Social Contracts for Non-Cooperative Games

Alan Davoust
Université du Québec en Outaouais
Gatineau, Quebec, Canada
alan.davoust@uqo.ca

Michael Rovatsos
University of Edinburgh
Edinburgh, U.K.
michael.rovatsos@ed.ac.uk

ABSTRACT

In future agent societies, we might see AI systems engaging in selfish, calculated behavior, furthering their owners' interests instead of socially desirable outcomes. How can we promote morally sound behaviour in such settings, in order to obtain more desirable outcomes? A solution from moral philosophy is the concept of a *social contract*, a set of rules that people would voluntarily commit to in order to obtain better outcomes than those brought by anarchy. We adapt this concept to a game-theoretic setting, to systematically modify the payoffs of a non-cooperative game, so that agents will rationally pursue socially desirable outcomes.

We show that for any game, a suitable social contract can be designed to produce an optimal outcome in terms of social welfare. We then investigate the limitations of applying this approach to alternative moral objectives, and establish that, for any alternative moral objective that is significantly different from social welfare, there are games for which no such social contract will be feasible that produces non-negligible social benefit compared to collective selfish behaviour.

CCS CONCEPTS

• **Computing methodologies** → **Cooperation and coordination**; *Philosophical/theoretical foundations of artificial intelligence*; *Multi-agent systems.*

KEYWORDS

Game theory, ethics, moral philosophy, agents

ACM Reference Format:
Alan Davoust and Michael Rovatsos. 2020. Social Contracts for Non-Cooperative Games. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375829

INTRODUCTION

As AI systems and autonomous agents begin to take more active roles in our society, there is renewed interest in the problem of instilling human moral values into artificial agents [1, 11], and in how to implement ethical decision-making algorithms aligned with such values [10, 15, 20, 25].

A recurring theme in this debate is the need for societal oversight through mechanisms such as public consultations, audits, and regulation [27, 36]. While such oversight appears essential for high-stakes contexts such as automated warfare or autonomous vehicles, there is arguably also a need to establish high-level moral guidelines that can be applied *without* human intervention in more mundane contexts such as route-finding in traffic, or energy management in the smart grid.

Assuming that autonomous agents are tasked to further their owners' interests, situations will most likely occur where the agents' self-interested goals are not aligned with the greater social good. *Social dilemmas* then arise, where the equilibrium between rational, self-interested strategies creates outcomes that are bad for everyone involved, as in the classic Prisoners' Dilemma [3]. In such games, the Nash Equilibrium, assumed to predict what rational strategic agents would choose to do, does not maximize the total payoffs of the players, i.e. it is not *efficient*. This problem is known to exist across many games, including traffic routing games, auctions, and others [7, 32].

Given this problem, an agent in such a game faces a moral dilemma, between doing what is considered *right* (at least in the classic utilitarian view of morality, i.e. maximising the sum of everybody's "happiness"), and what is best for itself (or its human owner). How could such agents be incentivized to *do the right thing*?

There are two general approaches to this problem: One is to program agents to behave according to particular moral values, e.g. applying specific solution concepts for games [21]. However, the designer's choice – to program agents in this way or as self-interested optimizers instead – then becomes a moral dilemma itself. The second approach is to alter the game in such a way that the moral value of an action aligns with the agents' rationality. This modification to the game comes as a scheme of rewards and punishments that modify the game's payoffs: examples include Pigouvian taxes [26] and approaches to solving congestion problems in networks [8, 12, 22]. However, the assumption that such rewards and punishments can be imposed on the game by some external authority (e.g., tolls to be set up and enforced on roadways), is a strong one, and limits the applicability of the approach.

In this paper we investigate a way of making such changes possible, which is to design the changes in an *incentive-compatible* way, i.e. so that all of the agents prefer the modified game to the original. This can be achieved for example by using only rewards, as in the k-implementation approach [24], but the resulting schemes are then costly to implement.

We propose a solution inspired by the notion of a *social contract*, aiming to satisfy both incentive-compatibility and budget balance. The key idea of a social contract is that a society puts a government in place and *willingly* submits to its authority in order to avoid anarchy. Transposed to the context of agent interactions, the idea

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375829

is that an agent society can willingly adopt rules that encourage morally sound behaviour, if these rules are designed in such a way that all agents are better off under the rules.

We formalise the idea in a game-theoretic framework: a social contract is a scheme to modify a game through rewards and punishments, such that (i) the morally desirable actions in the modified game are the rational choice for self-interested players; (ii) all agents rationally prefer the modified game to the original game; and (iii) the scheme does not cost anything beyond the utility produced by the game itself.

One difficulty in this formalisation is that we must define what morally sound (or "right") actions are. We take a consequentialist view of morality, where the moral value of an action is judged according to the desirability of the outcome it produces. In the classic utilitarian view [23], the moral objective is to maximize social welfare, i.e. the sum of payoffs across all players, and the players' moral imperative is to choose actions that will bring about this objective.

However, within this consequentialist view, we can adopt other moral objectives, such as Rawlsian "maximin" fairness [28], which aims to maximize the welfare of those who are worst off, egalitarian outcomes (aiming for equal payoffs for all), or *Nash welfare*, which aims to maximize the product of utilities [5]. For each of these objectives, we can define the players' moral imperative accordingly.

As it turns out, this choice of a moral objective is crucial to the feasibility of our approach. Our first result is that with social welfare as a moral objective, a social contract that meets our requirements is feasible for any game, under mild conditions on the players' rationality. Unfortunately, for any other moral objective, there will be games where our requirements for a social contract cannot be met. Finally, we show that 'weaker' social contracts (improving the situation but falling short of optimality) usually exist for those objectives, although their social benefits may be arbitrarily small.

RELATED WORK

In political economy, Pigou [26] set an important foundation with the idea of taxing negative externalities, i.e. costs of an economic player's actions borne by external parties. Pigouvian taxes force economic players to face both the costs and the benefits of their actions, incentivizing them to maximize the net utility that their actions produce. This idea has been applied to a variety of games, including the management of road traffic [31, 33].

Following the idea of pricing roads to manage congestion, the inefficiency of equilibria in more abstract congestion games has been extensively studied [8, 12, 30, 33]. The problem is to devise a scheme of taxes on the resources involved in the game, in order to minimize congestion. These schemes minimize overall costs, provided that the taxes can be returned to the players: otherwise the losses incurred by taxes may offset the benefits of routing efficiency [9].

All of these approaches assume social welfare maximization as the overall objective, and assume that the chosen solution can be enforced by some external authority. The only exception in the literature, to our knowledge, is the k-implementation problem [24], which is defined as the problem of implementing an arbitrarily chosen outcome in a game at a minimal cost. The scenario considered is that an external party to the game wants the players to choose some particular joint action, and offers rewards (strictly positive utility) to specific players for specific joint actions. The goal is to alter the game to make the chosen action a dominant strategy equilibrium, at a minimal cost. It is shown that an action can be implemented for free if and only if it is a Nash equilibrium. In our setting, we require the implementation to be free (*no deficit* requirement), but gain some flexibility by allowing both rewards and punishments.

Similar problems have been tackled in the area of multi-agent learning (MAL) through *reward shaping*, which focuses more on the way strategic agents learn some globally desirable behaviour based on rewards received during the learning process [17, 35]. However, in this setting rewards do not need to have a tangible reality and budget considerations can safely be ignored.

Where our approach differs from most existing work is in investigating the individual and collective rationality of the agents joining the modified game. In this sense our problem resembles *mechanism design* [18], in the sense that we want to implement a globally desirable outcome given agents' individual choices. However, classical mechanism design results do not apply to our setting, since we only consider the possibility of redistributing the game payoffs (we cannot design an entirely new game or use a direct-revelation mechanism). We also assume publicly known utility functions, and leave private ones for future work.

Rationality considerations are also prominent in the literature of social contract theory. Buchanan [4] models the political social contract as a situation where the parties seeking to create a 'state' are faced with a simple Prisoners's Dilemma: Individual and joint rationality lead them to agree to cooperate and establish an institution to enforce the terms of the agreement. Heckathorn and Maser [16] argue that the agreement to cooperate is the result of a bargaining process where the players settle on one of many possible contracts, and that this choice will be dictated by a form of *political rationality*, where the chosen contract must be compatible with each party's belief in its bargaining strength. Gibbons [14] considers a similar model, but views the ruler instituted by the social contract as an additional agent with its own incentives and payoffs, and argues that the terms of the contract should be viewed as an equilibrium in a repeated game between all parties including the ruler.

These approaches resemble our model, although contrary to Gibbons we do not consider the presence of an external ruler as key: in a setting of autonomous agents, purely technical mechanisms could be used to enforce short-lived agreements. We build on these ideas by exploring how they apply to different models of morality and of rationality.

The final connection worth mentioning is to cooperative game theory, which considers how agents should share the benefits of cooperation – relying on the same assumption of transferable utility. In a sense, our agents must redistribute these benefits in such a way that all agents are willing to participate: this is in essence the definition of the *core* for a cooperative game [13]. However, cooperative game theory is not concerned with how the value of a coalition is derived from a non-cooperative game definition, or how different models of rationality might impact the value of a coalition

(e.g. because of how different coalitions might strategically interact). To avoid these complexities, we consider an "all or nothing" model, where either all agents agree to the social contract, or just strategically act on their own, which allows us to explore a wider range of possible models of rationality.

FORMAL SETUP

We consider a normal-form game setting where a game $G = \langle N, A, u \rangle$ involves n agents $N = \{1, \ldots, n\}$, where each agent i has a set of possible actions $A_i = \{\alpha_1^i, \ldots, \alpha_k^i\}$ and $A = \times_{i=1}^n A_i$. The set of strategy profiles (or strategies) for player i, S_i, is the set of probability distributions over A_i, and $S = \times_{i=1}^n S_i$ is the set of strategy profiles for the game.

When a joint action $a \in A$ is played, each agent i receives a payoff $u_i(a)$ given by a utility function $u_i : A \to \mathbb{R}$. We assume that utility is *transferable* between agents, in the form of money or some similar currency. For simplicity we slightly abuse notation and denote the expected payoff for each player i under strategy profile s as $u_i(s)$.

The study of inefficient equilibria [19] for different games is based on two fundamental assumptions, (1) that *rational* agents will play Nash equilibrium strategies in the game, and (2) that the *moral* action would be to play a social welfare-maximizing strategy instead. We consider each of these in turn.

The assumption that agents will play a Nash equilibrium in any game is present in most of the related literature, and is reasonable when a single equilibrium in dominant strategies exists: it only really requires agents not to play dominated strategies. It is harder to justify for arbitrary games with possibly many Nash equilibria (infinitely many in some cases).

What, then, is a good model of rational (self-interested) behaviour? For the software agents that populate our considered agent society, *some* strategy selection algorithm must be implemented, and unfortunately there is no clear "best" solution. We could consider various refinements of Nash equilibria, or a learning algorithm.

In order to account for the agents' decision procedure and explicitly discuss its properties, we represent it by a function $\pi : \mathcal{G} \to S$ ($\mathcal{G}$ is the set of all games) whereby the agents will select a strategy profile $s = \pi(G)$. In our setting it will be important that this function is *common-knowledge* and *deterministic*. This does not tie our work to any particular solution concept, but implies that the mapping of a game $G \in \mathcal{G}$ to the (possibly mixed) strategy profile $\pi(G)$ is known by all agents, who can then correctly predict their expected payoff from playing G. This removes *strategic* risk [29], which is problematic in our setting because the associated uncertainty does not follow well-defined probabilities.

However, we do assume throughout the paper that the agents are *minimally rational*, in the sense that they will not play dominated strategies or expect others to do so.

Another aspect of the agents' rationality is that their beliefs of the other players' intentions induces a preference relation between games. We denote that an agent i (weakly) prefers some game G_1 to another game G_2 as $G_1 \succeq_i G_2$, based on a comparison between $u_i(\pi(G_1))$ and $u_i(\pi(G_2))$.

The second assumption, that agents should play a social-welfare maximizing action, reflects a classic utilitarian view of morality. As discussed previously, this can conflict with alternative notions of fairness such as that put forward by [28]. In order to explore alternative moral objectives, we formalize the moral objective in our framework as a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ that associates a "moral" value with a payoff vector. A morally optimal action $a^* = \arg\max_{a \in A^*} f(\boldsymbol{u}(a))$ for a game G is any action that maximises $f(\boldsymbol{u}(a))$, where $\boldsymbol{u}(a) = \langle u_1(a), \ldots, u_n(a) \rangle$.

SOCIAL CONTRACT DESIGN

Suppose that the strategy profile selected by agents in the game is not optimal with respect to the moral objective f:

$$f(\boldsymbol{u}(\pi(G)) < f(\boldsymbol{u}(a^*))$$

We would like to modify the game so that agents will have an incentive to behave in ways that will bring about the moral objective, and do so in a way that all agents will (rationally) prefer the modified game to the original one. Additionally, we would like the scheme to be implemented using only the utility generated by the original game. The resulting problem can be defined as follows:

Definition 1 (Social Contract Design Problem)**.** Given a game $G = \langle N, A, u \rangle$, a moral objective f, and a decision procedure π, the *social contract design problem* is to find a modified game $G' = \langle N, A, u' \rangle$ such that:

(Effective Deterrent) the strategy profile chosen by the agents in G' has the same moral value as the optimal joint action in G:

$$f(\boldsymbol{u'}(\pi(G')) = f(\boldsymbol{u}(a^*))$$

(No Deficit) The game only redistributes the utility generated by the original game:

$$\forall a \in A. \sum_i u_i'(a) \leq \sum_i u_i(a)$$

(Individual Rationality) Agents (weakly) prefer G' to G:

$$\forall i \in N. G' \succeq_i G$$

The *effective deterrent* requirement reflects the primary purpose of the social contract, which is to deter agents from choosing those actions deemed immoral (per the chosen moral standard f) in the original game, instead choosing a morally optimal joint action a^*. However, since the social contract involves a scheme of punishments and rewards that modify the payoffs of a^*, the moral value of this action may have changed in G'. To account for this, we must reformulate our objective to focus on moral value rather than on the specific action: if the morally optimal action a^* in G had some moral value $f(\boldsymbol{u}(a^*))$, our aim is to ensure that agents choose some strategy profile s in G' with the same moral value (in expectation) as a^* has in G.

To avoid implementing schemes that would be costly to implement (e.g. simply giving agents a large enough reward to play the morally optimal action), the *no deficit* requirement ensures that rewards and punishments be produced by redistributing the utility produced by the game, i.e. the payoffs of the original game G. In fact, as we will see, in most cases we can achieve the stronger requirement of *budget-balance*, i.e. that any punishment be redistributed to other players as reward.

The intuition behind *individual rationality* is that we want agents to *voluntarily* commit to the social contract, i.e. G' should be designed in such a way that the players will (rationally) prefer playing G' over G. We interpret this as meaning that agents should have a higher expected payoff playing their role in G' than in G. The difficulty is, however, that the expected payoff of playing a game is not well defined for an arbitrary game. We must therefore consider the implications of the players' (joint) decision function π.

Finally, as we will show below, meeting all these requirements may not be possible in certain games for certain moral objective functions. However, it may still be worth creating a social contract, if we can ensure that the modified game will be, if not optimal, at least better than the original game in terms of the moral objective.

This problem is defined as the design of a *weak* social contract, which is identical to a (strong) social contract, except that it substitutes the effective deterrent requirement with a weaker version, whereby

$$f(\boldsymbol{u}(a^*)) > f(\boldsymbol{u'}(\pi(G')) > f(\boldsymbol{u}(\pi(G)))$$

i.e. the resulting moral value of (the chosen action in) G' is higher than that of G, but less than the value of the *optimal* action in G.

RESULTS

Social Welfare Based Moral Value

Our main result is that we can guarantee the feasibility of a social contract under two conditions: (1) the agents' decision procedure π must be *common-knowledge* and *deterministic*; and (2) the moral objective f must correspond to maximizing social welfare, i.e. either f is equated with social welfare, i.e.

$$f(\boldsymbol{u}(a)) = \sum_i u_i(a) \quad .$$

or equivalently f is expressed as a strictly increasing function of social welfare, i.e.

$$f(\boldsymbol{u}(a')) > f(\boldsymbol{u}(a)) \Leftrightarrow \sum_i u_i(a') > \sum_i u_i(a)$$

As discussed previously, the first condition ensures that the agents can correctly predict which strategies the other agents will select, and thus compute their expected payoff from playing G. Their preference relation between games is then straightforward: they will prefer any game over G where they (believe that they) will obtain more than in $\pi(G)$.

The second condition means that a^* maximizes social welfare: even if f is not equated with social welfare, the condition guarantees that it has the same maxima. Therefore, a^* produces (non-negative) surplus utility compared to $\pi(G)$. This surplus can be redistributed to all the players, so that by playing a^* in the new game, each player obtains as much as in the original equilibrium $\pi(G)$, plus a share of the surplus. The payoffs for all other joint actions can then be redistributed in proportion to the new payoffs of a^*, making a^* dominant, unless there are several social welfare-maximizing actions. In this case, we have multiple Nash equilibria, and a coordination problem. In order to ensure that the players coordinate on one of these joint actions (it can be chosen arbitrarily), a small amount of utility can be deducted from the payoffs in all of the others, thus restoring strategic dominance of the chosen action.

We formalize this result with the following theorem.

Theorem 1. *With social welfare as a standard of moral value, and any deterministic, common-knowledge decision procedure π, for any game G there exists a solution to the social contract design problem.*

Proof. For any arbitrary game G, the modified game G' can be constructed as follows:

(1) The common-knowledge decision procedure π gives us the strategy profile $\pi(G)$ that the players would use in G. Compute its expected payoff $u_i(\pi(G))$ for each player i.
(2) Select an arbitrary social welfare-maximizing joint action a^* of G.
(3) Compute the value
$$\sigma = \sum_i u_i(a^*) - \sum_i u_i(\pi(G))$$
It is easy to see that $\sigma \geq 0$, by the linearity of expectation, and the inequality is strict unless the strategy profile $\pi(G)$ maximizes social welfare. Intuitively, σ is the surplus social welfare produced by a^* compared to the players' payoff expectations in G given π.
(4) Define the payoffs of a^* in G' as follows:
$$u'_i(a^*) = u_i(\pi(G)) + \frac{\sigma}{n}$$
(5) For all other joint actions $a_j \neq a^*$, $u'_i(a_j)$ is obtained by scaling $u'_i(a^*)$ to the social welfare produced by a_j:
$$u'_i(a_j) = \frac{\sum_i u_i(a)}{\sum_i u_i(a^*)} u'_i(a^*)$$
.
(6) If a^* was the only social welfare-maximizing action in G, then a^* is now a dominant-strategy equilibrium. Otherwise, we subtract a small amount ϵ from all players' payoffs for all social welfare-maximizing actions except a^*; making this action a dominant-strategy equilibrium.

It is easy to see how this construction satisfies the effective deterrent requirement: $\pi(G')$ is the pure strategy profile a^*, and we have

$$\begin{aligned} f(\boldsymbol{u'}(a^*)) &= \sum_i (u'_i(\pi(G)) + \frac{\sigma}{n})\sigma + \sum_i u_i(\pi(G)) \\ &= \sum_i u_i(a^*) \qquad \text{(per the definition of } \sigma) \\ &= f(\boldsymbol{u}(a^*)) \end{aligned}$$

It is also clear that the *no deficit* requirement is satisfied: the total utility produced by an action in G is simply redistributed, as shown for a^* above. The final tweak of subtracting ϵ from the payoffs of other social-welfare maximizing actions makes this scheme *not* budget-balanced. Budget balance could be preserved by implementing an external coordination device for a^* [2].

Individual rationality is obtained by giving every player in G' their expected payoff in the original game G[1]. The fact that $\pi(G)$ is common knowledge and deterministic implies that all players can compute their expected payoff from playing G and will reach the same conclusion. This guarantees that the sum of the values they calculate coincides with the expected payoff of an actual strategy

[1]Note that *any* redistribution of the surplus creates a social contract satisfying the stated requirements: in this proof we simply distribute it equally to all players (step 4), but other distributions could be considered.

profile of G, so that the inequality $\sum_i u_i(a^*) \geq \sum_i u_i(\pi(G))$ holds. □

Alternative moral objectives

We now show that the positive results obtained when applying social welfare as a more standard are invalidated when applying any alternative moral standard. Intuitively, a moral objective is different from social welfare if we can find some game with two different actions, where one action is preferred under social welfare, and the other is preferred under the other objective. Formally, we will say that a moral objective f differs significantly from social welfare if:

$$\exists(\boldsymbol{v}, \boldsymbol{w}) \in \mathcal{R}^n \times \mathcal{R}^n, \sum_i v_i > \sum_i w_i \wedge f(\boldsymbol{v}) < f(\boldsymbol{w})$$

Theorem 2. *For any moral value function that differs significantly from social welfare, there exist games for which no solution to the social contract design problem exists.*

Proof. We construct a counter-example as follows:

(1) Since f differs significantly from social welfare, there must exist payoff vectors of length n, $\boldsymbol{v}$ and $\boldsymbol{w}$, such that $\sum_i v_i > \sum_i w_i$ and $f(\boldsymbol{v}) < f(\boldsymbol{w})$. Since $\sum_i v_i > \sum_i w_i$, there is at least one component k of these payoff vectors such that $v_k > w_k$.
(2) Let G a game with n players and two actions α and β for each player. The game only has two possible outcomes: if player k plays α then the payoffs are as in the vector $\boldsymbol{v}$ defined above, otherwise they are as in $\boldsymbol{w}$. Since $v_k > w_k$ player k's dominant strategy is to play α (with payoffs as in $\boldsymbol{v}$), and all players should expect this to happen with certainty. On the other hand the morally optimal actions are those where player k plays action β.

We can now prove that there is no suitable social contract for this game. Suppose that there existed a modified game G' that met the stated requirements. Let $\boldsymbol{v'}$ the equilibrium payoffs in G'. To satisfy the *effective deterrent* requirement, they must have the same moral value as $\boldsymbol{w}$, which is different from (strictly greater than) the moral value of $\boldsymbol{v}$. Therefore the payoff vector $\boldsymbol{v'}$ must be different from $\boldsymbol{v}$, and since $\sum_i v'_i \leq \sum_i v_i$ (*no deficit* requirement), we have $\sum_i v'_i < \sum_i v_i$ and there must be one component j such $v'_j < v_j$. Therefore, player j cannot rationally prefer game G' to game G, which contradicts the assumption that G' satisfied the requirement of individual rationality. □

This negative result tells us that for some games, we will not be able to establish a social contract that would produce a morally optimal outcome, unless moral value is maximized by the same actions that maximize social welfare (rendering the distinction between the two objectives moot).

The key difficulty in this case is that in order to improve fairness[2], we might be forced to decrease some players' equilibrium payoffs, who would not agree to this change.

[2]Here and in the following discussion (where it is clear from the context), we will use the term "fairness" to refer to an alternative moral value function, e.g. maximin fairness or Nash social welfare.

Notably, our counter-example illustrates the extreme case, where a dominant strategy *maximizes* social welfare in the original game: as a result we simply cannot modify the payoffs without breaking the individual rationality requirement. However, if the equilibrium outcome of the original game does not maximize social welfare, then we can at least create a *weak* social contract as defined above, unless the new fairness objective is somehow "incompatible" with an increase in social welfare.

Intuitively, if the chosen action in G doesn't maximize social welfare, then it means there is a strictly positive surplus that can be redistributed to improve fairness, and the "compatibility" notion means that there exists *some* allocation of this surplus that will actually advance the moral objective. For example, if our moral objective is to maximize maximin fairness, then the surplus should be allocated to whichever player has the lowest payoff.

The requirement that a moral objective function is *compatible* with social welfare can be formalized by the following (sufficient) condition on the gradient of f: at any point in $\mathbb{R}^n$, f must be differentiable and ∇f must have at least one positive coordinate i: this means that allocating positive utility to the player i will increase f (at least locally).

This now gives us the following theorem:

Theorem 3. *With any a standard of moral value compatible with social welfare, and any common-knowledge deterministic decision procedure π, for any game G where $\pi(G)$ does not maximize social welfare, there exists a solution to the weak social contract design problem.*

Proof. The proof is a minimal adaptation from the proof of theorem 1, where the construction is the same except for step 4: in step 4, (all or part of) the surplus σ is distributed to the different players in whichever way maximizes the considered moral objective. An increase in the objective function f is guaranteed by the condition that the moral objective is *compatible* with social welfare, and the *weak effective deterrent* is thus satisfied. The other conditions are satisfied in the same way as in theorem 1. □

Thus, although it may not always be possible to obtain morally optimal outcomes through a social contract, it is still almost always worth establishing a social contract because there will almost always be *some* improvement to be gained. Our final result qualifies this statement, by showing that the resulting improvement can be arbitrarily small.

What we mean specifically is that given a game where the equilibrium outcome is "unfair" (i.e. has a low value according to a moral objective such as maximin utility), and there is another action that is "fairer", the fairness of the latter action has absolutely no bearing on what social contracts will be feasible: the limiting factors to the feasible social contracts are (i) the utility surplus available from the social welfare maximising action, and (ii) the fairness of the original outcome $\pi(G)$. This is because the latter is a starting point to define the equilibrium payoffs of the modified game, and the former will determine how little or how much we can improve over the original game.

We can formalise this result by comparing the difference between the fairness of the chosen action $\pi(G)$ and that of the fairest action

a^* with the improvement provided by the best feasible social contract. We consider a game G, a common-knowledge deterministic decision procedure π, and a moral objective function f which differs significantly from social welfare and is continuous over $\mathbb{R}^n$. We assume that the outcome $\pi(G)$ is not optimal in terms of f, so that there is a non-zero "cost of anarchy" $c = f(u(a^*)) - f(u(\pi(G)))$ (which would be zero under a (strong) social contract). We also assume that $\pi(G)$ does not maximize social welfare.

Theorem 4. *For any $\epsilon > 0$ there exists a game G such that: if c is the cost of anarchy, and m the maximal improvement of f afforded by a feasible social contract on G, then $\frac{m}{c} < \epsilon$.*

Proof. As in theorem 2, we have some payoff vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, such that $\sum_i v_i > \sum_i w_i$ and $f(\boldsymbol{v}) < f(\boldsymbol{w})$, and $v_j > w_j$ for some j, and define a game G where $\boldsymbol{v}$ and $\boldsymbol{w}$ are the payoff vectors for the social-welfare maximizing outcome and the fairest outcome, respectively. We define a third possible outcome with payoff vector $\boldsymbol{z}$ as the equilibrium payoff of G. The point of the proof is to choose $\boldsymbol{z}$ that achieves the desired property.

As $c = f(\boldsymbol{w}) - f(\boldsymbol{v})$, we want $\frac{m}{c} < \epsilon$ to hold, i.e. $m < c\epsilon$.

Let $\epsilon_1 = c\epsilon$. Since f is continuous around $\boldsymbol{v}$, for any value ϵ_1 there exists some $\epsilon_2 > 0$ such that:

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{v} - \boldsymbol{x}\|_1 < \epsilon_2 \Rightarrow |f(\boldsymbol{v}) - f(\boldsymbol{x})| < \epsilon_1$$

where $\|\|_1$ represents the city block distance on $\mathbb{R}^n$

Assume player j is the only player whose actions affect the outcome, and has three actions available, which will lead to the three outcomes $\boldsymbol{v}$, $\boldsymbol{w}$, and $\boldsymbol{z}$ respectively. We can now define $\boldsymbol{z}$ as: $z_j = v_j + \frac{\epsilon_2}{2n}$, and $\forall i \neq j, z_j = v_j - \frac{\epsilon_2}{n}$. We now have $z_j > v_j > w_j$, meaning that the action leading to payoff $\boldsymbol{z}$ is dominant for player j.

Since $\boldsymbol{z}$ is the equilibrium payoff, the surplus social welfare is $\sigma = \sum_i v_i - \sum_i z_i < \epsilon_2$, and m is entirely determined by the improvement in fairness that can be achieved by allocating this surplus to different components of $\boldsymbol{v}$. We obtain $\|\boldsymbol{v} - \boldsymbol{z}\|_1 < \epsilon_2$, and thus $|f(\boldsymbol{v}) - f(\boldsymbol{z})| < \epsilon_1 = c\epsilon$.

□

DISCUSSION AND CONCLUSION

In this paper we define a social contract scheme for agent societies: a systematic modification of a game that incentivizes morally sound behaviour, with two key additional requirements: agents would (rationally) voluntarily enter this contract, and the scheme must not cost anything to implement (beyond the utility produced by the game outcomes).

The main idea is that morally sound behaviour, which by default is defined as maximizing social welfare, produces more utility than what self-interested agents would have chosen in the original game. This surplus utility can be redistributed to the agents as a reward for choosing a moral action. The rewards make agents prefer the new game, and are generated by the game itself, thus satisfying both our key requirements.

However, the results do not carry over to other moral objectives, and in all cases there are difficulties around the rationality assumptions that we ascribe to the agents, and their implications for our individual rationality requirement.

This requirement states that the social contract must be designed so that rational agents prefer the modified game to the original, and our intuitive solution is that the utility surplus generated by moral actions can provide them with higher expected utility in playing the modified game (where the moral objective is a dominant-strategy equilibrium).

However, the difficulty is in defining their expected utility from the original game. In an arbitrary game, the expected payoff for a player is not well defined, due to the uncertainty around the different players' strategies. Outside of simple cases – e.g., if there is a dominant strategy equilibrium – common notions of rationality do not tell us how agents should choose their strategy, in particular in the presence of multiple equilibria.

There are several ways to solve this problem. The first, which underpins our results, is simply to eliminate the strategic uncertainty. This requires some common-knowledge and deterministic decision procedure, so that all agents share an understanding of how they would play the game and what their expected payoffs are. It may seem unlikely for humans, but for software agents it is more reasonable, as their decision procedure would be a computer program, which they could conceivably share during the negotiation of a social contract [34].

Another solution would be that the agents be *averse* to strategic risk. Following [29], we can consider a game with strategic risk as a kind of lottery (albeit without well-defined probabilities), and with comparable value in the possible outcomes, risk-averse agents would prefer the social contract, where the outcome is almost certain (as it is a dominant-strategy equilibrium), to this lottery. An example is the model of rationality defined by Von Neumann and Morgenstern [37], where a rational player would remain in a coalition as long as they were guaranteed a payoff greater than the worst possible payoff that they could guarantee themselves, i.e. the "maximin" payoff that they could obtain assuming the other players' worst-case behaviour.

Conversely, it would be more difficult to create a social contract for *risk-seeking* agents would prefer the lottery to the social contract, unless the social contract provided them with a considerable amount of surplus. Similarly, if the agents' decision procedure was not deterministic, they could all overestimate their expected utility from the original game, and there might not be enough surplus to convince them to adopt the social contract.

Regarding the moral objective, the difficulty is that for other objectives than social welfare, such as distributive fairness (e.g. Nash social welfare or maximin fairness), the most desirable outcomes are not necessarily those that maximize social welfare, and implementing a social to ensure fairness does not guarantee a utility surplus that can be used to encourage self-interested agents to join. In the extreme case, the original game could have a clear outcome where social welfare is maximized, and then no social contract can be implemented where all agents are better off.

Finally, we note that our results have so far shown the existence (or not) of solutions, but do not offer a clear procedure to select one of the possible solutions. In practice, implementing a social contract would probably involve some negotiation over the surplus utility produced by the "better" outcome. A concrete mechanism for this could be an n-player bargaining game [6], where the disagreement payoff is given by $\pi(G)$.

REFERENCES

[1] Colin Allen, Gary Varner, and Jason Zinser. 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12, 3 (2000), 251–261.

[2] Robert J Aumann. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society* (1987), 1–18.

[3] Robert Axelrod. 1984. *The Evolution of Cooperation.* Basic Books, New York.

[4] James M Buchanan. 1975. *The limits of liberty: Between anarchy and Leviathan.* Number 714. University of Chicago Press.

[5] Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D Procaccia, Nisarg Shah, and Junxing Wang. 2016. The unreasonable fairness of maximum Nash welfare. In *International Conference on Economics and Computation.* ACM, 305–322.

[6] Suchan Chae and Jeong-Ae Yang. 1994. An N-person pure bargaining game. *Journal of Economic Theory* 62, 1 (1994), 86–102.

[7] George Christodoulou and Elias Koutsoupias. 2005. The price of anarchy of finite congestion games. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing.* ACM, 67–73.

[8] Richard Cole, Yevgeniy Dodis, and Tim Roughgarden. 2003. Pricing network edges for heterogeneous selfish users. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing.* ACM, 521–530.

[9] Richard Cole, Yevgeniy Dodis, and Tim Roughgarden. 2006. How much can taxes help selfish routing? *J. Comput. System Sci.* 72, 3 (2006), 444–467.

[10] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral Decision Making Frameworks for Artificial Intelligence.. In *AAAI.* 4831–4835.

[11] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14, 3 (2004), 349–379.

[12] Dimitris Fotakis, George Karakostas, and Stavros G Kolliopoulos. 2010. On the existence of optimal taxes for network congestion games with heterogeneous users. In *International Symposium on Algorithmic Game Theory.* Springer, 162–173.

[13] James W Friedman. 1986. *Game theory with applications to economics.* Vol. 87. Oxford University Press New York.

[14] Robert Gibbons. 2001. Trust in social structures: Hobbes and Coase meet repeated games. In *Russell Sage foundation series on trust, Vol. 2. Trust in society,* K. S. Cook (Ed.). Russel Sage Foundation, 332–353.

[15] Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian Charles Williams. 2016. Embedding Ethical Principles in Collective Decision Support Systems.. In *AAAI,* Vol. 16. 4147–4151.

[16] Douglas D Heckathorn and Steven M Maser. 1987. Bargaining and constitutional contracts. *American Journal of Political Science* (1987), 142–168.

[17] Chris Holmes Parker, Adrian K Agogino, and Kagan Tumer. 2016. Combining reward shaping and hierarchies for scaling to large multiagent systems. *The Knowledge Engineering Review* 31, 1 (2016), 3–18.

[18] Matthew O Jackson. 2014. Mechanism theory. *SSRN 2542983* (2014).

[19] Elias Koutsoupias and Christos Papadimitriou. 1999. Worst-case equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science.* Springer, 404–413.

[20] Derek Leben. 2017. A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology* 19, 2 (2017), 107–115.

[21] Joshua Letchford, Vincent Conitzer, and Kamal Jain. 2008. An "ethical" game-theoretic solution concept for two-player perfect-information games. In *International Workshop on Internet and Network Economics.* Springer, 696–707.

[22] David Mguni, Joel Jennings, Emilio Sison, Sergio Valcarcel Macua, Sofia Ceppi, and Enrique Munoz de Cote. 2019. Coordinating the Crowd: Inducing Desirable Equilibria in Non-Cooperative Systems. In *International Conference on Autonomous Agents and MultiAgent Systems.* 386–394.

[23] John Stuart Mill. 2016. Utilitarianism. In *Seven masterpieces of philosophy.* Routledge, 337–383.

[24] Dov Monderer and Moshe Tennenholtz. 2003. k-Implementation. In *Proceedings of the 4th ACM conference on Electronic commerce.* ACM, 19–28.

[25] Ritesh Noothigattu, Snehalkumar'Neil'S Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D Procaccia. 2018. A voting-based system for ethical decision making. In *AAAI.*

[26] Arthur C Pigou. 1920. *The economics of welfare.* Macmillan and co., London.

[27] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.

[28] John Rawls. 1971. *A theory of justice.* Harvard University Press.

[29] AE Roth. 1988. The expected value of playing a game. In *The Shapley Value: Essays in Honor of Lloyd S. Shapley.* Vol. 51. 70.

[30] Tim Roughgarden. 2007. Routing games. *Algorithmic game theory* 18 (2007), 459–484.

[31] William H Sandholm. 2007. Pigouvian pricing and stochastic evolutionary implementation. *Journal of Economic Theory* 132, 1 (2007), 367–382.

[32] Ramteen Sioshansi, Shmuel Oren, and Richard O'Neill. 2008. The cost of anarchy in self-commitment based electricity markets. *Competitive Electricity Markets: Design, Implementation and Performance* (2008), 245–266.

[33] MJ Smith. 1979. The marginal cost taxation of a transportation network. *Transportation Research Part B: Methodological* 13, 3 (1979), 237–242.

[34] Moshe Tennenholtz. 2004. Program equilibrium. *Games and Economic Behavior* 49, 2 (2004), 363–373.

[35] Kagan Tumer and Adrian Agogino. 2006. Agent reward shaping for alleviating traffic congestion. In *Workshop on Agents in Traffic and Transportation.* 87.

[36] Andrew Tutt. 2016. An FDA for algorithms. *Admin. Law Review* 69, 83 (2016).

[37] John Von Neumann and 0 Morgenstern. 1944. *Theory of games and economic behavior.* Princeton UP.