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:*

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

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 **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 **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, $\mathbf{s}^* = (s_1^*, s_2^*, \ldots, s_n^*)$ 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* $\left([2], S, \left(V^1, V^2\right)\right)$ *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* $\mathbf{s}^* = (s^*, \ldots, s^*)$ *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 $r > 0$, $f^i(rs_1, \ldots, rs_n) = r^k f^i(s_1, \ldots, s_n)$, 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, $\mathbf{x} = (x^1, \ldots, x^n)$ 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 $r \geq 0$:

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

Where the inequality follows from the fact that $\mathbf{s}^*$ is a multiplicative Kantian equilibrium of G. It follows that $\mathbf{f}\left(\mathbf{s}^*\right)$ 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 $\mathbf{f} = \left(f^1, f^2\right)$ 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:

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

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

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

Where the last non-equality follows from (17). It follows that $\left(\tilde f\,(2/3)\,, \tilde f\,(4/3)\right)$ 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 $\left(s_1^*, s_2^*\right) = \left(1/\sqrt 2, 1 + 1/\sqrt 2\right)$ 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:

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

Where the last non-equality follows from (18). So $\left(\tilde f\,(1/\sqrt 2)\,, \tilde f\,(1 + 1/\sqrt 2)\right)$ 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>