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

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

JOHN E. ROEMER
Yale University

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

COMMENT ON MATTHEW BRAHAM AND MARTIN VAN HEES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 6: The equilibrium outcomes.

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

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

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

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

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

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

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

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

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

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

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

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

COMMENT ON ITAI SHER

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

Sher writes:

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

My responses:

1. Existence

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

it implements most literally Kant's instruction that 'one should take that action that one would will be universalized'.

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

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

2. Efficiency

Sher's claim that it is easy to construct examples of inefficient Kantian equilibrium is bizarre. What I prove is that if the game is strictly

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

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

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

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

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

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

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

3. Strategic Non-Equivalence

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

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

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

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

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

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

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

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

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

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

for example, the work of the evolutionary psychologist Michael Tomasello, whom I discuss extensively in *HwC*).

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

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

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

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

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

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

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

Near the end of his paper, Sher writes:

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

Indeed!

COMMENT ON PETER VALLENTYNE

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Indeed, we can 'implement' *any social welfare function* with this procedure. Simply endow each player with the social preference order and have him play his strategy to achieve the highest ranked strategy profile, according to this order, that he can induce by his behavior alone. Trivially, any strategy profile that maximizes the social ordering is a Nash equilibrium of this game! The converse, however, is false. In particular, (*Confess*, *Confess*) is also a Nash equilibrium of this game, because if either player changes his strategy to *Silence*, the new profile (*Silence*, *Confess*) is

dominated by (*Confess*, *Confess*) according to $\succsim_{lex}$. So, there is a *bad* Nash equilibrium as well as a *good* one of the altered game. If we pursued this discussion further, and tried to construct a game that would implement $\succsim_{lex}$ in the more demanding sense that *every* Nash equilibrium of the game is a maximal element of $\succsim_{lex}$, we would be led down the path to Maskin-type implementation theory.

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

COMMENT ON JEAN-FRANÇOIS LASLIER

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

I will, however, comment more generally on the evolutionary approach to cooperation. There is no doubt that over time, our species has increased its degree of cooperation immensely. In pre-historic times, the extent of cooperation was limited to one's small band of at most several hundred souls. Today, we have cooperation within nations with over a billion souls. A most significant form of that cooperation is taxation, which now, in the most advanced countries, collects approximately one-half of the national product, and re-allocates it for the public good. This is a twentieth-century innovation. Karl Marx's theory was that cooperation increases as history progresses because of technological development: economic structures develop only so long as they encourage the further development of the productive forces, and are then 'burst asunder' when they become fetters on that development. To link this to the evolution of cooperation one would have to theorize why more advanced productive forces necessarily require more cooperation to operate. Marx gave some examples (feudalism gives way to capitalism which gives way to socialism), but the necessary link to cooperation is, I think, not satisfactorily explained.

Although I agree with Tomasello and others that our species evolved, through selective adaptation, as a cooperative species—in contrast even to other great apes—that cannot explain the rather short time span (say,

10,000 years) in which the extent of human cooperation has increased so dramatically. Something like the kind of mechanism that Marx offers is necessary to explain such rapid social evolution.

REFERENCES

Ahlquist, John S., and Margaret Levi. 2013. *In the Interest of Others: Organizations and Social Activism*. Princeton, NJ: Princeton University Press.

Cohen, Gerald A. 2009. *Why Not Socialism?* Princeton, NJ: Princeton University Press.

Roemer, John E. 1994. *A Future for Socialism*. London: Verso.

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

Roemer, John E. 2020. "What is Socialism Today? Conceptions of a Cooperative Economy." Cowles Foundation Discussion Paper No. 2220. Yale University, New Haven, CT.

John E. Roemer is the Elizabeth S. and A. Varick Stout Professor of Political Science and Economics at Yale University. He is a Fellow of the Econometric Society, and has been a Fellow of the Guggenheim Foundation, and the Russell Sage Foundation. His research concerns political economy, and distributive justice. He is the author of numerous books, including, among others, *How We Cooperate* (2019), *Sustainability for a Warming Planet* (2015), *Democracy, Education, and Equality* (2006), *Political Competition: Theory and Applications* (2006), and *Equality of Opportunity* (1998).

Contact e-mail: <john.roemer@yale.edu>