention during the last century, it has mostly been to cast him as a fairly minor nineteenth-century socialist (for example, Berlin 2019). In this important new book, Alberto Mingardi sets out both to bring Hodgskin and his accomplishments out of obscurity and to reframe the English journalist, economist, and activist as an important figure in the classical liberal tradition of political economy, stretching from Adam Smith to Friedrich Hayek, and beyond.

In chapter one, we meet Hodgskin, the man—failed naval officer, self-trained economist, journalist, and polemicist. Hodgskin, Mingardi tells us, was born to lower middle-class circumstances in Chatham, Kent, in 1787. After a brief and bitter stint in the Royal Navy that would later inspire his first theoretical work, *An Essay on Naval Discipline* (1813), he matriculated at the University of Edinburgh as a literature student. He never graduated, but he did form new acquaintances that led to his introduction to Francis Place (1771–1854), an important figure in London's radical working-class intellectual circles. With Place's help, Hodgskin launched a journalistic career by joining London's *Morning Chronical* as parliamentary reporter in 1822. It was from this perch as a political and economic journalist that Hodgskin made his principal contributions to political economy, both through his role in the Mechanics' Institute in London and, most significantly, through the three major treatises that he published between 1825 and 1832. In the following four chapters, which together form the core of his book, Mingardi investigates in detail the doctrines, arguments, and intellectual contexts of these three works.

Labour Defended Against the Claims of Capital (1825) contains the core of Hodgskin's polemic against the capitalism of his time. Moreover, it is largely on account of its impassioned defense of economic reform that scholars have cast Hodgskin as a socialist. According to Mingardi, this portrayal is a deeply mistaken one that has long obscured Hodgskin's unique (and important) place in the history of political economy. Its error lies not in reading Hodgskin as an enemy of 'idle' capitalists and landlords but rather in inferring from that enmity a friendliness towards the kind of economic collectivism characteristic of socialism. By Hodgskin's lights, these economic villains are able to oppress the poor only insofar as the state intervenes on their behalf, as it inevitably does. If the state would simply refrain from meddling in economic matters, workers would enjoy the just fruits of their labor, and prosperity and liberty would overcome poverty and dependence. Hodgskin, as Mingardi reads him, is not so much opposed to capitalism per se as to *crony* capitalism in which the government props up the lazy and well-connected at the expense of the industrious common worker. If labor and the practical knowledge that powers it were permitted their proper rewards, there would be no need for any mechanism of distribution apart from the voluntary exchanges among worker-owners. Here, as Mingardi emphasizes, we begin to see the outsized influence of Adam Smith's economic thought on Hodgskin's doctrines.

The Smithian seed that took root in *Labour Defended* came to fruition in Hodgskin's *Popular Political Economy* (1827), to which Mingardi turns in his third chapter. Hodgskin's aim in this work, Mingardi explains, is to offer an account of political economy that is popular not in the sense of being watered-down or accessible but rather in the sense of being written "from the point of view of the people" rather than from the point of view of the crony-capitalist class (66). From this point of view, the fundamental question of political economy is distinctly Smithian: How can diffuse knowledge, including both theoretical understanding and know-how, guide an economic order towards prosperity and independence for the workers who possess and rely on that knowledge? In Hodgskin's political context, the fierce public debate over the protectionist Corn Laws, through which Parliament heavily taxed grain imports to maintain high grain prices for British landowners, constituted the backdrop against which this question, and others like it, arose. According to Hodgskin, the Corn Laws and similar acts of protectionism harm workers by "curb[ing] the spirit of enterprise, and imped[ing] production, by checking the progress of knowledge and the acquirement of skill" (79). For Hodgskin, as for Smith, there is an invisible hand at work in markets, a hand which, if permitted by the state to operate freely, will facilitate both the productive use of individual knowledge by laborers and the distribution of social knowledge through prices.

A similar Smithian spirit pervades Hodgskin's defense of free banking, to which Mingardi turns in chapter four. Parliament had begun in 1825 to increase the power and authority of the state-run Bank of England, partially in response to a banking crisis which had scuttled several large private banks that same year. Hodgskin argued that this policy "confers on […] government a boundless power of working mischief" (105). This power is liable to cause inflation to a degree that would never occur if legislators would simply leave banking and banknotes, like all other facets of economic life, to private markets. Far from providing a way for the rich to ride high at the expense of the poor, private banking was, in Hodgskin's judgment, just one more dimension of the mutually beneficial market society that would emerge if only governors would let it. The state, in the sphere of banking as in other spheres, can only make things worse.

Although the Smithian strain in Hodgskin's economic thought is very pronounced, it coexists with a distinct, though not unrelated, Lockean strain. Mingardi details in chapter five how this dimension of Hodgskin's thinking is foundational to his account of property rights and their natural (as opposed to political) character. In *The Natural and Artificial Right of Property Contrasted* (1832), Hodgskin argues, following John Locke, that workers enjoy natural property rights in whatever they produce through their labor. This is because each person owns herself, and labor extends the self to encompass what her labor produces. Indeed, the conceptual relationship that Hodgskin posits between personal identity and natural property rights is so tight that to doubt the reality and absolute strength of those rights is to be irrational or even "insane" (Hodgskin 1832, 30). Property theory is therefore a kind of demonstrative natural science that has nothing at all to do with politics. Indeed, by Hodgskin's lights, any attempts by governments to pass positive law defining property rights—attempts of the kind endorsed by Jeremy Bentham, the archvillain of *Natural and Artificial*—can be nothing less than irrational and anti-scientific usurpations of natural right: there is, and can be, no artificial right of property at all.

Having considered in detail all three of Hodgskin's primary works, Mingardi concludes with some reflections on Hodgskin's influence. He suggests that, despite Herbert Spencer's claims to the contrary, there is reason to suppose that Hodgskin influenced Spencer, whom he knew personally in the context of his career as a journalist. In particular, Mingardi hears echoes of Hodgskin's voice in Spencer's critique of statism, which, like Hodgskin's own critique, turns on the idea of economic complexity. This idea would, moreover, come to define Hayek's economic outlook nearly a century later. It is in virtue of these relationships that Hodgskin occupies an important place in the development of classical liberalism as we know it today.

Mingardi's book offers the most complete and coherent reconstruction of Hodgskin as a systematic figure so far. Moreover, the book retires—hopefully once and for all—the lazy assumption that since Hodgskin supported working-class causes and opposed the capital structure of his time and place, he must have been a socialist. As Mingardi makes clear, Hodgskin sought not to replace concentrated aristocratic control of property with concentrated democratic control, but rather, to liberate economic life from centralized control of any kind. Far from offering some kind of proto-Marxism, Mingardi's Hodgskin channels the legacies of Smith and Locke into a radical form of anti-statist classical liberalism. Mingardi pierces the interpretive fog of assumed socialism that has crept up around Hodgskin over the decades; consequently, his book merits wide attention, close study, and vigorous engagement.

In the remainder, I'd like to draw out and reflect on a tension between the Lockean and Smithian dimensions of Hodgskin's thought to which Mingardi draws attention. In a Lockean voice, Hodgskin claims that natural property rights form an eternal, complete, and rationally accessible system—that is, a moral science. Property rights are thus radically prepolitical, and there is neither need nor license for legislatures to alter them. But in his more Smithian (and proto-Hayekian) passages, Hodgskin argues that economic relationships emerge from and constitute a spontaneous order that individual minds cannot possibly comprehend. Both of these lines of thought support Hodgskin's anti-statist—and, in particular, anti-legislative—position, but they do so on different and seemingly incompatible grounds. From the Lockean perspective, there is no legitimate legislative task with respect to property rights, because any minimally rational individual can grasp and respect them. But from the Smithian perspective, there is no need for legislatures to meddle in economic matters because spontaneous order settles all economic questions conclusively and to the benefit of all. Despite rendering similar policy prescriptions, it is not immediately clear how these two approaches can accommodate one another.

Mingardi's text suggests a solution rooted in Hodgskin's conception of human life as governed by natural laws. Just as legislators ignore the natural law of individual motivation and action when they fail to recognize the roots of property in the self-owned individual and her labor, they ignore the natural law of human exchange when they attempt to regulate markets or, as they are especially wont to do, give special privileges to particular market actors. The natural law of property creation and the natural law of exchange, however, are not nomological isolates; to the contrary, they constitute, respectively, the micro-level and macro-level dimensions of one and the same comprehensive natural law of human behavior. From the birth of property at laborers' hands to the most complex (private) banking transactions, all economic behavior follows from and expresses the same rational and determinate law.

Although Hodgskin's endorsement of this comprehensive law-governed economic-cum-social science may reconcile his Lockean and Smithian commitments, it raises two problems. First, it seems to commit Hodgskin to a kind of determinism about human action that is nowhere present in either Locke or Smith and which threatens human freedom and responsibility. Second, insofar as Hodgskin attempts to derive his political conclusions from what naturally *does* happen in human life, he threatens to crowd out the normativity necessary for the judgments about what *should* happen in human life on which those conclusions depend. I very much hope that scholars will take up these problems in Hodgskin, and that they will rely on Mingardi's fine book as they do so.

REFERENCES

Berlin, Isaiah. 2019. "Socialism and Socialist Theories." In *The Sense of Reality: Studies in Ideas and their History*, edited by Henry Hardy, 96–145. Princeton, NJ: Princeton University Press.

Hodgskin, Thomas. 1832. *The Natural and Artificial Right of Property Contrasted*. London: B. Steil.

Layman, Daniel. 2020. *Locke Among the Radicals: Liberty and Property in the Nineteenth Century*. New York, NY: Oxford University Press.

Stack, David. 1998. *Nature and Artifice: The Life and Thought of Thomas Hodgskin (1787–1869)*. Rochester, NY: Boydell and Brewer.

Daniel Layman is Assistant Professor of Philosophy at Davidson College. He is the author of *Locke Among the Radicals: Liberty and Property in the Nineteenth Century* (New York, NY: Oxford University Press, 2020) and the co-author (with Michael Huemer) of *Do Governments Have Moral Authority? A Debate* (New York, NY: Routledge, forthcoming).
Contact e-mail: <dalayman@davidson.edu>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 191–196.
https://doi.org/10.23941/ejpe.v13i2.517

Review of Janek Wasserman's *The Marginal Revolutionaries: How Austrian Economists Fought the War of Ideas*. New Haven, CT: Yale University Press, 2019, 354 pp.

OLA INNSET
National Library of Norway

The starting point for Janek Wasserman's new group history of the Austrian School of Economics is the appearance of Friedrich Hayek's *The Road to Serfdom* ([1944] 2007) at the top of Amazon's bestseller list in June 2010. The spike in sales for Hayek's anti-collectivist tract, some sixty-six years after its original publication, was due to the radical-right TV personality, Glenn Beck, who devoted an hour-long episode of his Fox News program to the book. The attention from Beck coincided with the Tea Party movement, the fallout of the 2008 financial crash, and the rise to fame of Republican politicians like Ted Cruz, Paul Ryan, and Ron Paul—the latter of which exclaimed, "we are all Austrians now", after winning the Republican presidential caucus in Iowa in 2012.

Wasserman's first publication was the acclaimed 2014 monograph *Black Vienna: The Radical Right in the Red City, 1918–1938*. Thus, he is in a privileged position to connect the current resurgence of the far-right in the USA and a school of economic thought originating in the Habsburg Empire of nineteenth-century Central Europe. In particular, Wasserman's ability to turn the loose and variated history of 'Austrianism' in both Austria and beyond into a cohesive narrative is impressive. It presents the idea of the 'Austrian School' without losing nuance or missing internal disagreements and other contingencies which punctuate its unique history. The Austrian School includes thinkers and characters as diverse as Friedrich von Wieser, Ludwig von Mises, Joseph Schumpeter, Oskar Morgenstern, Friedrich von Hayek, and Murray Rothbard. Wasserman does a great job at showing the reader both the differences and the many instances of common ground between them all.

The first four of the seven chapters are set in Vienna, and we learn about what Wasserman calls "The Prehistory and Early Years of the Austrian School" (roughly 1870–1890), "The Golden Age" (1890–1918), "Austria's End" (1918–early 1930s), and "Depression, Emigration, and Fascism" (the war years up until 1945). These chapters give a dense and ultimately satisfying account of the creation and development of an Austrian School of Economics. At the outset, the Austrian School was characterized by the development of marginal utility and Carl Menger's scornful attacks on the German Historical School. The famous *Methodenstreit* is given a rather sparse retelling, which perhaps leaves something to be desired for those seeking new insights on this seminal controversy about the relationship between theory and empiricism in economics. This could also be said of Wasserman's later retelling of both stages of the Socialist Calculation Debates, which are brief and not very novel. In this way, this book does not appear to be directed towards experts on these controversies or even to specialists on the history of economic thought. Rather it is written to uncover the meaning and context of the Austrian School for those who aren't already well-versed in intellectual debates over how to interpret this tradition—a category which includes the vast majority of academic economists, historians, and of course, the wider public.

Wasserman's retelling of the birth of the Austrian School and its role in Austria for some sixty years serves the primary function of contextualizing a school of thought which has recently popped up among far-right activists and politicians both in the USA and in Europe. References to both Hayek and Mises have been frequent in these circles for some time, in part due to the influence of various institutes and think tanks established in both their names. In the first chapters of the book, Wasserman delivers lively portraits of these bright, upper-class men and their involvements in imperial politics during the liberal golden age of the Habsburg Empire. He then moves on to post-empire Austria and the next generation's organizing of private seminars and research institutes financed by the Rockefeller Foundation, accompanied by drinking songs in the coffee houses and restaurants of the Ringstrasse. All of this is perhaps covered even better in, for instance, Erwin Dekker's *The Viennese Students of Civilization* (2016). But Wasserman's account is in many ways a set-up for what happens next: the intriguing story about what transpired when the events of World War II saw these people transplanted to the entirely different context of post-war USA.

In some ways, Wasserman's story is one of progressive radicalization from the 1870s up to the present moment. Socialists, Marxists, and other leftists were always antagonists for the Austrians, and even Carl Menger's spats with the German Historicists touched upon the subject of the role of the state in the economy. Eugen von Böhm-Bawerk was the School's leading light after Menger and produced one of the earliest and most serious critiques of Marx's *Capital*, the 1898 volume *Karl Marx and the Close of His System* (Böhm-Bawerk [1896] 1949). Yet, early Austrians like Böhm-Bawerk accepted both Marxists and other socialists in their seminars, which included Austro-Marxists like Otto Neurath and Rudolf Hilferding, and even the Russian Bolshevik Nikolai Bukharin. Early members of the Austrian School were not averse to social policies, and someone like Friedrich von Wieser also displayed sympathies for social democracy. Disagreements with socialists on friendly terms appear to have ended with Mises, who made his fame as a critic of socialism in the 1920s. Wasserman claims that Mises already had a more dogmatic approach in Vienna, and that he refused to admit those who did not share his political views into his seminar.

After arriving in the USA during the war, Austrians like Hayek, Schumpeter, Morgenstern, Fritz Machlup, and Gottfried Haberler were more immediately successful than Mises. Like Quinn Slobodian in *Globalists* (2018)*,* Wasserman highlights the stories of Machlup and Haberler, who are less famous than Hayek and Mises today, but who were arguably more influential: Haberler, within the new institutional structure of international trade policy around the General Agreement on Trade and Tariffs (GATT), and Machlup, as the leader of the Bellagio group, who advised the influential Group of 30 (G30) during the build-up to the end of the Bretton Woods agreement.

All the Austrians were hugely successful in securing positions and patronages in their new circumstances, something that was often belied in their own self-presentation as outcasts and misunderstood geniuses—stories of which have then been repeated by sympathetically inclined biographers. The conventional narrative of the Austrian economist being 'in the wilderness' post-emigration is thoroughly debunked by Wasserman, who writes:

> When they were not presiding over the American Economic Association (AEA), the Econometric Society, the International Economics Association (IEA), or National Bureau of Economic Research (NBER), the US government called on their expertise, candidates from conservative parties in the United States and United Kingdom sought their advice, and the International Monetary Fund (IMF), the General Agreement on Tariffs and Trade (GATT), and the World Bank solicited their opinions on trade and monetary matters. (235)

On top of this came the lavish funding from business interests sympathetic to the Austrians' anti-left politics. This was especially the case with Mises, who was on a yearly honorarium from the National Association of Manufacturers (NAM) as a consultant, was paid a stipend from Leonard Reed of the Foundation for Economic Education, and had his salary as a visiting professor at NYU covered by private contributions. Younger Austrians in exile, like Haberler, Machlup, Morgenstern, and indeed even Hayek, tended to see Mises' laissez-faire views as outdated, and his whole approach to economics and politics as highly dogmatic. However true, that was perhaps also what became Mises' attraction in his new country. Like Mises, Hayek moved away from pure economics and became more of a political philosopher, also engaging in activism through the Mont Pèlerin Society and the various think tanks and foundations that grew out of it. Nonetheless, Hayek's somewhat more nuanced views, and the bestowal of the 1974 Bank of Sweden prize upon him, made him into less of a fringe figure than Mises.

The latter's ideological fervor and lack of mainstream recognition instead led to the development of a 'heterodox' school of American economists devoted to Mises' ideas. This group ended up diverging in significant ways from the more diverse group of actual Austrians who had taken part in Mises' seminars in Vienna. A telling anecdote told by Wasserman relates to a planned birthday celebration for Mises, in which his Austrian and American friends and followers disagreed markedly on who ought to be invited.

Another angle on this important aspect of the story is given through Wasserman's dive into Fritz Machlup's attempt in the early 1980s of writing the history of the Austrian School. Drafts of Machlup's paper were shared with other Austrians, all of whom found it difficult to know who to include in the School. Machlup ended up making a distinction between 'Austrians', like himself, Hayek and Mises, who had been in Vienna and contributed to the development of the School; those he called 'un-Austrian Austrians', a category referring to people like Morgenstern and Schumpeter who were clearly from the Austrian School but who nonetheless departed from some of the teachings and followed diverging intellectual paths; and lastly, 'non-Austrian Austrians', a label designating the group of Americans worshipping at the altar of Mises. The various splits within this faction are covered well by Wasserman towards the end of the book. The most enduring split is perhaps that between overt racists and far-right activists building on Murray Rothbard's legacy in various Ludwig von Mises Institutes, and scholars connected to the Koch-funded Mercatus Institute at George Mason University, who took a hermeneutic turn, and ended up proposing a revamped Hayekianism (Boettke 2018).

One part of this mosaic, which is somewhat missing, is the so-called Virginia School of Constitutional Economics. James M. Buchanan and colleagues are certainly mentioned here and there; but the way in which this school specifically builds on Austrian ideas is not given much attention. Perhaps this is what gets lost in Wasserman's contention that Austrian economics eventually came to signify the sect built around Mises in the USA, and its various splits and off-shoots. All in all, Wasserman's masterful book paints a much needed critical yet scholarly picture of the Austrian School. His book is not a polemic against the Austrians; but unlike many of the accounts written by people personally connected to the School, he brings attention to these thinkers' privileged backgrounds and lifestyles, their fundamentally elitist politics, and the important connections to wealthy benefactors with clear political agendas. The last part of the book focuses on disagreements between far-right 'Austrians' and their more centrist counterparts, both in the USA and in Europe. Splits and conflicts have indeed taken place, also in Europe, where a large portion of the members of the German Hayek Institut recently ceded from the organization in protest to the close ties of some members to the far-right political party Alternative für Deutschland. Wasserman has therefore succeeded in demonstrating that the contested legacy of the Austrian School bears direct relevance to an understanding of modern politics.

REFERENCES

Boettke, Peter J. 2018. *F. A. Hayek: Economics, Political Economy and Social Philosophy.* London: Palgrave Macmillan.

Böhm-Bawerk, Eugen von. (1896) 1949. "Karl Marx and the Close of His System." In *Karl Marx and the Close of His System by Eugen von Böhm-Bawerk and Böhm-Bawerk's Criticism of Marx by Rudolf Hilferding*, edited by Paul M. Sweezy, 1–118. New York, NY: Augustus M. Kelley.

Dekker, Erwin. 2016. *The Viennese Students of Civilization: The Meaning and Context of Austrian Economics Reconsidered.* Cambridge, MA: Cambridge University Press.

Hayek, Friedrich A. (1944) 2007. *The Road to Serfdom.* Edited by Bruce Caldwell. Chicago, IL: Chicago University Press.

Slobodian, Quinn. 2018. *Globalists: The End of Empire and the Birth of Neoliberalism.* Cambridge, MA: Cambridge University Press.

Wasserman, Janek. 2014. *Black Vienna: The Radical Right in the Red City, 1918–1938.* Ithaca, NY: Cornell University Press.

Ola Innset holds a PhD in history and civilization from the European University Institute. His most recent books are *Reinventing Liberalism: The Politics, Philosophy and Economics of Early Neoliberalism (1920–1947)* (Springer, 2020) and a Norwegian monograph on the market turn in Norwegian politics entitled *Markedsvendingen: Nyliberalismens historie i Norge* (Fagbokforlaget, 2020).
Contact e-mail: <ola.innset@nb.no>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 197–202.
https://doi.org/10.23941/ejpe.v13i2.521

Review of Cheryl Misak's *Frank Ramsey: A Sheer Excess of Powers*. Oxford: Oxford University Press, 2020, 500 pp.

DAVID C. COKER
George Mason University

It would appear that Frank Ramsey is about to become famous, all over again. In the recent, and excellent, critical biography of Keynes by Zachary Carter (*The Price of Peace*), Ramsey is invisible—the index lists a single reference to him. The mention is that, along with Bertrand Russell, this "young Cambridge philosopher" helped usher Wittgenstein's groundbreaking *Tractatus Logico-Philosophicus* into print in English (Carter 2020, 113). This may be the last time Ramsey's role in the goings-on at Cambridge at this time can be so overlooked. Ramsey appears poised to step once more into the spotlight. And the reason is the new biography by Cheryl Misak, *Frank Ramsey: A Sheer Excess of Powers*.

There is a certain attractiveness to the idea of *genius.* How could Keats have written such a quantity of immortal poetry, and yet have died at twenty-five? In the same vein, we might ask how did Ramsey manage to make major and enduring contributions to so many disciplines, and have died just short of his twenty-seventh birthday? That is the teaser for this book: an up-close and personal look at genius in action. Thus, the book moves in two dimensions: it is a regular biography, but it is also a chronicle of how Ramsey's ideas developed, and how they influenced (and were influenced by) the major thinkers who were his compatriots at Cambridge University in the nineteen twenties, between the world wars.

Misak asks, rhetorically, how can there be much of a biographical story considering how young Ramsey was when he died? and, how intensely involved must he have been—given the level of his achievement—with purely intellectual pursuits? Yet the biographical portions of the book turn out to be immensely interesting. This is partly because his life was surprisingly eventful, but also because his social and intellectual interactions included a host of vital players of the time: the philosophers Bertrand Russell, G. E. Moore, and Ludwig Wittgenstein; the economists J. M. Keynes, Arthur Pigou, and Piero Sraffa; and the mathematicians G. H. Hardy and J. E. Littlewood. One might add to this list a number of individuals from the Bloomsbury group (in addition to Keynes of course). Despite his youth, Ramsey was on a level playing field and even dominated his interactions with many of these thinkers.

Frank Ramsey was born on February 22, 1903. He was born into a family that Noel Annan, a provost at King's College, Cambridge, called part of the "intellectual aristocracy" (Annan 1955). His father, Arthur Ramsey, was a mathematics Fellow and President of Magdalene College, Cambridge, and his mother, a social reformer, held a degree from Oxford. Arthur apparently was quiet and austere, and curiously (considering Frank's mathematical gifts) not the model for young Frank, while his mother came from an athletic and outgoing family. Intellectual achievement extended through the rest of the children as well: Frank was the oldest; Frank's younger brother, Michael, became Archbishop of Canterbury; Bridget became a physician; and Margaret, an Oxford economics don. The debates between the devout Michael and the atheistic Frank would continue, as friendly banter, throughout their lives. (Unfortunately for Michael, Frank had allies in both of his sisters.) Ramsey's early school years are covered in surprising, and surprisingly interesting, detail. Academically he was marked by a general precocity, perhaps most notably in mathematics, which in his later years was always referred to as the "work" (50, 85, 91) he felt he should be doing, but worried he was constantly neglecting. This was at least partly because his interests ranged so widely; but it also reflected a particularly intense immersion, by his college years, in philosophy (though he was first 'wrangler' in mathematics when his examination was revealed). Philosophy is thus one of the central concerns in the book. Cambridge was a hotbed of philosophical thinking at the time, with G. E. Moore and Bertrand Russell already significant presences there, and with Wittgenstein on the horizon. It is Wittgenstein's intricate relationship with Cambridge that has much to do with Ramsey's, and this book's, story.

For the economic reader, there is the interest in Ramsey's relationship with Keynes (and Ramsey's rooming neighbor, Pigou). Keynes saw the promise in Ramsey almost immediately and brought him over to King's College (at Cambridge). In the exclusive economics club that Keynes oversaw (the meetings of which were held in Keynes' rooms), Keynes and Ramsey frequently had the last word. When Keynes had a tricky math issue before him, Ramsey was consulted. Keynes in a letter called Ramsey "certainly far and away the most brilliant undergraduate who has appeared for many years in the border-county between Philosophy and Mathematics" (112). When Keynes published his book on probability in 1921, it was met with a chorus of praise. Keynes had said in conversation that probability was, at that time, on the level of astrology. He was out to rectify the problem. Bertrand Russell was typical when he called the book "undoubtedly the most important work on probability that has appeared for a very long time", a book which "it is impossible to praise too highly" (113). Ramsey was not so sure. He reviewed the book, critically. And more than a year later Ramsey presented a paper to The Apostles, where he also offered a differing view of the subject. Ramsey thought Keynes claimed too much for induction (for instance, that it could be grounded in the "limited variety of properties in nature", 115). He also thought Keynes' belief in the logical properties of probability was ungrounded. Ramsey reserved this "frequency" basis of probability (Keynes' emphasis, 117) for physics. But Ramsey was more interested in the subjective side of probability, the side he felt characterized human agency. He would go on to write an important paper on the topic. Keynes rebuffed most criticism of his book, but Ramsey's troubled him. "Keynes had been hearing from almost everyone that the *Treatise* was a major achievement until it fell apart under the criticism of his favourite undergraduate" (118). From this point on, Ramsey was Keynes' go-to person for vetting papers for *The Economic Journal*.

Misak is good on these intellectual arguments, as she is a more-than-adequate storyteller regarding the biographical elements. But she is herself a professor of philosophy, and it is the philosophical dimensions that are the true *raison d'être* of the book. The considerable biographical detail will hold the interest even of the casual reader—for instance, one who is merely reading to fill in gaps about the Bloomsbury group. I, for one, found all that material extremely diverting. But for the reader who has even a passing interest in analytical philosophy, or better, pragmatism, the book is a real find. Misak has an argument to make, one she has been honing over the course of several books. (To more fully investigate her arguments, I recommend her 2016 *Cambridge Pragmatism: From Peirce and James to Ramsey and Wittgenstein*, as well as her earlier 2013 *The American Pragmatists*. Both are exciting reads even for the non-philosopher.) That argument revolves around Ramsey picking up on elements from the American, Charles Sanders Peirce, and transforming and building upon them. This incorporating of American pragmatism into Ramsey's thinking in turn exerted pressure on the analytical philosophy then practiced at Cambridge (Moore and Russell), striking at the heart of some of their foundational propositions. But, the most interesting dynamic for Misak is Ramsey's influence on Wittgenstein, which she takes to be considerable, even definitive. Ramsey and Wittgenstein had an extended and intensive interaction, over almost all of Ramsey's Cambridge years. It was Ramsey, at the tender age of eighteen, who was tapped to translate Wittgenstein's *Tractatus Logico-Philosophicus* (a title coined by Moore). He was, apparently, the only person who had the requisite knowledge of German, and the philosophical insight, sufficient for the task. His challenges to Wittgenstein, Misak argues, began to change Wittgenstein's views. And it is one of the tragic dimensions of Ramsey's early death that this influence, which Misak feels was guiding Wittgenstein in a productive direction, ceased to be.

But, as the above paragraph on Keynes and probability hopefully indicates, there is much for the economist in these pages as well. Ramsey published two mathematical papers in Keynes' *The Economic Journal* (in reality, two separate parts of one original lengthy paper). Misak calls them "founding ideas of the sub-disciplines of optimal taxation and optimal saving" (126). They still have currency. One of these papers—"A Mathematical Theory of Saving" (1928)—has been described by Partha Dasgupta as "one of the dozen most influential papers of the twentieth century" (314). It has come down to us as the Ramsey–Cass–Koopmans model. The other—"A Contribution to a Theory of Taxation" (1927)—has been called by Stiglitz "a landmark in the economics of public finance" (2015, 235).

I have given the impression that this is a book heavy on the academics, but that should be qualified. There is a lot of talk about Bloomsbury, and therefore about sexual freedom and the many instances that made up that freedom. There is unhappiness (Ramsey struggled coming into his own, as a sexual person). There is much talk of psychoanalysis, and of extended stays in Vienna to imbibe the same. There is, as anyone who has read Pigou's biography might guess, a great deal of walking in the mountains. There is Arthur's (Ramsey's father) tragic inattentiveness while driving his car. And there is Ramsey living openly with a woman while (mostly) maintaining his happy marriage. So, it isn't all just philosophy.

But the philosophy does make for compelling reading. And for the very academically inclined, Misak asks a handful of specialists to write on topics she feels less confident about, in a series of boxes scattered throughout the text. But on her own ground, Misak generates intellectual excitement: Ramsey, after completing the translation of Wittgenstein's *Tractatus,* reviews it. For Misak, that review:

> […] still stands as one of the most important commentaries. Indeed, we will see that Ramsey's persistent objections to the theory of meaning and truth set out in the *Tractatus* were largely responsible for Wittgenstein's turn away from the *Tractatus* and towards what we think of as the later Wittgenstein. This was one of the most important shifts in the history of philosophy. Wittgenstein was himself largely responsible for the way philosophy unfolded in Cambridge and beyond. Ramsey's book, had it been completed, might have reset this major […] course of philosophy. (xxvi)

So, the story of Ramsey and his influence is still being written. Ramsey's book referred to above didn't appear until the 1990's! The piecemeal publication of many of Ramsey's articles, and his early death, have somewhat dulled his impact. This has resulted in what one theorist dubbed "the Ramsey Effect" (xxv), where one's newly minted insight is found to have been thought of by Ramsey decades before. There is much that hasn't been mentioned. On the personal side, there is the interaction with the obsessed and often distressing Wittgenstein (waiting for his arrival in 1929, Keynes wrote to his wife: "Ludwig … arrives tomorrow … Pray for me!", 346). On the academic side, there are connections with the Vienna Circle, and through them or Sraffa or some other conduit, to influence on von Neumann and Morgenstern, and their joint book *Theory of Games and Economic Behavior.* One might even wonder if Ramsey's subjective critique of Keynes' probability book had outcomes later, in the psychological and philosophical dimensions of *The General Theory.* But to speculate on these and other questions one must read this excellent book.

REFERENCES

Annan, Noel G. 1955. "The Intellectual Aristocracy." In *Studies in Social History: A Tribute to G. M. Trevelyan*, edited by John H. Plumb, 256–283. London: Longmans.

Carter, Zachary D. 2020. *The Price of Peace: Money, Democracy, and the Life of John Maynard Keynes.* New York, NY: Random House.

Misak, Cheryl. 2013. *The American Pragmatists.* New York, NY: Oxford University Press.

Misak, Cheryl. 2016. *Cambridge Pragmatism: From Peirce and James to Ramsey and Wittgenstein.* New York, NY: Oxford University Press.

Ramsey, Frank P. 1927. "A Contribution to the Theory of Taxation." *The Economic Journal* 37 (145): 47–61.

Ramsey, Frank P. 1928. "A Mathematical Theory of Saving." *The Economic Journal* 38 (152): 543–559.

Stiglitz, Joseph E. 2015. "In Praise of Frank Ramsey's Contribution to the Theory of Taxation." *The Economic Journal* 125 (583): 235–268.

David C. Coker is a Bradley Fellow and PhD candidate in Economics at George Mason University, and is currently teaching as adjunct at the University of Maryland, Baltimore County. His paper on John Rawls and economics is forthcoming in *RHETM*, and his paper (with Ross Emmett) on Frank Knight and James Buchanan is forthcoming in *Œconomia*.
Contact e-mail: <dcoker2@masonlive.gmu.edu>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 203–213.
https://doi.org/10.23941/ejpe.v13i2.534

Review of Valentin Beck, Henning Hahn, and Robert Lepenies' (eds.) *Dimensions of Poverty: Measurement, Epistemic Injustices, Activism.* Cham: Springer, 2020, 412 pp.

S. SUBRAMANIAN
Independent Scholar

This book is the second volume in a series titled *Philosophy and Poverty*, the objective of which, we are told, is to provide "a forum for the whole range of philosophical research on poverty and poverty alleviation, broadly construed" (ii). As such, the project is a large one, in both scope and ambition. The essays in the present volume represent a selection from the proceedings of the Conference on Dimensions of Poverty held in Berlin in 2017, and are intended to cover the themes of "Measurement, Epistemic Injustices, Activism". Apart from an editorial introduction, the book features twenty contributions distributed across five thematic concerns: "Poverty as a Social Relation", "Epistemic Injustices in Poverty Research", "Philosophical Conceptions in Context", "Measuring Multidimensional Poverty", and "Country Cases". One must expect that the nature, quality, and range of the essays collected in the book must inevitably be dictated by the papers presented at, and available from, the conference from which they have originated, which presumably constitutes a model different from one in which a book's editors have the freedom of commissioning papers. The first model has limitations not shared by the second one, and I imagine this must be borne in mind when judging whether what I have referred to as the 'scope and ambition' of the project are quite realised in the volume under review.

The editorial introduction is a useful review of the themes sought to be addressed in the book, and of the specific chapters in it. Parts I and III ("Poverty as a Social Relation" and "Philosophical Conceptions in Context", respectively) are perhaps best read together as reflecting a collection of concerns on some conceptual issues underlying the phenomenon of poverty. Part II, on "Epistemic Injustices in Poverty Research", consists of a set of four papers that must be welcomed—at least for their motivating intention—as suggesting engagement with concerns that are not part of the common currency of poverty studies. Parts IV and V, on multidimensional poverty measurement and country case studies, respectively, are relatively more 'mainstream' aspects of contemporary poverty analysis.

Parts I and III of the book, comprising six essays, are—as mentioned earlier—thematically connected by a concern, broadly, with philosophical approaches to answering the question 'what is poverty?' Among themselves, the relevant papers cover a set of well-worn issues, including absolute poverty, relative poverty, the capabilities approach to poverty assessment, poverty as a social relation, the place of social networks and social capital in understanding poverty, the enhanced effectiveness of aid when it is participatory and collaborative, and the importance of human rights as a normative guide to philanthropic impulse. Much of this work is unexceptionable, but not particularly novel, nor arresting. Indeed, one emerges from these reflections in a spirit of some jadedness, which is perhaps excusable in light of this sort of observation: "[…] human rights are morally important" (150). There is a suggestion in much of this of what one might call 'Poverty for Moral Dummies'. The complaint is not so much with the authors as with the apparent continuing need to address poverty in these terms for those that might still be interested in the subject, or worse, work on it. Of the more important conceptual issues undergirding the notion of poverty on which Amartya Sen has written are those pertaining to the 'capability perspective' and the distinction between absolute and relative poverty. On the capability approach, I am unable to see in this volume much advancement on, or useful alternative or complement to, what Sen has already said on the subject. On the distinction between absolute and relative poverty, some of the contributions have made me wonder if I have myself ever properly understood Sen on this important question.[1]

Certain issues in these philosophical reflections on poverty which I missed—and in no particular order of perceived importance—are the following.

(1) *An assessment of the axiomatic bases of poverty measurement.* In both unidimensional and multidimensional poverty measurement, there has been a long tradition of rationalizing poverty indices in terms of the axioms on which they have been built. As it happens, virtually every one of these axioms—focus,[2] normalization,[3] symmetry,[4] continuity,[5]

[1] Sen (1979).
[2] Hassoun (2014).
[3] Basu (1985), Subramanian (2009b).
[4] On the 'anonymity' principle in 'liberal constitutions', see Loury (2000).
[5] Donaldson and Weymark (1986), Atkinson (1987).

transfer,[6] decomposability,[7] scale invariance,[8] and replication invariance[9]—has been subjected to scrutiny in the literature. An assessment of the ethical appeal and logical coherence of the axiomatic basis of aggregation in the poverty measurement literature would have been a welcome and integral component of any critical treatment of 'philosophy and poverty'.

(2) *A more nuanced appreciation of the role of income in poverty assessment.* As is well-known, and following principally on Sen's contributions to the conceptualization of poverty, there is now a substantial body of thought which prioritizes a 'capability' approach to poverty assessment over a 'resourcist' approach. It would be fair, I think, to suggest that the volume under review reflects this priority. This is not to say at all that the contributors to the volume have taken a uniformly dismissive view of the role of income (or more broadly, resources) in accounting for poverty, as evident, for example, in Jonathan Wolff: "As I have suggested, income adjustment, and, perhaps, the availability of new financial products is a very helpful way of addressing poverty" (37); or, again, Sanjay Reddy: "There is no necessary conflict between having a concern with the avoidance of income poverty and with recognizing that there are diverse non-income concerns that must enter into poverty assessment, too" (217). Having said this, I believe it is possible that the volume might have achieved a greater and more subtle balance of perceptions if it had reflected the sort of critique of the 'capability perspective' that philosophers like Thomas Pogge[10] have attempted in the past, wherein it is suggested that capability theorists have perhaps exaggerated the contrast between resources and capabilities in favour of the capability interpretation. Such an approach might also have facilitated an interesting consideration of the philosophical dimensions of a very simple measure of money-metric well-being that has been advanced by the economist Kaushik Basu (2001, 2006, 2013), which he calls 'the quintile income statistic' (the average income of the poorest 20 per cent of a population). The quintile income statistic offers the interesting possibility of viewing income not just as a means to an end (as in the usual 'identification-cum-aggregation' approach to poverty measurement), but as an end in itself, so that command over a reasonable level of income is seen as a valued human functioning

[6] Chateauneuf and Moyes (2005).
[7] Kanbur (2006).
[8] Kolm (1976a, 1976b), Krtscha (1994).
[9] Hassoun and Subramanian (2012).
[10] Pogge (2010a). See also Kelleher (2015).

in and of itself. This is the sort of perspective that could lead logically to a consideration of 'basic income' as a policy instrument for the alleviation of poverty, and to political-economy considerations of feasibility, fiscal deficits, and the means of financing income support for the poor through, among other things, enhanced redistributive taxation. Philanthropy is discretionary, but taxation—of wealth and inheritance, for instance—is mandatory, and a more forceful means of seeking justice in the presence of undeserved want.

(3) *Poverty and social groups*. I am not suggesting an absence of this thematic concern in the volume, so much as underlining the desirability of an altogether more active and explicit presence of such concern in a disquisition on philosophy and poverty. What is the political economy of caste, race, and ethnicity in an understanding of poverty?[11] How is group membership exploited and conflict fostered by the motive of establishing ownership over resources?[12] How are marginalized groups (typically tribals and forest-dwellers) subjected to even greater oppression than would be warranted by the endemic prejudice that obtains against them, when local governments collaborate with multinational corporations in appropriating control over natural resources?[13] How might 'group-affiliation' externalities mediate the measurement of poverty?[14] How might inter-*group* inequalities in the distribution of poverty call for targeting strategies that are different from what might be dictated solely by a concern with inter-*personal* inequalities?[15] How might poverty-alleviation measures be influenced by an engagement with the ubiquitous but often-missed presence of 'horizontal inequality'?[16] What values come into play in an assessment of poverty which is systematically informed by inter-group differentials in the societal distribution of burdens and advantages?[17] These are some at least of the questions of philosophical interest that a concern with social groups should engender in analysing the phenomenon of poverty.

Part II of the volume deals with "epistemic injustices" in poverty research. Franziska Dübgen makes the point that epistemic and cognitive

[11] See, among others, Sundaram and Tendulkar (2003), Thorat and Newman (2010), Ashwini Deshpande (2011, 2013), Motiram and Singh (2012), and Satish Deshpande (2013).

[12] Mitra and Ray (2014).

[13] See, among others, Padel and Das (2010), Pogge (2010b), and Karat (2012).

[14] Subramanian (2009a).

[15] Keen (1992), Dasgupta and Kanbur (2005).

[16] Stewart (2005).

[17] See, for example, Dworkin (1977), and Galanter (2002).

injustices are reflected in the marginalization of Southern researchers; in making 'others' the object rather than subject of research; in the neglect of subaltern knowledge and experience; and in the misconstruction of social reality. Jonathan Chimakonam specifically highlights the neglect of the Kenyan philosopher Odera Oruka and his theory of 'the human minimum' in the canvas of poverty studies. Sharon Adetutu Omotoso draws on the discourse on hair to draw a distinction between the 'Hairy' (associated with "Scholarly African Feminists" [SAF], 122) and the 'Hairless' (associated with "Indigenous-Survivalist African Feminists" [I-SAF], 122), and to focus attention on one aspect of deprivation that is frequently missed out in poverty studies, which she calls 'intellectual poverty'. Intellectual poverty is seen by the author to be a feature of both the 'Hairless' (the relatively less privileged materially deprived African women) and the 'Hairy' (the relatively more privileged, elitist class of African women, often with stable and well-paying jobs), but in different ways. For the 'Hairless', intellectual poverty is manifested as ignorance and lack of access to information which leads to irrationality, myopia, and reduced skills in problem-solving. For the 'Hairy', intellectual poverty is manifested in a lack of knowledge of, and attendant empathy for, the predicament of their 'Hairless' sisters. The resulting conflict between the SAF and I-SAF groups, stemming from intellectual poverty of one type or the other, is seen as being needless and unfortunate, and an impediment to combating the feminization of poverty. Finally, Mitu Sengupta considers the 'post-development' critique of poverty studies, in which much of poverty research from the Global North is perceived as being so shot through with epistemic injustice as to offer little hope for remediation in the form of less insular and more inclusive modes of understanding poverty. To address the question of whether academics (especially Western academics, as one understands) should engage in activism, the author reviews the post-development critique, in a somewhat autobiographical vein, by tracing her own association with poverty research under the aegis of the ASAP (Academics Stand Against Poverty) project. From what one can tell, the answer to the question resides in noting that there is good activism and bad activism, and activism which is humble and activism which isn't, so that if and where Western academic activism is of the good-and-humble type, "what's not to like" in it?—as the title of the article suggests.

As I have mentioned earlier, the issue of epistemic injustice in poverty research is not one that is commonly encountered in poverty studies. To the extent that this is true, it is certainly welcome that the problem has been flagged in this volume. Having said that, I believe it is not just interesting but also relevant to ask to what extent the phenomenon has actually been addressed in this collection—not least since it occupies an important place in its stated concerns (it is part of the book's title, and an entire section has been devoted to it). A quick count in the "About the Contributors" section (xi–xvi) yields thirty contributors in all, of whom twenty-five seem to be operating from the UK, the USA, Canada, Germany, France, Belgium, and Austria, while only four are from the global South: South Africa, Chile, Cameroon, and Nigeria, and one is from Taiwan. It is not just a matter of counting nationalities in the list of contributors. Rowntree and Booth are mentioned often in this volume in connection with identifying the absolute poor: unless I have missed something, I haven't seen a reference to the great Dadabhai Naoroji's ([1901] 1969) work in what amounted to deriving an 'absolute poverty line' for colonial India. (I know: Dadabhai who?) And it is not just the failure of the North to keep track of relatively rarefied and culture-specific work that occurs in the South. It is also failure of citation of significant peer research, following, for all one knows, from failure to even take note of research on areas that fall squarely within the ambit of the North's own research concerns. For instance, how much citation of work by Indian scholars working from India is there on unidimensional (money-metric) poverty and multidimensional poverty in India, as carried, for instance, in one of India's foremost social-science journals, the *Economic and Political Weekly*?[18] It is good to be concerned with 'epistemic injustices' in poverty research. I state that without irony. I would merely add that it is also good to do something about it.

Part IV of the book on measuring multidimensional poverty has the largest number of contributions: seven. The first of these is by Sabina Alkire, another reminder "of the ongoing work on poverty research as it relates to multidimensional poverty measurement" (198). Sanjay Reddy, in his essay, offers an attractively brief critique of the protocols of

[18] Some of the earliest work on unidimensional poverty assessment by reference to a poverty line has been done in India. Prominent examples would include the poverty line advanced by the National Planning Committee under Jawaharlal Nehru in 1938; that advanced by the Indian Planning Commission (1962); a spate of papers—often in the form of debate—on the subject in the columns of the *Economic and Political Weekly (EPW)* in the early 1970s (special mention may be made of the work of Dandekar, Rath, Rudra, Minhas, and Bardhan); and several more perspectives on the issue, in the same journal, in the years to come (a small sample would include Sitaramam et al. 1996; Srinivasan 2007; Krishnaji 2012; and Subramanian 2014, 2015); and elsewhere (for example, Swaminathan 2010). The *EPW* has also been a platform for work on multidimensional poverty (as in Jayaraj and Subramanian 2009, and Sengupta 2016).

unidimensional (money-metric) measurement that preside over the approaches adopted by the World Bank and the governments of India and the USA. Much of this criticism is a continuation and summarisation of work he has himself done earlier, and it is a pity that his objections have still not been met in the vastly misleading work on money-metric poverty which continues to be done in the World Bank and official Indian and US traditions. (One is reminded here of the humourist Allan Laing's take on how Damon Runyon might have reacted to Henry James' prose: "[…] however long [he] snow[s], I am not getting [his] drift, so what is the use of going around with [him]?"[19]) Caroline Dotter and Stephan Klasen use Indian data to illustrate what happens when we change indicator thresholds for certain dimensions of multidimensional poverty to reflect the possibility that absolute requirements in the space of functionings may elicit varying (relative) requirements in the space of resources.

The next three essays in Part IV deal with how one may come up with a list of dimensions to be employed in multidimensional poverty measurement. Xavier Godinot and Robert Walker reflect on how such a list might emerge, not through the fiat of 'experts', but through a participative and collaborative procedure of consultation with those that actually experience poverty, by way of a strategy which they call the "Merging of Knowledge" (269). In a second contribution, by Francesco Burchi, Nicole Rippin, and Claudio Montenegro, dimensions are sought to be identified by attempting to trace an agreement on chosen dimensions by locating overlapping areas of congruence in countries' constitutional provisions, supplemented by what the authors refer to as "the public consensus approach" (286), participatory studies, and surveys. Nicolas Brando and Katarina Fragoso suggest that the dimensions of deprivation reckoned in extant measures are inadequate because of their overwhelming focus on material and biological deprivations, to the neglect of 'relational' deprivation, as manifested in the lack of control and autonomy needed to convert resources into functionings, even in the presence, formally, of access to these resources.

In a spirit akin to that in which I have made some observations earlier on the concern for epistemic injustice in poverty studies, I would like to draw attention to a refreshingly honest and straightforward comment made by Godinot and Walker on the quest for a generally acceptable list of dimensions for use in multidimensional poverty measurement:

[19] Laing (1951, 178).

> And, one might hope, if the policy community is open to novel dimensions suggested by the research, developing appropriate indicators will be undertaken in partnership with people having direct experience of poverty. The fact that it takes courage to write the preceding sentence which the reader might find ludicrously optimistic, underlines just how far the rhetoric of participation in international governance is distanced from practice. (271)

The above quote provokes speculation on yet another aspect of paternalism and the Hegemony of the Expert in poverty analysis. Multidimensional poverty measurement of a certain sort requires us to specify dimensions of deprivation, indicators within each dimension, and thresholds within each indicator—as well as various possible compromises between 'intersection' and 'union' approaches to the identification of the multidimensionally poor. Many of these choices have already been made in the publications of supra-national institutions. The requirement of 'comparability' across countries will surely gravitate in the direction of eliciting compliance from all countries on norms, procedures, and measures. This has already happened with unidimensional (money-metric) poverty, and it should be no surprise if a similar thing should also happen with multidimensional poverty (if it has not already happened). I speak of a situation in which 'If It Is Unidimensional Poverty, It Must Be The World Bank's Dollar-A-Day', and 'If It Is Multidimensional Poverty, It Must Be The UNDP's MPI', just as 'If It Is Tuesday, This Must Be Belgium'.

Returning to Part IV of the book, the seventh contribution deals with a particularly problematic dimension of deprivation, one—such as longevity or child mortality—which requires unavoidable engagement with the activity of valuing life. For those that value life infinitely, multidimensional poverty measures cannot entertain dimensions which entail a trade-off between the value of life and values in other dimensions. Is the project of multidimensional poverty measurement thereby unsalvageably jeopardised? Nicole Hassoun, Anders Herlitz, and Lucio Esposito address this interesting question, and suggest that multidimensional poverty measurement is still possible even when the comparability of the value of life with other values is denied, so long as incommensurability is combined with the notion that the value of life trumps all other values (though it may not be possible to say by 'how much').

The collection winds up, in Part V, with three useful country studies of multidimensional poverty in Cameroon, Germany, and Bangladesh.

By way of summing up, I would say that the volume under review reflects a certain drying-up of fresh research perspectives on poverty, as well as a certain tiredness with its subject of enquiry. This is manifested in various instances of convolution, repetition, triteness, and fine-tuning. The book does not want for effort or good intent, but in terms of outcome it is, in the end, disappointing—with few but honourable exceptions (in which I would specifically include the contributions of Reddy; Godinot and Walker; Burchi, Rippin, and Montenegro; Hassoun, Herlitz, and Esposito; and Hans Mpenya, Francis Baye, and Boniface Epo).

REFERENCES

Atkinson, Anthony B. 1987. "On the Measurement of Poverty." *Econometrica* 55 (4): 749–764.

Basu, Kaushik. 1985. "Poverty Measurement: A Decomposition of the Normalization Axiom." *Econometrica* 53 (6): 1439–1443.

Basu, Kaushik. 2001. "On the Goals of Development." In *Frontiers of Development Economics: The Future in Perspective,* edited by Gerald M. Meier and Joseph E. Stiglitz, 61–86. New York, NY: Oxford University Press.

Basu, Kaushik. 2006. "Globalization, Poverty, and Inequality: What is the Relationship? What Can Be Done?" *World Development* 34 (8): 1361–1373.

Basu, Kaushik. 2013. "Shared Prosperity and the Mitigation of Poverty: In Practice and in Precept." Policy Research Working Paper No. 6700. The World Bank, Washington, DC.

Chateauneuf, Alain, and Patrick Moyes. 2005. "Measuring Inequality Without the Pigou-Dalton Condition." Research Paper No. 2005/02. World Institute for Development Economics Research (UNU–WIDER), Helsinki.

Dasgupta, Indraneel, and Ravi Kanbur. 2005. "Community and Anti-Poverty Targeting." *The Journal of Economic Inequality* 3 (3): 281–302.

Deshpande, Ashwini. 2011. *The Grammar of Caste: Economic Discrimination in Contemporary India.* New Delhi: Oxford University Press.

Deshpande, Ashwini. 2013. *Affirmative Action in India.* New Delhi: Oxford University Press.

Deshpande, Satish. 2013. "Caste and Castelessness: Towards a Biography of the 'General Category'." *Economic and Political Weekly* 48 (15): 32–39.

Donaldson, David, and John A. Weymark. 1986. "Properties of Fixed-Population Poverty Indices." *International Economic Review* 27 (3): 667–688.

Dworkin, Ronald. 1977. *Taking Rights Seriously.* Cambridge, MA: Harvard University Press.

Galanter, Marc. 2002. "Righting Old Wrongs." In Martha Minow's *Breaking the Cycles of Hatred: Memory, Law, and Repair*, edited by Nancy L. Rosenblum, 107–131. Princeton, NJ: Princeton University Press.

Hassoun, Nicole. 2014. "An Aspect of Variable Population Poverty Comparisons: Does Adding a Rich Person to a Population Reduce Poverty." *Economics & Philosophy* 30 (2): 163–174.

Hassoun, Nicole, and S. Subramanian. 2012. "An Aspect of Variable Population Poverty Comparisons." *Journal of Development Economics* 98 (2): 238–241.

Jayaraj, D., and S. Subramanian. 2009. “A Chakravarty-D’Ambrosio View of Multidimensional Deprivation: Some Estimates for India.” *Economic and Political Weekly* 45 (6): 53–65.

Kanbur, Ravi. 2006. “The Policy Significance of Inequality Decompositions.” *The Journal of Economic Inequality* 4 (3): 367–374.

Karat, Brinda. 2012. “Of Mines, Minerals and Tribal Rights.” *The Hindu*. Accessed May 15, 2012. https://www.thehindu.com/opinion/lead/of-mines-minerals-and-tribal-rights/article3419034.ece.

Keen, Michael. 1992. “Needs and Targeting.” *The Economic Journal* 102 (410): 67–79.

Kelleher, J. Paul. 2015. “Capabilities versus Resources.” *Journal of Moral Philosophy* 12 (2): 151–171.

Kolm, Serge-Christophe. 1976a. “Unequal Inequalities. I.” *Journal of Economic Theory* 12 (3): 416–442.

Kolm, Serge-Christophe. 1976b. “Unequal Inequalities. II.” *Journal of Economic Theory* 13 (1): 82–111.

Krishnaji, Nidadavolu. 2012. “Abolish the Poverty Line.” *Economic and Political Weekly* 47 (15): 10–11.

Krtscha, Manfred. 1994. “A New Compromise Measure of Inequality.” In *Models and Measurement of Welfare and Inequality*, edited by Wolfgang Eichhorn, 111–119. Heidelberg: Springer-Verlag.

Laing, Allan M. 1951. “Damon Runyon on Henry James.” In *The Phoenix Book of Wit and Humour*, edited by Michael Barsley. London: Readers Union, Phoenix House.

Loury, Glenn C. 2000. “Racial Justice.” Lecture given during the W. E. B. Du Bois Lectures at Harvard University, Cambridge, MA, April 17, 2000. http://www.bu.edu/irsd/files/DuBois_3.pdf.

Mitra, Anirban, and Debraj Ray. 2014. “Implications of an Economic Theory of Conflict: Hindu-Muslim Violence in India.” *Journal of Political Economy* 122 (4): 719–765.

Naoroji, Dadabhai. (1901) 1969. *Poverty and Un-British Rule in India*. New Delhi: Publications Division, Ministry of Information and Broadcasting, Government of India.

Padel, Felix, and Samarendra Das. 2010. “Cultural Genocide and the Rhetoric of Sustainable Mining in East India.” *Contemporary South Asia* 18 (3): 333–341.

Planning Commission. 1962. “Perspectives of Development, 1961–1976: Implications of Planning for a Minimum Level of Living.” Reprinted in *Poverty and Income Distribution in India*, edited by Thirukodikaval N. Srinivasan, and Pranab K. Bardhan, 1974, 9–38. Calcutta: Statistical Publishing Society.

Pogge, Thomas. 2010a. “A Critique of the Capability Approach.” In *Measuring Justice: Primary Goods and Capabilities*, edited by Harry Brighouse, and Ingrid Robeyns, 17–60. Cambridge: Cambridge University Press.

Pogge, Thomas. 2010b. *Politics as Usual: What Lies Behind the Pro-Poor Rhetoric*. Cambridge, MA: Polity Press.

Sen, Amartya. 1979. “Issues in the Measurement of Poverty.” *The Scandinavian Journal of Economics* 81 (2): 285–307.

Sengupta, Anindita. 2016. “Gender Inequality in Well-being in India: Estimates from NFHS Household-level Data.” *Economic and Political Weekly* 51 (13): 43–50.

Motiram, Sripad, and Ashish Singh. 2012. “How Close Does the Apple Fall to the Tree? Some Evidence from India on Intergenerational Occupational Mobility.” *Economic and Political Weekly* 47 (40): 56–65.

Sitaramam, Vetury, Sharayu A. Paranjpe, T. Krishna Kumar, Anil P. Gore, and J. G. Sastry. 1996. "Minimum Needs of Poor and Priorities Attached to Them." *Economic and Political Weekly* 31 (35/36/37): 2499–2505.

Srinivasan, Thirukodikaval N. 2007. "Poverty Lines in India: Reflections after the Patna Conference." *Economic and Political Weekly* 42 (41): 4155–4165.

Stewart, Frances. 2005. "Horizontal Inequalities: A Neglected Dimension of Development." In UNI-WIDER's *Wider Perspectives on Global Development*, 101–135. New York, NY: Palgrave Macmillan.

Subramanian, S. 2009a. "Poverty Measurement in the Presence of a 'Group-Affiliation' Externality." *Journal of Human Development and Capabilities* 10 (1): 63–76.

Subramanian, S. 2009b. "Revisiting the Normalization Axiom in Poverty Measurement." *Finnish Economic Papers* 22 (2): 89–98.

Subramanian, S. 2014. "The Poverty Line: Getting It Wrong Again… and Again." *Economic and Political Weekly* 49 (47): 66–70.

Subramanian, S. 2015. "Once More Unto The Breach… The World Bank's Latest 'Assault' On Global Poverty." *Economic and Political Weekly* 50 (45): 35–40.

Sundaram, Krishnamurthy, and Suresh D. Tendulkar. 2003. "Poverty among Social and Economic Groups in India in 1990s." *Economic and Political Weekly* 38 (50): 5263-5276.

Swaminathan, Madhura. 2010. "The New Poverty Line: A Methodology Deeply Flawed." *Indian Journal of Human Development* 4 (1): 121–125.

Thorat, Sukhadeo, and Katherine S. Newman, eds. 2010. *Blocked by Caste: Economic Discrimination in Modern India*. New Delhi: Oxford University Press.

S. Subramanian (b. 1953) studied at Loyola College (University of Madras), the Indian Institute of Management (Ahmedabad), and the London School of Economics and Political Science. He is a former Indian Council of Social Science Research (ICSSR) National Fellow, and a former professor of the Madras Institute of Development Studies. Subramanian is an elected fellow of the Human Development and Capability Association (HDCA). He has worked extensively on measurement and other aspects of poverty, inequality, and demography, and on topics in collective choice theory, welfare economics, and development economics. His published work has appeared in journals such as the *Journal of Development Economics*, *The Journal of Development Studies*, *Social Choice and Welfare*, *Theory and Decision*, and *Mathematical Social Sciences*. Some of his books are *Rights, Deprivation, and Disparity: Essays in Concepts and Measurement* (New Delhi: Oxford University Press, 2006), *The Poverty Line* (New Delhi: Oxford University Press, 2012), and *Inequality and Poverty: A Short Critical Introduction* (Singapore: Springer, 2019). In 2015, he was appointed as a member of the Advisory Board of the World Bank's Commission on Global Poverty under the Chairmanship of Sir Anthony Atkinson. He lives and works in Chennai.
Contact e-mail: <ssubramanianecon@gmail.com>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 214–222.
https://doi.org/10.23941/ejpe.v13i2.535

Review of Thomas Piketty's *Capital and Ideology*. Translated by Arthur Goldhammer. Cambridge, MA: The Belknap Press of Harvard University Press, 2020, 1093 pp.

FRANCESCO GUALA
University of Milan

PIKETTY'S JUST SOCIETY

Six years ago, the translation of *Capital in the Twenty-First Century* (Piketty 2014) came out in bookstores all over the world. It would quickly become one of the publishing phenomena of the decade, selling millions of copies and receiving endorsements by celebrities, politicians, and Nobel laureates. In spite of its size (about a thousand pages), *Capital in the Twenty-First Century* was an interesting example of social science accessible to the average reader. Using dozens of graphs and very little theory, Piketty put forward and defended a simple thesis: after a period of decline in the middle of the twentieth century, inequality in the accumulation of capital has started to rise again, and will continue to do so in the future.

In the recessive climate that followed the collapse of financial markets, the success of *Capital in the Twenty-First Century* was not difficult to explain. The main message—inequality has increased a lot—was accompanied by another popular claim: inequality has increased *too much*, and we must do something about it. Being an essentially factual book, however, *Capital in the Twenty-First Century* did not argue these latter claims. And it did not indicate what sort of reforms could reduce inequality without affecting well-being or creating other injustices. The numerous articles written by Piketty in magazines and newspapers after the release of the book shed some light on his political orientation, but left several questions unanswered: Why do we want more equality? Which inequalities are unjust, and which ones are not? How much inequality are we willing to tolerate, and when does inequality become excessive? And, if inequality is excessive, what can we do about it?

Seven years later, Piketty has published another ambitious and demanding work. The title—*Capital and Ideology*—evokes Marx again. And the weight is again impressive: 1,100 pages, in the English edition pub-

EJPE'S NOTE: This review is a revised translation of a review by Francesco Guala (2018) published on the online Italian blog *Doppiozero* on February 22, 2018.

lished by Harvard University Press. *Capital and Ideology* departs even further from the canons of standard books of economics. It deals not only with economic history, but also with the history of ideas and of institutions and contains a political manifesto for a new 'socialism for the twenty-first century'. And yet again, despite its size and breadth, *Capital and Ideology* does not fully satisfy readers' curiosities, as I will explain shortly.

First, I will summarize the contents of the book, as far as it is possible. The seventeen chapters of *Capital and Ideology* are organized in four parts. The first three parts are essentially historical and provide an overview of 'inequality regimes' from the Middle Ages to the present day. By 'regime' Piketty means the set of institutions that determine the production and distribution of wealth in each society. Each regime is partly dependent on the scientific and technological knowledge of the time, without being entirely determined by it. And each regime produces its own 'ideology', that is, a set of beliefs, theories, arguments aimed at justifying the prevalent forms of inequality. Departing from classic Marxism, Piketty emphasizes the autonomy of ideologies from the forms of production. There is no determinism but a relationship of mutual interdependence.

The first three quarters of the book—about seven hundred pages—are only a starter for the main course served in the last part, "Rethinking the Dimensions of Political Conflict" (chapters 14–17). Here Piketty outlines his proposals, a set of reforms aimed at stopping the growth of inequality and creating the conditions for a truly just society. The pivotal mechanism is the wealth tax, a progressive tax on the savings of the richest citizens that would erode the accumulation of capital over the years. The proceeds from the wealth tax would be used to provide a basic starting endowment for young citizens upon reaching the age of maturity. But there is more than this. In the fourth part of the book, Piketty deals with broader issues such as the change in the social base of traditional parties, the emergence of 'nativist' populism, the free movement of individuals across borders, and the problem of global warming.

The main function of the early chapters is preparatory: they are meant to set the stage for Piketty's proposals. The message is simple and not very controversial: the current increase in inequality is not a natural phenomenon and is not inevitable. On the contrary, it is a contingent development that people accept both for lack of imagination (there does not seem to be an alternative) and because they have been persuaded by fallacious arguments. These arguments are precisely the ideology we must get rid of. Piketty repeatedly emphasizes that he intends to use the term 'ideology' in a neutral, non-derogatory sense: any regime has its own ideology. This historicist approach, however, cuts both ways, insofar as it suggests a relativist interpretation: If there is no difference in value between different ideologies, if Piketty is describing without passing any judgment, how are we supposed to take his proposal? According to his own approach, he is just producing another ideology. The fact that it is more egalitarian than others does not seem to be a good reason to accept it. Unless…

Unless equality and justice are the same for Piketty, and the goal of designing a more equal society is considered so obvious that it does not even deserve a discussion. This is the impression one gets from reading the first part of *Capital and Ideology*. 'More equal' and 'more just' are basically used as synonyms, and Piketty never cares to tell us what justice is or why equality is just. We gain some insights only after almost a thousand pages. At this point (968–969), Piketty finally explains that a just society is not characterized by absolute equality. It is rather a society in which inequalities are functional to improve the well-being of those who are worst off. A footnote refers to John Rawls, who articulated this idea in the most comprehensive and influential way half a century ago (Rawls 1971). And this is all the political philosophy that you will find in *Capital and Ideology*.

This is a serious fault. There is plenty of evidence (for example, Hochschild 1981; Kluegel and Smith 1986; Le Grand 1991; Miller 1992; Konow 2003) that the majority of people do not simply identify justice with equality. They rather think that a just society may well be unequal if the differences are justified. To convince the average citizen, therefore, one cannot appeal to justice in a generic way. A positive argument in favor of more extensive appropriation and redistribution of income by the state is required. The problem today is not mainly to convince voters that it is possible to reduce inequality, as Piketty obstinately argues for a thousand pages. You must convince them that it is right.

Common sense morality in the economic realm is based on two fundamental principles—the principle of equality and the principle of merit or productivity (for example, Mitchell and Tetlock 2009). The first one says that a commodity or service which has not been produced by anyone in particular ought to be distributed equally. The second one says that whoever has produced a good or service has a special entitlement, a right to use, or a priority of ownership over those who have done nothing. The problem is how to solve the numerous cases in which the two principles are in conflict—when wealth is produced through cooperation, for example.

Rawls prioritized the principle of equality and believed that the principle of productivity should be violated in order to achieve more equality. The only reason for respecting productivity is that it may work as an incentive for the most talented or industrious members of society. If he had to give up most of his earnings, for example, Lionel Messi might put less effort in training, and in the end might produce football of a lower quality. Similarly, a very talented surgeon with an ordinary salary might prefer to spend the weekends with his family, rather than in the operating room. One of the problems with this approach, noticed by Rawls' critics, is that it seems to assume that inequality is acceptable only as a necessary evil. But most people disagree: they think that it is right for a surgeon to earn more than his colleagues, if he is better at his job and works harder than they do.

One may try to circumvent the problem by linking economic justice directly with commitment, talent, or merit. Equality would then become a relative principle: it would not prescribe the same income for everyone, but the same income for those who have the same merit. Piketty, however, does not like this solution: he dismisses 'meritocratic hypocrisy' in a few paragraphs, as an ideology that has been created merely to justify the position of the winners. Questions of consistency aside (is 'ideology' derogatory now?), he is probably right not to follow this argumentative route. Three decades of philosophical debates have demonstrated that egalitarian theories based on merit—such as Arneson's (1989), Dworkin's (2002), or Roemer's (1998) 'luck egalitarianism', for example—suffer from numerous problems. On the one hand, desert is not aligned with the productivity principle, for success in market economies rewards the ability to satisfy others' preferences, which may depend (and often does depend) on factors that are unrelated to merit, such as luck. On the other hand, trying to compensate for the 'distortions' of luck would raise thorny issues. If he had been born with a single arm, the surgeon would not have been able to operate. Although he does not deserve to have both arms, is this original luck a good reason to take away a chunk of income at the end of the year? If the answer is positive, where are we going to stop? The surgeon was also lucky to have parents who encouraged him to study medicine. He was lucky to have a teacher who inspired him to specialize in surgery instead of geriatrics. He was lucky to be born in a country where medical studies are accessible to many. And so on and on, endlessly.

One may say that a precise quantification of merit is not essential. Once we recognize that our individual achievements are largely determined by factors that are beyond our control, we can agree that a significant egalitarian compensation is in order. How much redistribution would be exactly just is a minor detail. But I fear that this reply is wrong, again: it ignores the fact that a significant redistribution of income already takes place and is taken for granted by most citizens. In most European countries, from thirty-five to fifty percent of gross domestic product is taxed and redistributed through direct transfers or via the provision of goods and services (a little less in the United States). The controversial issue is not *that* we should redistribute, but *how much*. In all countries most taxation is born by the wealthiest sectors of the population (for example, Joumard, Pisu, and Bloch 2012). In Italy, forty percent of citizens pay almost all income tax, to give an idea, while more than half do not contribute anything (for example, Italian National Institute of Statistics 2017). So, the question is not whether the wealthiest citizens should pay more than others—they already do—but how much *more* they should pay, compared to the status quo.

This is the key question for Piketty, as indicated by numerous clues. Most notable is his obsession for the highest income bracket, the super-rich of the infamous 'one percent'. Statistically, the distribution of the population across income brackets has the shape of an elephant: it begins with a large hump on the left (low and medium-low incomes); it drops in the middle (medium-high incomes); and then rises steeply on the far right (the 'trunk' of the elephant). One of the most significant phenomena of the past thirty years has been the growth of the trunk, that is, the disproportionate increase in the number of families with incomes over two hundred thousand dollars. In relatively dynamic countries, such as the United States, the elephant's hump has also shifted to the right, which means that a significant number of families have made a transition from the middle or lower-middle class to the upper or upper-middle class. But, interestingly, Piketty does not care about the hump—he only looks at the trunk.

Why this obsession? Piketty presumably does not simply hate the rich and is not ideologically anti-capitalist. Although he sometimes presents his proposals as revolutionary, in reality his tax on capital would increase average taxation on wealth to five percent, from current rates of about two to three percent. His wealth tax would reach ten percent in the case of large estates, a policy that would progressively erode the accumulation of capital but would not 'transcend' capitalism as Piketty boasts (989).

In my view, his obsession is to be explained differently: Piketty probably senses that the real reason to worry about the super-rich is neither economic nor moral but has to do with the delicate social and political equilibria of democratic countries. Historically, democracy has emerged and has worked well in societies with a strong middle class. The shifts in income described above—the sinking of the elephant's 'hump', in particular—have reduced the political center of gravity upon which democratic countries are based. This is a historic change of great importance, with potentially devastating effects (for example, Fukuyama 2011).

The divergence of interests between classes that differ too much in terms of lifestyle, culture, and the ability to influence political decisions, tends to generate enormous tensions that are difficult to control. In many cases, it promotes social segregation, endemic violence (high crime, private protection agencies), and the emergence of 'strong men' who try to combine nationalistic populism with a defense of the economic interests of the oligarchy.

Those who care about democracy are entitled to worry. Perhaps Piketty is among them, even if he does not say it explicitly. It is a pity that he does not try to articulate his concern for the super-rich more clearly, not the least because some of the arguments are ready to use. 'Democratic egalitarianism' would offer solid arguments in support of his reforms. The main message of democratic egalitarians like Elizabeth Anderson (1999) and David Miller (1999) is that the Left should not pursue equality as an end in itself. The preservation of social cohesion within national communities, without which democracy cannot function, should be the main target of progressive parties. Such a goal does not require that we are able to measure desert, because it has little to do with it. Nor does it require that we identify justice with equality in the economic sense. Freedom and political justice, primarily, require that we impose a cap on excessive inequality.

I have dwelled on these issues because the main shortcoming of Piketty's book is the lack of an underlying theory. *Capital in the Twenty-First Century* was a factual book, and this theoretical deficit could be forgiven. But *Capital and Ideology* is explicitly a political text: those who expected a theoretical leap, unfortunately, will be disappointed.

We still have Piketty's recipes for reform, of course: from this point of view, what is the overall verdict? Piketty has the great skill of simplifying complex arguments using large numbers. The sections dedicated to the tax on accumulated wealth are convincing, in the sense that the reader is persuaded that in principle it can be done. For income tax, Piketty revives the top brackets of the Old Left. The proceeds would be used to finance a form of minimum income, as many representatives of the New Left advocate. But above all, they would promote inter-generational redistribution and partially neutralize hereditary privilege. These reforms would allow to bridge the gap, both in income and in opportunities, between younger and older people which in many countries continues to increase over time. Finally, a more robust participation of workers in the management of large firms—modelled on German and Swedish corporate law—would impose natural limitations on managers' salaries.

Piketty does not spend much time discussing the obstacles that such reforms would face. Aversion to taxation in many societies is correlated with lack of trust in state institutions. Unfortunately, Piketty always plays the role of advisor to an enlightened prince or citizenship, who once persuaded would have no problem applying the just reforms. In the ideal world of Piketty, the government is always benevolent. Citizens trust politicians and do not punish them when they raise taxation. Unions never defend unproductive rent and never lead companies to bankruptcy with the collusion of politicians. The rich are the source of all evil, and if we get rid of them, everything will be fine.

Things are a little more complicated, as politicians know well. Citizens do not trust governments blindly, even when the latter try to defend the rights of the poor. Citizens often do not trust each other and fear that transfers of resources will turn into rents. Such hurdles would be even more impervious if the redistribution of resources took place on a global scale. Piketty is in favor of transgressing national borders, both in the movement of people and in fiscal solidarity. But in his idealized world, citizens have no qualms about giving up part of their sovereignty. Hungarian or Spanish voters are willing to be governed by a Dutch or Finnish prime minister, and when the latter imposes sacrifices or spending cuts, they accept them cheerfully. European elections are held in this fictional world even if candidates are unable to address their voters in an understandable language. Finally, in the world of Piketty, populists cannot exploit people's resentment against foreigners to gain seats in parliament.

I suspect that political scientists will find the chapters devoted to trans-national justice, populism, and global warming rather naive. It is a pity because *Capital and Ideology* is full of interesting ideas. The problem is that the analysis of economic reforms—in particular, of the tax policies—does not justify a volume of over a thousand pages. Piketty should have written a more compact book centered on the last chapter, perhaps introduced by a short summary of the previous ones. The great historical fresco of the first seven hundred pages deserves a separate outlet and does not add much to Piketty's theses. An explicit defense of the idea of justice underlying the reforms, on the other hand, would have greatly strengthened his political proposal. But for this we will have to wait until the next book, which the prolific Piketty will undoubtedly write soon.

REFERENCES

Anderson, Elizabeth S. 1999. "What Is the Point of Equality?" *Ethics* 109 (2): 287–337.

Arneson, Richard J. 1989. "Equality and Equal Opportunity for Welfare." *Philosophical Studies* 56 (1): 77–93.

Dworkin, Ronald. 2002. *Sovereign Virtue: The Theory and Practice of Equality*. Cambridge, MA: Harvard University Press.

Fukuyama, Francis. 2011. "Poverty, Inequality, and Democracy: Dealing with Inequality." *Journal of Democracy* 22 (3): 79–89.

Guala, Francesco. 2018. "Piketty e L'uguaglianza." *Doppiozero*. Published on February 22, 2018. https://www.doppiozero.com/materiali/piketty-e-luguaglianza.

Hochschild, Jennifer L. 1981. *What's Fair? American Beliefs about Distributive Justice*. Cambridge, MA: Harvard University Press.

Italian National Institute of Statistics. 2017. "La Redistribuzione del Reddito in Italia." Press release No. 201597, June 21, 2017. https://www.istat.it/it/files/2017/06/CS_-Redistribuzione-reddito-in-Italia_2016.pdf.

Joumard, Isabelle, Mauro Pisu, and Debbie Bloch. 2012. "Tackling Income Inequality: The Role of Taxes and Transfers". *OECD Journal: Economic Studies* 2012 (1): 37–70.

Kluegel, James R., and Eliot R. Smith. 1986. *Beliefs about Inequality: Americans' Views of What Is and What Ought to Be.* New York, NY: Aldine De Gruyter.

Konow, James. 2003. "Which Is the Fairest One of All? A Positive Analysis of Justice Theories." *Journal of Economic Literature* 41 (4): 1188–1239.

Le Grand, Julian. 1991. *Equity and Choice: An Essay in Economics and Applied Philosophy*. London: Routledge.

Miller, David. 1992. "Distributive Justice: What the People Think." *Ethics* 102 (3): 555–593.

Miller, David. 1999. *Principles of Social Justice*. Cambridge, MA: Harvard University Press.

Mitchell, Gregory, and Philip E. Tetlock. 2009. "Disentangling Reasons and Rationalizations: Exploring Perceived Fairness in Hypothetical Societies." In *Social and Psychological Bases of Ideology and System Justification*, edited by John T. Jost, Aaron C. Kay, and Hulda Thorisdottir, 126–157. New York, NY: Oxford University Press.

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Translated by Arthur Goldhammer. Cambridge, MA: The Belknap Press of Harvard University Press.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: The Belknap Press of Harvard University Press.

Roemer, John E. 1998. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Francesco Guala is a philosopher and experimental economist interested primarily in the foundations and the methodology of social science. He teaches in the Department of Philosophy at the University of Milan (Italy). He is the author of many articles in scientific and philosophical journals, and of two monographs, *The Methodology of Experimental Economics* (Cambridge, 2005) and *Understanding Institutions* (Princeton, 2016). In 2011 he co-edited with Daniel Steel *The Philosophy of Social Science Reader* (Routledge).

Contact e-mail: <francesco.guala@unimi.it>
Website: <users.unimi.it/guala>

Erasmus Journal for Philosophy and Economics,
Volume 13, Issue 2,
Winter 2020, pp. 223–228.
https://doi.org/10.23941/ejpe.v13i2.536

Review of Robert B. Talisse's *Overdoing Democracy: Why We Must Put Politics in its Place*. New York, NY: Oxford University Press, 2019, ix + 198 pp.

ELIAS ANTTILA
Vrije Universiteit Amsterdam

The point of departure of Robert B. Talisse's *Overdoing Democracy* will feel familiar to even casual followers of American political news outlets or political online culture: since Donald Trump's election to the US presidency in 2016—and as evidenced by it—the American political spectrum has expanded further in both directions, which has resulted in increased animosity between political partisans. "Democratic politics", it is said, "is tearing us apart" (3). Talisse's book makes essentially two claims. First, it argues that, tragically, this is in fact the case: to their detriment, various kinds of political polarization have become a trend in Western democracies. Second, it provides a practical solution that involves forgoing political conversations and finding non-political activities to do together to heal the political divide. The book has three parts. The two parts of the book that each develop one of these arguments are prefaced with a separate background section, which provides new readers with some very basic democratic theory, making *Overdoing Democracy* accessible to a wider audience while still providing new academic insights.

The background section of the book (Part I: Framing the Thesis), which provides appropriate groundwork in democratic theory for the rest of the book, emphasizes accessibility. Readers completely new to democratic theory or political philosophy will find concrete, relatable examples in Talisse's discussion of the nature of democracy. In these examples and in their surrounding discussion, Talisse makes the conscious assumption that democracy is desirable and valuable on the whole, both as a form of government and as a principle of social life. *Overdoing Democracy* does not seek to persuade those inclined to distrust democracy on account of recent world politics or explain *why* democracy is a "capital social good" (12). Talisse briefly contrasts his support of democracy to anarchism, although in its brevity this contradistinction is misleading: in one sense, although anarchists all share an opposition to parliamentary politics on the grounds that they create hierarchy, viewing democracy as a principle of social life, as Talisse suggests, seems to match with contemporary left-anarchism very well.[1]

Anarchists aside, Talisse continues to frame his thesis with a discussion of the democratic views of Jane Addams and John Dewey. For Talisse, Addams and Dewey represent an overly trusting view of the power of democracy, which seeks to solve the social and political problems that arise in democracies by prescribing *more* democracy in public life. (At this point, it is not clear what practical interventions might follow from this prescription.) For Talisse, the main challenger to Addams' and Dewey's democratic philosophy is (not anarchists but) elitist-minded anti-democrats, who bracket the public as gullible masses to be managed, not empowered. These anti-democrats do not go by name in Talisse's text, but rather seem to be adherents to a kind of folk view. The conflict between Addamsian and Deweyan democrats and the elitist anti-democrats that Talisse puts forth could be understood as a version of the conflict between democrats and technocrats. Regardless, Talisse intends his own position to lie in-between these two views. Like the democrats, he places great value in democracy, and believes in both the intrinsic and instrumental value of democracy. However, unlike the democrats, he thinks that in the face of democratic hiccups such as polarization, in prescribing 'more democracy', democracy 'overdoes' itself, and consequently further polarizes the public to the detriment of good governance. Therefore, Talisse adds to his slogan 'democratic politics is tearing us apart' that it "must be put in its place" (11).

The rest of the book is devoted to arguing that democratic politics is in fact polarizing the electorate and that this is undesirable, and to explaining what can be done about it. In what follows, I recount the argument before raising some concerns. Part II of the book (Diagnosis) explicates and argues for this slogan in two parts. In saying that democratic politics is in fact 'tearing us apart', Talisse supports the claim that the political spectrum has become further polarized, and describes what that polarization is like. This argument is framed by introducing the concepts of political saturation and political reach. Political saturation (of social life) refers to the phenomenon where political projects come to dominate all or many aspects of social life. Political reach refers to the physical and social locations where the duties, obligations, and responsibilities of citizens are exercised. Using UK and US data, Talisse argues that political

[1] See, for example, Graeber (2010) for an accessible description.

projects have come to overshadow all social encounters, which are becoming increasingly guided by the need to stay true to political 'allegiances'. For example, consumer-brand allegiances may also double as political allegiances, because consumers now view choosing and committing to brands to be politically charged. Talisse invites us to think about this kind of homogenization as a concentration of political allegiance within another allegiance (namely, consumer-brand allegiance). The same homogenization is also seen in physical spaces (a postal code can accurately predict the political beliefs of its inhabitants) and virtual spaces (social media and its users divide users into politically homogeneous social groups). Talisse ends this section by extending his argument to claim that the combination of this homogenization of spaces and the creation of relevance for political allegiances at every turn is, in fact, what drives the overextension of politics and belief polarization.

In order for polarization and the overextension of politics to be relevant, next, Talisse needs to argue that, all things considered, these phenomena are undesirable. To do this, Talisse first invokes a difference between political polarization and belief polarization. Political polarization indicates how divided political allegiances are, including the distance of party platforms (platform polarization); how absolute political allegiances are about their platform (partisan polarization); and how distrusting political opposites are of each other (affective polarization). Belief polarization, on the other hand, indicates the process by which people come to hold stronger versions of their views after discussing them with like-minded people. These two categories also seem to happily differentiate between polarization in cases where agents adopt more of a particular set of beliefs (political polarization) and cases of single-issue polarization. In discussing belief polarization, Talisse invites us to consider beliefs as affects that intensify as interlocutors become increasingly polarized. These political affects work primarily to affirm a group identity, which can then distort the views of those who do not belong in the group. Eventually, this causes members of the public to lose the ability to engage in rational discussion with others. Talisse speculates that the current political landscape even provides the right conditions for a civil war. This speculation aside, I note that Talisse's view of the dynamics of belief polarization is similar to that of Dan Kahan and his work on 'cultural cognition'.[2]

[2] See, for example, Kahan et al. (2012).

Moving to the final part of the book, Talisse's last argument is prescriptive. Because Talisse observes that belief polarization is often conditional on inhabiting an environment that corroborates existing views, to put politics in its place, he suggests that heterogenous members of the public must find mutually engaging non-political activities to do together. Mutual non-political activities serve to both prevent polarization and depolarize those who already have become polarized, because they cultivate an attitude of 'civic friendship'. Civic friends respect each other as people who have equal say in shaping communities and society even if they may not personally like each other. This practice and the habit of civic friendship is intended to reel in the overextension of the saturation and reach of democratic politics, which is at the centre of political breakdown. And this, in turn, should fix the overreach of democracy. "Putting politics in its place" (31) means, unlike Addams' and Dewey's suggestion, prescribing *less* democracy, not more.

While Talisse's diagnosis is coherently argued, a few concerns shadow each of his arguments. First, while Talisse's claim that talking politics is rarely done in mixed company seems plausible, the empirical evidence provided with these three examples is somewhat inconclusive and incomplete. For example, the data Talisse provides does not accurately track the changes to how, where, and with whom political discussions occur over time, and just how much political brand allegiance is happening (and where and by whom). In the United States, for example, it is worth keeping in mind that most Americans who experience suffrage are still politically rather disengaged (when measured, for example, according to voter turnout, see DeSilver 2017) and do not like to talk about politics (Pew Research Centre 2019). It is true that polarization is taken to be an established reality in UK and US political settings, but a watertight conjunction or separation of conventional wisdom and the studied reality would require further evidence.

However, even granting the empirical premise of political polarization, Talisse's second argument also raises concerns: while it seems intuitive and plausible that polarization leads to a loss of abilities to engage with political opposites and that polarization is ultimately undesirable, these theses are ultimately a speculation. A countering speculation might claim, for example, that as political saturation and reach extend and as belief polarization occurs, the electorate discusses politics more and becomes more adept at reasoning against or together with our opponents; the more we argue, the better we get. In following the common narrative that polarization generally harms democracy, Talisse also does not discuss potential benefits that might arise from political conflicts, whose necessity and value are theorised and stressed, for example, by agonistic pluralists (see, for example, Wenman 2013). It is unclear why it is necessarily undesirable to form or hold extreme beliefs: radical ideals cannot be ruled out a priori as undesirable. In favouring agreement over disagreement and non-extreme beliefs over extreme beliefs, Talisse seems to unjustifiably favour centrism or the maintenance of a political status quo. (He strongly denies that his argument is centrist or conservative, but this insistence is not further justified.) Thus, it is difficult to accept uncritically the thesis that polarization is intrinsically undesirable.

However, even granting that political polarization is wholly undesirable, Talisse's prescriptive argument also faces difficult challenges. The prescription to find mutual non-political activities does not capture the imbalance of the stakes for each party in coming together. In many encounters, if not all, requesting neutrality is a non-neutral request. According to Talisse, some views are not to be tolerated, but it is not clear where this line is to be drawn, who gets to draw it, and why.

A second concern for his prescriptive argument, as Talisse himself points out, is that it is difficult to imagine what the apolitical activities could be. Voluntary work, games, entertainment, food, recreational activities, and talking about these seem to all be politically ripe, as Talisse notes. This concern is briefly taken up, explicating that it is hard work, but this note does not address the issue adequately. Talisse's argument here could be helped by considering *how* political some activities are, and especially in which ways they are political. The presumption here is that because an activity is political, it will polarise further or have disastrous consequences. This may often be the case; but in the cases that it is not, how did that happen?

A third challenge to the prescriptive argument could be made in terms of feasibility: how could Talisse's prescription be practically implemented? As the argument stands, Talisse seems to leave it up to the public to adopt his prescription, but why or how they could do that isn't clear. Along the lines of the second counter-argument, just how will people come together if they are already so polarized? More concretely, what might a policy intervention that takes on board Talisse's prescription look like?

To conclude, Talisse's book provides readers new to democratic theory, and political philosophy in general, an accessible entry point, one that is especially topical in the United States and some other Western democracies. Talisse's philosophical slogans are catchy, and they are sufficiently explained. This introduction chooses the familiar idea that political differences and differences in what we believe are becoming even farther apart from each other, and that this is to the detriment of a functioning democracy. This argument has some empirical merit, but its normative claim, which is universally against polarization, is not sufficiently justified. A concern about the argument's prescriptive component lies in its positing an (arguably false) equality between the costs of engaging in non-political activity. The theory also faces a cyclical difficulty: if political saturation and reach loom larger than ever, coming together over non-political things seems improbable. Throughout his book, Talisse does insist that the task of overcoming polarization will not be easy, as evidenced by the problems raised here and in his chapters. However, while solutions to polarization may be difficult to uncover, so are its problematizations.

REFERENCES

DeSilver, Drew. 2017. "In Past Elections, U.S. Trailed Most Developed Countries in Voter Turnout." *Pew Research Center*, May 15, 2017. https://www.pewresearch.org/fact-tank/2020/11/03/in-past-elections-u-s-trailed-most-developed-countries-in-voter-turnout/.

Graeber, David. 2010. "Are You An Anarchist? The Answer May Surprise You!" *The Anarchist Library*, November 9, 2009. https://theanarchistlibrary.org/library/david-graeber-are-you-an-anarchist-the-answer-may-surprise-you.

Kahan, Dan M., Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa L. Ouellette, Donald Braman, and Gregory Mandel. 2012. "The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks." *Nature Climate Change* 2 (10): 732–735.

Pew Research Center. 2019. "Public Highly Critical of State of Political Discourse in the U.S." *Pew Research Center*, June 19, 2019. https://www.pewresearch.org/politics/2019/06/19/public-highly-critical-of-state-of-political-discourse-in-the-u-s/.

Wenman, Mark. 2013. *Agonistic Democracy: Constituent Power in the Era of Globalization.* Cambridge: Cambridge University Press.

Elias Anttila is a PhD candidate in philosophy at the Vrije Universiteit Amsterdam. They have interests in democratic theory and epistemic injustice.
Contact e-mail: <o.e.anttila@vu.nl>