OPEN ACCESS

KING'S College LONDON

King's Research Portal

Document Version
Other version

Link to publication record in King's Research Portal

Citation for published version (APA):
Hargreaves Heap, S., & Ismail, M. (in press). No-harm principle, rationality, and Pareto optimality in games. *SYNTHESE*.

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 15. Jan. 2025

Liberalism, rationality, and Pareto optimality

Shaun Hargreaves Heap[1] and Mehmet S. Ismail[2]

Department of Political Economy, King's College London, London, UK

23 January 2021

Abstract

Rational players in game theory are neoliberal in the sense that they can choose any available action so as to maximize their payoffs. It is well known that this can result in Pareto inferior outcomes (e.g. the Prisoner's Dilemma). Classical liberalism, in contrast, argues that people should be constrained by a no-harm principle (NHP) when they act. We show, for the first time to the best of our knowledge, that rational players constrained by the NHP will produce Pareto efficient outcomes in *n*-person non-cooperative games. We also show that both rationality and the NHP are required for this result.

Keywords: Classical liberalism, no-harm principle, Pareto optimality, rationality, non-cooperative games

[1] s.hargreavesheap@kcl.ac.uk
[2] mehmet.s.ismail@gmail.com

1. Introduction

It is well known from game theory that there is no necessary connection between the Nash equilibrium of a non-cooperative game and Pareto optimality. The prisoners' dilemma illustrates this point well. It is also an important insight. Since Pareto optimality or efficiency is an appealing normative standard for judging outcomes, the insight sets an agenda for possible policy interventions to secure Pareto improvements in settings like the prisoners' dilemma. In this paper, we show a contrary result: when individual rational action in games is additionally constrained by the no-harm principle of classical liberalism, the associated no-harm equilibria are Pareto optimal. Or to put this slightly differently, we provide a non-cooperative foundation for Pareto optimality in games through the addition of the 'no-harm principle' to the standard assumption of individual rationality. We also show that it is the combination of individual rationality and the no-harm principle that secures this result. The no-harm principle by itself does not.

The importance of this new result depends, of course, on the appeal of the 'no harm principle'. The principle was most clearly stated by J. S. Mill in *On Liberty*, his classic defence of a liberal society founded on individual freedom:

> The sole end for which mankind are warranted, individually or collectively, in interfering with the liberty of action of any of their number, is self-protection. That the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others. His own good, either physical or moral, is not sufficient warrant (Mill, 1859).

That a person should be free to do whatever they please so long as it does not harm others is constitutive of all modern liberal societies. Indeed, much of the business of the judiciary in liberal societies involves deciding when one person's exercise of liberty may or may not be reasonably said to have conferred a 'harm' upon another. What counts as a 'harm' is, of course, naturally controversial and this is why the courts get involved. The point to notice, however, about the courts' involvement in such matters is that it arises because they are upholding the 'no harm principle'. It is a constitutive principle and so, as such, it ought to appeal to people who are voluntarily members of liberal societies. (This, though, is not to say it always does!)

Unlike the problems the courts face when deciding on a 'harm', we argue that, in the context of game theory where outcomes identify pay-offs for all players, a 'harm' to someone through another's action can be defined uncontentiously. A harm occurs to a person when their pay-offs are diminished by the actions of another. While this is an advantage of a game theoretic representation of social interactions, it is nevertheless not immediately obvious how to formalize the no-harm principle within standard game theory. The difficulty is that the principle requires each person to consider how an action might or might not harm another before deciding whether to take it; and, if harm is identified through a negative change in pay-offs for the other person, then this requires that there be a reference point pay-off for the other person. But what reference point should be used when deciding whether an action produces a negative change in the other's pay-off? Conventionally, when players decide what to do in a game, there is no reference point of this kind. All actions are potentially feasible by definition and

have equal footing in this sense. None is obviously a reference point or a status quo in this sense, unless arbitrarily so defined/labelled, from outside the game.

Our approach to this problem is to allow that any set of players' actions within the game could identify a status quo or reference outcome for such calculation. We now follow Brams (1994) in constructing an extensive form game on the basis of sequential deviations by the players from any such status quo. Unlike Brams (1994) who calls the outcome of this procedure non-myopic equilibrium, we introduce the no-harm principle as a constraint on play in this extensive form game (see also Brams and Wittman, 1981). The equilibrium of this game is what we call a no-harm equilibrium. Of course, there may be no deviation that satisfies individual rationality and the no-harm principle, in which case the status quo is the equilibrium of this game and it is a no-harm equilibrium. But in some cases the equilibrium of this game is another set of actions. In which case, this new set is the no-harm equilibrium associated with that status quo; and that status quo is not a no-harm equilibrium.

By considering every possible status quo, this procedure allows us to identify the no-harm equilibria of the game. Of course, which will obtain will depend on the actual status quo. We have nothing to say about this. What we show is that each such no-harm equilibrium is Pareto optimal. In short, in so far as people are rational and constrained by the no-harm principle in this way, then the no-harm equilibria that arise in the game are Pareto optimal. So, whatever is the actual status quo, the equilibrium outcome in the game when individual rationality is constrained by the no-harm principle will be Pareto optimal.

2. The no-harm principle in non-cooperative games

The tools of game theory did not exist when the no-harm principle was first introduced. And, it is not immediately obvious how one might define this key principle of classical liberalism in an *n*-person and / or dynamic setting where players' actions affect one another and players can condition their actions on the actions of others. To the best of our knowledge, we introduce the first definition of the NHP in such a strategic environment.

***Preliminaries**:*

Let $G = (A, u_i)_{i \in N}$ denote a normal form game in which $N = \{1, 2, \dots, n\}$ denotes a society whose members are called players, $u_i : A \to R$ player i's Bernoulli utility function representing a strict ranking over the consequences, A_i finite pure action set of player i, and $A = \times_{i \in N} A_i$ set of pure action profiles. Let $a \in A$ denote a pure action profile in a normal form game G.[3] As is standard in game theory, every action profile is associated with an outcome, which is a pay-off profile.

A profile a Pareto dominates a' if for all i $u_i(a) \geq u_i(a')$ with at least one strict inequality. A profile is called Pareto optimal or efficient if there is no other profile that

[3] Our definitions can be extended to the games with cardinal von Neumann-Morgenstern utilities in a straight-forward way. We keep the current framework for its simplicity.

Pareto dominates it. A profile is called weakly Pareto optimal if there is no other profile in which everyone is strictly better off.

***Construction of the extensive form alternating-offers game**:*

For a given status quo profile a in G, we define an extensive form game denoted by $\Gamma(a)$, where players take turns in a given order to decide whether either to change their action and so 'move' to another available profile in G or 'stay'. If they 'move', we say they propose a different outcome and when a player decides to 'stay', we say they agree to that outcome and it becomes the updated status quo. If the subsequent players in turn decide to 'stay' at the (possibly updated) status quo, then this outcome is implemented, and the players collect the pay-offs in that agreed outcome.

As an example, consider the prisoners' dilemma (PD), where we assume for the purposes of the illustration (C,C) is the status quo.

	C	D
C	3,3	1,4
D	4,1	2,2

In the extensive form game that begins with the status quo (C,C), Row player might unilaterally switch their action to D, hence 'moving' to (D,C). Column player might then unilaterally move to (D,D), where Row could choose to 'stay'. As a result of Row's decision to 'stay', (D,D) becomes the updated status quo. Column then decides whether to 'move' from or 'stay' at (D,D); when Column 'stays', (D,D) is the agreed outcome for both players and they receive payoffs of (2, 2).

***No-cyclicity condition**:*

We assume that the game tree constructed in this manner from any status quo, a, in G does not include "cycles". This is to ensure that it is a standard extensive form game. If a game tree allows the same move by the same player to be repeated, then we call this tree *cyclic*. We assume that our construction of the game tree is *acyclic*—i.e., a valid move that turns an action profile a' into another action profile a cannot be repeated by the same player in the extensive form game.

The no-cyclicity condition, for every status quo profile a in game G, allows the construction of the *extensive form alternating-offers game*, $\Gamma(a)$. Let $\sigma_i(a)$ denote a pure strategy of player i and $\sigma(a)$ denote a pure strategy profile in $\Gamma(a)$.[4]

To give an example from the PD above, a cycle can be created by the following sequence from the status quo (C,C): (D,C), (D,D), (C,D), and back to (C,C), *ad infinitum*. No-cyclicity condition says that the game tree does not allow Row player to move from (C,C) to (D,C) for the second time in order to prevent indefinite cycles.

[4] For a definition of extensive form games, see, e.g., Osborne and Rubinstein (1994).

***No-Harm Principle**:*

Our specification of the no-harm principle applies to a player's 'stay' decision in this extensive form alternating offer game. We make this assumption for two reasons. First, in a dynamic strategic setting, the classical liberal has no reason to be concerned with the properties of any transitional (i.e., non-status quo) states in this extensive form game, particularly if they are purely mental constructs. It is the eventual outcome that matters. Second, an individual can only contribute through their own decision to making an outcome the final outcome by deciding to 'stay'; it is his or her 'stay' decision where the classical liberal should apply the no-harm principle.

Definition 1. Consider $\Gamma(a)$ where $a \in A$. Let b and b' be two action profiles in G; and it is player j's turn in $\Gamma(a)$ to decide between moving or staying having inherited b' from earlier player decisions in $\Gamma(a)$. Action profile b is the current status quo. If j chooses to stay at b' then j's decision is said to satisfy the *no-harm principle* with respect to b if $u_i(b') \geq u_i(b)$ for every player i *not* equal to j. That is, when j agrees to the proposed outcome b' by deciding to stay this does not harm other players with respect to the current status quo b. We say that $\Gamma(a)$ satisfies the *no-harm principle* (NHP) if every stay decision satisfies the NHP with respect to the current status quo.

Of note, the no-harm principle implies neither Pareto optimality nor even Pareto improvement from a status quo. In section 4, we illustrate that assuming the no-harm principle in an extensive form game $\Gamma(a)$ may lead a society to a Pareto inferior outcome compared to the initial status quo a.

Solution concept:

We assume that players in games are individually rational and farsighted in the usual sense of subgame perfection and are additionally constrained by the no-harm principle (NHP) in their action choices in the extensive form alternating-offer game based on the current status quo, a. A strategy profile $\sigma(a)$ is called subgame perfect equilibrium (SPE) if it is a subgame perfect equilibrium in $\Gamma(a)$, which is a strategy profile that forms a Nash equilibrium in every subgame (Nash, 1953; Selten, 1965).

Definition 2. Let $a \in A$ be a status quo. A strategy profile $\sigma(a)$ is called *no-harm equilibrium* (NHE) if:

(1) $\Gamma(a)$ satisfies the no-harm principle, and
(2) players are farsightedly individually rational.

The NHE depends, of course, on the status quo, a. But for now it is important to note that the no-harm principle per se does *not* require Pareto optimality of the accepted offer—players simply act independently and maximize their individual utility, so they can stay wherever they want as long as the outcome does not harm others with respect to the status quo: i.e. there could always be other outcomes that are as good for the individual who decides to 'stay' and which would be better for the other players. Nor does the NHP by itself imply a Pareto improvement of outcomes, as we illustrate in section 4. Of note, we do not require that players try to maximize the pay-off of others.

Finally, note that a no-harm equilibrium is a subgame perfect equilibrium *under the constraint* that it satisfies the no-harm principle. However, an NHE is *not* equivalent to

a strategy profile that is both a subgame perfect equilibrium and satisfies the no-harm principle, because in general there may be no SPE that satisfies the NHP, but as we show next an NHE always exists.

3. Existence, uniqueness, and Pareto optimality

In this section, we show that the no-harm equilibrium exists in normal form games, discuss under what conditions the uniqueness of the NHE outcome is guaranteed from a status quo, and illustrate the Pareto optimality of the NHEs.

Theorem 1. For every action profile a in a normal form game, there exists a no-harm equilibrium in pure strategies with respect to the initial status quo a.

Proof. First, we show that given an action profile a, the game $\Gamma(a)$ always possesses a pure subgame perfect equilibrium. This is true because $\Gamma(a)$ is a well-defined finite extensive form game with perfect information. To see this, notice that the root of the game is a and that there is a unique order of players who take their turns sequentially by construction of $\Gamma(a)$. Because there are finitely many players and we assume the no-cyclicity condition, the game $\Gamma(a)$ ends after finitely many steps. This implies that there is always a subgame perfect equilibrium in pure strategies.

Next, we assume that players act according to the NHP, which puts a constraint on their choices in $\Gamma(a)$. This implies that the extensive form game tree is essentially smaller under the NHP than the game tree of $\Gamma(a)$. The constrained game tree is still a well-defined finite extensive form game with perfect information. Let $\sigma^*(a)$ be a subgame perfect equilibrium in the constrained game, which exists by the same arguments as above. We next show that $\sigma^*(a)$ is a no-harm equilibrium in $\Gamma(a)$.

There are two cases to consider. First, consider the first player, say j, who moves at the root of $\Gamma(a)$. If $\sigma^*(a)$ prescribes player j to stay at a, which is the status quo, then this by definition satisfies the NHP. Moreover, staying at a must be farsightedly rational because otherwise there would be another outcome in the NHP-constrained game tree that gives j a strictly better pay-off, which cannot be because $\sigma^*(a)$ is an SPE in the NHP-constrained game tree. Second, If $\sigma^*(a)$ prescribes player j to move from the status quo a, then it must be that j strictly prefers $\sigma^*(a)$ to a in the NHP-constrained game tree. This concludes our proof that $\sigma^*(a)$ is a no-harm equilibrium in $\Gamma(a)$. QED

Theorem 2. For every action profile a in G, there is a unique no-harm equilibrium outcome.

Proof. Given an action profile a, the game $\Gamma(a)$ possesses a pure subgame perfect equilibrium as shown in the proof of Theorem 1. We next show that this subgame perfect equilibrium outcome is unique. The reason is that no matter which player moves on a (possibly terminal) node either (i) the player has a unique pure best response or (ii) the pure best responses all lead to the same outcome because the preferences of the players are strict in G. Thus, the subgame perfect equilibrium outcome in $\Gamma(a)$ is unique. Analogously, the subgame perfect equilibrium outcome in $\Gamma(a)$ whose game tree is constrained by the NHP must also have a unique outcome. Together with Theorem 1, this implies that the NHE outcome must be unique. QED

Theorem 3. A strategy profile is Pareto optimal if and only if it is a no-harm equilibrium.

Proof. We first show that if an action profile a is Pareto optimal then it is a no-harm equilibrium at a. To reach a contradiction, suppose that it is not, and the outcome of the no-harm equilibrium from a is given by $a' \neq a$. We know that the NHE is unique because the subgame perfect equilibrium is unique due to the fact that the preferences in G are strict (see Theorem 2). Then, a' would Pareto dominate a, which is a contradiction to the assumption that a is Pareto optimal. This is because if the players who choose to stay at or before a' prefer a' to a, then, together with the no-harm principle, a' Pareto dominates a. If those players who choose to stay do not prefer a' to a, then it is not rational for the players to stay at or before a'. Next, we show that for a generic action profile a that is not Pareto optimal, a no-harm equilibrium will lead to a Pareto optimal outcome. Suppose that $\sigma^1(a)$ is the no-harm equilibrium from a. Let $a^1 = \sigma^1(a)$ (with a slight abuse of notation) be the action profile that leads to the NHE from a. As above, a^1 Pareto dominates a, because otherwise the no-harm principle and rationality assumptions would preclude moving from a to a^1. Next, consider $\sigma^2(a^1)$ and let $a^2 = \sigma^2(a^1)$, which is the NHE from a^1. Again, a^2 must Pareto dominate a^1 or a^1 is Pareto optimal. This process will end at some k such that $\sigma^k(a^{k-1}) = a^k$ where a^k Pareto dominates a^{k-1}, and a^k itself is Pareto optimal because there are finitely many action profiles and hence there exists at least one Pareto optimal outcome. It is left to show that $a^k = a^2 = \sigma^1(a)$. This is true because no player would choose to stay at a^2 while (i) there is a path from a^2 to a^k, (ii) a^2 is the status quo, so the NHP applies, and (iii) all players, including the player who presumptively has chosen to stay, strictly prefers the latter to the former. QED

One might wonder what happens to the NHEs when there are indifferences between the outcomes in G. Then, it is well known that subgame perfect equilibria in Γ need not be unique, hence possible multiplicity of NHEs in G. In this case, for analogous reasons as in the proof of Theorem 3 we can conclude that every Pareto optimal profile would be an NHE. Moreover, for every profile there would always an NHE that is Pareto optimal, though some might be only weakly Pareto optimal.

4. Illustrations

The Prisoners' Dilemma

We first go back to the PD. It follows CC, CD and DC are Pareto efficient and so all are NHEs when they are status quo, but DD is not Pareto efficient and is not an NHE. Nevertheless, although it is clear DD is not Pareto efficient, it is perhaps not immediately obvious why deviation from DD satisfies both (1) NHP and (2) farsighted rationality.

Consider, for example, a deviation from DD by, say, Row to C. This deviation may seem to be precluded because it does not immediately satisfy Row's individual rationality—since Row is worse—off at CD. Nevertheless, to see whether it might satisfy Row's farsighted individual rationality (2), we need to consider what Column does at CD because CD may not be stopping point. Indeed, Column cannot 'stay' at CD because CD harms Row relative to the status quo of DD and so will not satisfy the no-harm principle (1). Hence if Column were to find themselves at CD, they would

have to 'move' and CC is the only option. Will Row stay at CC? CC satisfies (1) the NHP. It also satisfies Row's farsighted rationality (2) because a 'move' to DC would produce a cycle back to DD from which no further deviation would be permitted as per the no-cyclicity assumption. Thus since CC is better for Row than DD, it is also farsightedly rational for Row to 'stay' at CC. By analogous reasoning the same applies to Column who chooses to 'stay' and in this way CC becomes an NHE. In other words, you have to trace through what happens with a deviation by Row using DD as the status quo before you can see that (2) is also satisfied by the deviation of Row to C from status quo DD; and DD is not an NHE. Instead, CC is the NHE associated with the status quo of DD.

Stag-Hunt and Hawk and Dove

It is well-known that Pareto optimality and the Nash equilibrium are logically distinct concepts in the sense that neither concept is a refinement of the other. As we show in Theorem 3 the NHEs coincide with Pareto optimal profiles. Thus, there is no logical relationship between the set of NHEs and the set of Nash equilibria. Two further illustrations bring this out. In the Stag-Hunt game, NHE is a case of Nash refinement, and in the Hawk-Dove game, NHE expands the Nash equilibria; whereas, as we have just seen, in the PD the Nash equilibrium is not an NHE.

Consider, first, the Stag-Hunt game:

	Stag	Hare
Stag	4,4	1,3
Hare	3,1	2,2

Clearly, irrespective of the status quo the alternating offers game will end up at (Stag, Stag), which is the Pareto dominant profile and also a Nash equilibrium. (Hare, Hare) is a Nash equilibrium but not an NHE.

Next, in the Hawk and Dove (Chicken) game, (Dove, Dove) is a NHE as well as the two Nash equilibria (H,D) and (D,H):

	Hawk	Dove
Hawk	1,1	4,2
Dove	2,4	3,3

A three-person example

We next illustrate the no-harm equilibria in a three-person game presented in Figure 1, where player 1 chooses a row, player 2 a column, and player 3 a matrix. Throughout this example, we assume that (A, D, L) is the initial status quo whose associated outcome is (6, 1, 2).

First, assume that player 1 moves first, player 2 second, player 3 third, and so on. The on-path moves in the no-harm equilibrium can be described as follows. Player 1 moves

L	C	D
A	3, 5, 1	6, 1, 2
B	2, 2, 3	1, 3, 4

R	C	D
A	8, 4, 5	4, 7, 7
B	7, 8, 8	5, 6, 6

Figure 1. No-harm equilibria in a three-person game

to (1, 3, 4), player 2 moves to (2, 2, 3), player 3 moves to (7, 8, 8), player 1 moves to (8, 4, 5), player 2 stays at (8, 4, 5), and finally player 3 also stays, making (8, 4, 5) the outcome of the no-harm equilibrium, which is Pareto optimal. The reason why player 2 stays at (8, 4, 5) is that player 2 cannot gain any payoff by moving to (4, 7, 7). This is because player 3 cannot stay at (4, 7, 7) because it does not satisfy the no-harm principle with respect to (6, 1, 2). If player 3 moves to (6, 1, 2) from (4, 7, 7), then player 1 would have to stay because the no-cyclicity condition precludes player 1 from moving again to (1, 3, 4). Then, player 2's best response would be to move to (3, 5, 1), followed by player 3's move to (8, 4, 5), where player 1 would stay. Player 2 would also have to stay at (8, 4, 5) because it is not possible to implement another outcome that satisfies the no-harm principle with respect to the updated status quo (8, 4, 5). Second, assume that player 3 moves first, player 1 second, player 2 third, and so on. The on-path actions of the no-harm equilibrium is given as follows. Player 3 moves to (4, 7, 7), player 1 moves to (5, 6, 6), player 2 moves to (7, 8, 8), where player 3 stays. Then player 1 also stays because there is no other profile that satisfies the no-harm principle with respect to (7, 8, 8). Third, if player 2 moves first, 3 moves second, and player 1 third, then the outcome of the no-harm equilibrium would be (7, 8, 8). This is because player 2 would stay at (6, 1, 2), and then player 3 would move to (4, 7, 7). The rest is analogous to the second case.

At the outset, it looks like player 2 and player 3 should be able to implement their most preferred outcome (7, 8, 8) in the game. However, as shown above this is not possible if player 1 is the first-mover at the initial status quo. This three-person example illustrates that the player ordering can affect the NHE associated with any status quo, but the ordering does not affect the conclusion that NHEs are Pareto optimal.

The role of the no-harm principle

The following game is a simple example to show that NHP is essential for Theorem 3.

	L	R
<u>L</u>	4,3	1,4
R	2,1	3,2

Without the no-harm principle, Theorem 3 would not hold. To see this suppose play starts at (1,4) (i.e. it is the current status quo): on grounds of individual rationality alone, Row would move to (3,2), where Column as well as Row would stay, making it the final outcome in this extensive form alternating offer game. (3,2) becomes the new status quo and it is also the subgame perfect equilibrium of the extensive form alternating offer game based on this status quo, so it is the final outcome. This is because (i) Column would not gain by moving to (2,1) from (3,2) because Row would not move to (4,3) as Row anticipates that Column would then go back to (1,4), and (ii) if Row moves back to (1,4) from (3,2), Column would stay there as well, which makes

Row worse off. Thus, without the no-harm principle and starting at (1,4), players would move to (3,2), and this is Pareto dominated by (4,3).

The role of individual farsighted rationality

Individual rationality is also a necessary assumption for Theorem 3 because a strategy profile might satisfy the no-harm principle alone and yield a Pareto inferior outcome with respect to the initial status quo. The following 2×2 game provides a simple example.

2,2	0,3
1,0	4,4

Suppose that (2,2) is the status quo. Consider the strategy profile in which Column moves from (2,2) to (0,3) where Row stays, making (0,3) the updated status quo. Row's decision to stay satisfies the no-harm principle since it does not harm Column player. Next, Column moves back to (2,2) and Row moves to (1,0) where first Column stays and then Row stays, making (1,0) the outcome of the alternating offers game. Column's decision to stay at (1,0) satisfies the no-harm principle since it does not harm Row player with respect to the updated status quo (0,3). Knowing this and if the players were farsightedly rational, they would *not* stay at (1,0). But in the absence of the assumption of rationality, the aforementioned moves cannot be ruled out and it results in an outcome satisfying the non-harm principle, (1,0), that is strictly Pareto dominated by (2,2), where the players started from.

5. Discussion and relevant literature

The no-harm principle directs players to take account of other's interest in a very particular way and it is well known from models of altruism and other regarding preferences that the orientation to the interests of others in such models can turn the Pareto superior action profile of CC in the PD into a Nash equilibrium. Since the no-harm principle has the same effect in the PD, it is worth commenting on how the no-harm principle differs from such models of altruism and other regarding preferences.

The no-harm principle is a rule-like constraint on individual action that someone who believes in or subscribes to classic liberalism will wish to follow. In contrast, in models of altruism and more generally models where individuals have pro-social preferences, an individual personally values the pay-offs enjoyed by others and this enjoyment typically grows with the size of the pay-off to others. There is no such monotonic relationship between another's pay-offs and an individual's likelihood of taking an action with the no-harm principle. An action either satisfies the no-harm principle or it does not and in the one case the action is permissible and in the other it is not. In this respect the no-harm principle is akin to a version of rule rationality: that is, if the language of preference satisfaction is retained individuals have a lexicographic preference for following a rule(s) and so when they act to satisfy their preferences, they act in accordance with the rule(s).

The no-harm principle is related in classical liberal political philosophy to the presumption that the State should not intervene in individual decision making when the

consequences of those decisions apply only to the individual(s) making the decisions. This non-intervention principle has, of course, since Sen (1970) featured prominently in the social choice literature. The no-harm principle has also been used more recently in this social choice literature (see, e.g., Lombardi et al., 2016). Mariotti and Veneziani (2009; 2013; 2020) introduce a notion called "Non-Interference" principle which roughly says that society's preferences should not change following a change in circumstances that affect only one individual and everyone else's preferences remain the same.[5] Recently, Mariotti and Veneziani (2020) show that there is inconsistency between their "Non-Interference" principle and the Pareto principle (i.e., if everyone in a society prefers an alternative x to y, then society should prefer x to y) in a non-dictatorship.

Our formalization of the no-harm principle differs from Mariotti and Veneziani's in two main respects, the framework and the conceptual definition. The most obvious difference between the two principles is that ours applies to actions within a game theoretical framework whereas theirs applies to the preferences within a social choice context. Conceptually, under our no-harm principle a player is allowed to choose any action as long as this action does not eventually harm (and may benefit) other players with respect to the status quo. However, the "Non-Interference" principle does not apply to a change in social situations that leave some members of the society better off.

Although we share the interest in the implications of subscribing to the tenets of classical liberalism with the social choice literature, the approach here is very different. We are *not* interested in the implications of classical liberalism for a social planner—as is the case in the social choice literature. Instead, we are concerned with the introduction of the no-harm principle as a constraint on individual decision making in games affects the equilibrium outcomes of those games.

In a recent and related development, Che, Kim, Kojima, and Ryan (2020) provided a characterization of the Pareto optima via utilitarian welfare maximization. While both our and their approaches are sequential in nature, the main difference between the two papers is that their framework is non-strategic whereas we provide a non-cooperative foundation for Pareto optimality via the no-harm principle.

In modeling the no-harm principle, we have followed the approach of Brams (1994). In this same tradition, Brams and Ismail (2019) show that there is always a non-myopic equilibrium that is Pareto optimal; though, not all Pareto optimal profiles are non-myopic equilibria, and not all non-myopic equilibria are Pareto optimal. The main difference in our results comes from the no-harm principle that we assume, which restricts the actions of the players to ones that satisfy the well-known principle of classical liberalism.

6. Conclusion

Game theory standardly makes no assumption about what motivates individuals to act other than they have preferences they seek to satisfy. While this is an admirably parsimonious assumption, it is also misleading when people either subscribe to the

[5] The formal definition of the Non-Interference notion is a little bit stricter than our verbal description.

political philosophy of classical liberalism or live in a society that is legally founded on the principles of classical liberalism. Such people are additionally constrained either legally or by their own beliefs to the no-harm principle. This is because the principle is the key constraint placed on the exercise of individual freedom by J. S. Mill in his classic manifesto for individual liberty: *On Liberty*. Thus, for those who live in a classically liberal society and/or who believe in classical liberalism, a question naturally arises: how is behaviour in games affected by the additional individual constraint on action supplied by the no-harm principle? We offer part of answer to this question.

We show with our operationalization of the no-harm principle, that this addition dramatically alters the predictions regarding what happens in games. In particular, the no-harm equilibria are always Pareto optimal. This stands in marked contrast to standard game theory where there is no necessary connection between Nash equilibria and Pareto optimality. It is important in the derivation of this result to note that our operationalization of the no-harm principle does not require Pareto optimality; nor does it even secure Pareto improvements from a starting position. It is the combination of individual rationality with the no-harm principle that secures Pareto optimality.

Our paper opens up two main directions for future theoretical research. First, what are the other frameworks in which strategic foundations of Pareto optimality can be studied? Second, we suggest a further exploration of classical liberal principles including the no-harm principle in game theory. The definitions of no-harm principle and the no-harm equilibrium can be extended to games under incomplete and perfect information as subgame perfect equilibrium is well-defined under these settings. Does the Pareto optimality of no-harm equilibria remain valid in those games?

It also suggests an important new direction for empirical research. It is well known from experiments, for example, that some people behave selfishly in public goods/PD interactions and others behave pro-socially by contributing to the public good. The pro-social contributions are typically understood through the prism of social preferences and selfishness is understood through the absence of such preferences. To what extent, then, might they be better understood through the differing sway or influence that the no-harm principle has on individuals? In particular, while the puzzle from these experiments from the perspective of standard game theory has centred on why subjects contributed anything to the public good, it changes with the result of this paper. Rather, the puzzling question becomes: why do so many subjects in these experiments, when they come from liberal societies, behave selfishly?[6]

[6] Amadae (2016) has an answer to this question: the rise and influence of Game Theory. In fact, the seeds of the analysis in this paper were sown by Amadae (2016) and Hargreaves Heap (2016). Amadae (2016) argues that game theory has encouraged a form of neo-liberalism that is distinct from classical liberalism precisely because game theory dispenses with the no-harm principle. She conjectures that the no-harm principle would dramatically alter the prediction of what rational individuals would do in a prisoners' dilemma. Hargreaves Heap (2016) reviewed this book and found this conjecture intuitively plausible and so, in effect, reproduced it in the review.

References

Amadae, S. M. (2016). *Prisoners of Reason: Game Theory and Neoliberal Political Economy*. Cambridge University Press.

Brams, S. J. (1994). *Theory of Moves*. Cambridge, UK: Cambridge University.

Brams, S. J., and Wittman, D. (1981). Nonmyopic equilibria in 2×2 games. *Conflict Management and Peace Science*, 6(1), 39–62.

Brams, Steven J., and Donald Wittman (1981). "Nonmyopic Equilibria in 2 × 2 Games." *Conflict Management and Peace Science* 6, no. 1 (Fall): 39–62.

Brams, S. J., and Ismail, M. S. (2019). Farsightedness in Games: Stabilizing Cooperation in International Conflict. MPRA Working Paper.

Che, Y. K., Kim, J., Kojima, F., and Ryan, C. T. (2020). Characterizing Pareto Optima: Sequential Utilitarian Welfare Maximization. *arXiv preprint arXiv:2008.10819*.

Gibbard, A. (1974). A Pareto-consistent libertarian claim. *Journal of Economic Theory*, 7(4), 388–410.

Hargreaves Heap, S. (2016). Review: Prisoners of Reason: Game Theory and Neoliberal Political Economy, by S. M. Amadae. *Journal of Economic Literature, 54*(4), 1392–1394.

Lombardi, M., Miyagishima, K., and Veneziani, R. (2016). Liberal egalitarianism and the Harm Principle. *The Economic Journal*, 126(597), 2173–2196.

Mariotti, M., and Veneziani, R. (2009). 'Non-interference' implies equality. *Social Choice and Welfare*, *32*(1), 123–128.

Mariotti, M., and Veneziani, R. (2013). On the impossibility of complete non-interference in Paretian social judgements. *Journal of Economic Theory*, 148(4), 1689–1699.

Mariotti, M., and Veneziani, R. (2020). The Liberal Ethics of Non-Interference. *British Journal of Political Science*, 50(2), 567–584.

Mill, J. S. (1859). *On Liberty*. London: JW Parker & Son.

Nash, J. (1953). Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, 128–140.

Osborne, M. J., and Rubinstein, A. (1994). *A Course in Game Theory*. MIT press.

Peacock, A. T., and Rowley, C. K. (1972). Pareto optimality and the political economy of liberalism. *Journal of Political Economy*, 80(3, Part 1), 476–490.

Selten, R. (1965). Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit: Teil i: Bestimmung des dynamischen preisgleichgewichts. Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics, (H. 2), 301–324.

Sen, A. (1970). The impossibility of a Paretian liberal. *Journal of Political Economy*, *78*(1), 152–157.