Value-Alignment Equilibrium
in Multiagent Systems

Nieves Montes[(✉)] and Carles Sierra

Artificial Intelligence Research Institute, IIIA-CSIC, Campus UAB, Bellaterra, Spain
{nmontes,sierra}@iiia.csic.es

Abstract. Value alignment has emerged in recent years as a basic principle to produce beneficial and mindful Artificial Intelligence systems. It mainly states that autonomous entities should behave in a way that is aligned with our human values. In this work, we summarize a previously developed model that considers values as preferences over states of the world and defines alignment between the governing norms and the values. We provide a use-case for this framework with the Iterated Prisoner's Dilemma model, which we use to exemplify the definitions we review. We take advantage of this use-case to introduce new concepts to be integrated with the established framework: alignment equilibrium and Pareto optimal alignment. These are inspired on the classical Nash equilibrium and Pareto optimality, but are designed to account for any value we wish to model in the system.

Keywords: Value alignment · Normative systems · Responsible AI

1 Introduction

In the last decades, research in Artificial Intelligence (AI) has been able to design and deploy increasingly complex systems, from robots and software agents, to recommendation algorithms and social networking apps [10]. Given that nowadays interaction with AI systems happens on a daily basis, a new challenge arises: how to ensure that all these systems, with all their complexity and power, behave in a way that is aligned with our human values. This requirement is referred to as the Value-Alignment Problem (VAP) [16], and is the focus of this work.

In this paper, we summarise a previously developed value alignment model [19] motivated by the assumption that we should be able to prove that any designed system is actually complying with our values [15]. It uses *norms* as the essential tools to supervise and limit autonomous agents' behaviour [11]. More importantly, it provides a precise definition of what it means for a norm to be aligned with a given value. We then provide a use-case based on the benchmark Iterated Prisoner's Dilemma game [3]. Inspired by the concepts of Nash equilibria and Pareto optimality in game theory, we introduce new concepts of value equilibria and optimality that build on top of the formal framework in [19].

2   Value Alignment Model

2.1   Revision of Our Background Formal Model

In order to introduce the necessary background, we provide a summary of the value alignment model that constitutes the starting point of this paper [19]. Its main underlying assumption is a *consequentialist view of values* [20], which expresses that the worthiness of any value is entirely determined by the outcomes of the actions that it motivates. Values, then, can serve as numerical quantifiers to assess how (un)desirable is a state in the world. In particular, we can use values to compare any two states and decide which of the two is preferred.

In the reviewed framework, the common conception of the world as a labelled transition system [8] is adopted. The world, then, is composed of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$ and a set of transitions $\mathcal{T} \subseteq \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. We refer to any transition $(s, a, s') \in \mathcal{T}$ with the notation $s \xrightarrow{a} s'$.

Values are conceived as mental constructs [12] that allow agents to decide which state of the world they prefer, according to their most prioritised values. This consideration motivates the following definition:

Definition 1 (from [19]). *A value-based preference* Prf *is a function over pairs of states that indicates how much preferred is one state over another in light of a particular value:* Prf $: \mathcal{S} \times \mathcal{S} \times G \times V \to [-1, 1]$, *where G is the set of agents and V is the set of values. The notation* $\mathsf{Prf}_v^\alpha(s, s')$ *indicates how much does agent α prefer state s' over state s with respect to value v.*

Value-based preferences, then, are bounded functions between -1 and $+1$. Positive (negative) preference indicates the post-transition state is more (less) desirable than the pre-transition state with respect to a specific value v. Preferences equal to 0 indicate that both states are identically preferred.

Values are held at the agent level. However, very rarely do agents' belief systems consist of a single value [18]. Aggregations over subsets of values and/or agents are also considered in [19]. However, we will not be employing any aggregation functions in this work. The interested reader is referred to the original paper for further details.

Now we have a clear view of the role of values when it comes to evaluating states. However, the states that can arise when letting a multiagent system evolve are dependent upon the *norms* in place, since these are the constructs which govern behaviour and therefore limit the actions that can be taken. Thus, value alignment is conceived as the alignment of a norm (or a set of norms) with respect to a value that is held in high regard. When a set of norms N is incorporated into the world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$, it is modified into a new, *normative* world $(\mathcal{S}, \mathcal{A}, N, \mathcal{T}_N)$ [1], where $\mathcal{T}_N \subseteq \mathcal{T}$ is the subset of all the original transitions allowed by N.

To evaluate how well or badly aligned is a norm $n \in N$, the transitions that can happen when this norm is enforced have to be evaluated. A norm is positively (negatively) aligned if it gives rise to transitions that move the system towards more (less) preferred states. However, beyond single transitions, the

long-term evolution of the world under the norms in place should be considered. This necessity motivates the following definition:

Definition 2 (from [19]). *A path p in the world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$ is a finite sequence of consecutive transitions $\{s_0 \xrightarrow{a_0} s_1, s_1 \xrightarrow{a_1} s_2, ..., s_i \xrightarrow{a_i} s_{i+1}, ..., s_f \xrightarrow{a_{f-1}} s_f\}$, where $p_F[i] = p_I[i+1]$, and $p_I[i]$ and $p_F[i]$ denote the pre- and post-transition states of the i-th transition.*

Paths are used to evaluate preferences over consecutive transitions, and support the formal definition of value alignment:

Definition 3 (from [19]). *The degree of alignment of norm $n \in N$ with respect to value $v \in V$ in the world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$ for agent $\alpha \in G$ is defined as the accumulated preference over all the paths in the normative world that results from implementing such norm:*

$$\mathsf{Algn}^\alpha_{n,v}(\mathcal{S}, \mathcal{A}, \mathcal{T}) = \frac{\sum\limits_{p \in paths} \sum\limits_{d \in [1, |p|]} \mathsf{Prf}^\alpha_v(p_I[d], p_F[d])}{\sum\limits_{p \in paths} |p|} \tag{1}$$

where paths *is the set of all paths in the normative world $(\mathcal{S}, \mathcal{A}, \{n\}, \mathcal{T}_n)$, and $|p|$ corresponds to the cardinality of p,* i.e. *the number of transitions in the path.*

Note that we exclude the possibility of infinite transition systems by considering all paths to be finite.

In this approach, the same exact weight is given to every single transition. Other suggestions are conceivable; for example one may want to consider whether preferences remain approximately stable along the paths or, conversely, there are large surges or sinks. In other fields, it is typical to consider a discount parameter that reduces the weight of transitions that happen into the distant future [17]. We acknowledge the existence of alternative approaches, but leave their exploration for future work.

Another issue to note with Definition 3 is that its notation indicates the alignment for a single agent α with respect to a single value v, since it takes the individual preferences with respect to that one value. However, as previously mentioned, preferences can be aggregated over agents and/or sets of values. Using such aggregated preferences would, consequently, result in alignment for sets of agents and/or with respect to sets of values.

Equation (1) states that alignment should take into account all possible transitions in the normative world. In general, this approach is not efficient, and to solve this issue, Monte Carlo sampling over all possible paths is recommended. Additionally, it is also advisable to keep the length of the paths fixed. These modifications lead to the following reformulation for the alignment:

$$\mathsf{Algn}^\alpha_{n,v}(\mathcal{S}, \mathcal{A}, \mathcal{T}) = \frac{\sum\limits_{p \in \text{paths}} \sum\limits_{d \in [1,l]} \mathsf{Prf}^\alpha_v(p_I[d], p_F[d])}{x \times l} \tag{2}$$

where l is the number of transitions in all paths, and x is the number of sampled paths.

In summary, the approach proposed in [19] provides a formal model to numerically quantify how compliant is a certain normative world designed towards some particular value. Differently to other works related to value alignment [2,7], such a model is separate from the decision-making process of the participating agents and their respective goals, and can hence be applied to any kind of agent and social space architecture.

2.2   Value-Alignment Solution Concepts

In this work, we model agents' interactions as normal-form games. In game theory, a stage game refers to each of the identical rounds played in one iteration. The Nash equilibrium is then defined as the set of players' actions such that no player can obtain a higher profit by unilaterally deviating from it, given the actions of all other players are fixed [14, Chapter 2]. In formal terms, Nash equilibria correspond to action profiles $(a_1^*, ..., a_{|G|}^*)$ such that, for all agents $i \in G$, it holds that:

$$r_i(a_1^*, ..., a_{i-1}^*, a_i^*, a_{i+1}^*, ..., a_{|G|}^*) \geq r_i(a_1^*, ..., a_{i-1}^*, a_i, a_{i+1}^*, ..., a_{|G|}^*) \qquad (3)$$

Nash equilibria may not represent the best option for players in terms of individual revenues, but profiles of joint actions for which no player has an incentive to unilaterally deviate. This solution concept represents *status quo* positions that persist despite not necessarily being the best solutions from a social perspective. In other terms, the Nash equilibria do not necessarily correspond with Pareto optimal outcomes [5], as in the case of the classical Prisoner's Dilemma game.

The concept of stage game Nash equilibrium is not directly applicable in our theoretical framework for two main reasons. First, in general, agents interact repeatedly. And second, we are interested in the alignment with respect to values, not in the game rewards themselves, even if alignments may be computed from rewards. However, Nash-like equilibria situations can be identified when the alignment satisfies an adapted version of Eq. (3). When considering this possibility, the argument a_i^* should not be identified as actions taken by agents in a single round, but rather by the strategies that individuals follow for the whole duration of the game. We conceive individual strategies as separate from norms. While strategies are part of the agents' internal decision process, we understand norms to be externally imposed constraints on the system, whose definition is the responsibility of the model's designer and beyond the control of the participating agents.

Therefore, given a set of norms N governing a multiagent system, agents adopt a particular strategy to play the game. Strategies are functions that take into account the past history of the game to return an action to be performed next [14]. The set of all strategies being played is $E = \{e_1, e_2, ..., e_{|G|}\}$, where e_i is the strategy followed by the i-th agent. Then, we can make the following definition:

Definition 4. *Given a normative system N and an alignment function with respect to the value of choice for agent i Algn_v^i, the alignment equilibrium is defined as the tuple of all individual strategies $\left(e_1^*, ..., e_{|G|}^*\right)$ such that, for all agents $i \in G$, it holds that:*

$$\mathsf{Algn}^i_{\left(e_1^*, ..., e_{i-1}^*, e_i^*, e_{i+1}^*, ..., e_{|G|}^*\right), v} \geq \mathsf{Algn}^i_{\left(e_1^*, ..., e_{i-1}^*, e_i, e_{i+1}^*, ..., e_{|G|}^*\right), v} \tag{4}$$

Note that the alignment equilibria depend on the alignment function of choice and the norms constraining the system. Another point to note is that, despite Eq. (4) is stated in terms of alignment with respect to individual agents, a natural extension arises when applied to subsets of agents or even aggregated for the whole society. Additionally, it can also be extended to sets of values.

Another important point is that, unlike the classical Nash equilibrium, whose existence is guaranteed for games played by a finite number of agents following mixed strategies [9], the properties of Eq. (4) have not been explored to the point of establishing conditions for existence. This is left as future work.

It is worth trying to understand the motivation behind Eq. (4). Classical game theory takes the view of trying to maximise one's own reward while minimising risks. The classical Nash equilibrium for a stage game is a possible response to this approach for a game that is played once. Our definition of alignment equilibrium, then, generalises the classical Nash equilibrium to account for games and situations that are presented repeatedly.

More importantly, Eq. (4) allows to examine a game from a different perspective other than that of the individual payoffs r_i, as we can analyse the alignment equilibria with respect to as many values as alignment functions we are able to come up with. Moreover, agents may consider different preferences or may take into account different variables to compute their preferences, and thus Eq. (4) need not be symmetric with respect to agents. By considering the particular instance where a game is played for a path of length one (a single round) and using the actual rewards as preferences, the classical concept of Nash equilibrium can be recovered.

Previously, we have mentioned the concept of Pareto optimality [5]. This solution concept from game theory can also be adapted to the value alignment framework. Classically, an action profile $(a_1^*, ..., a_{|G|}^*)$ is said to lead to a *Pareto optimal outcome* if there is no other profile $(a_1, ..., a_{|G|})$ such that:

$$r_j\left(a_1, ..., a_{|G|}\right) > r_j\left(a_1^*, ..., a_{|G|}^*\right) \text{ for at least one agent } j \in G, \text{ and} \tag{5}$$

$$r_i\left(a_1, ..., a_{|G|}\right) \geq r_i\left(a_1^*, ..., a_{|G|}^*\right) \text{ for all agents } i \in G \tag{6}$$

In other words, in classical game theory a Pareto optimal outcome is one where we cannot improve anyone's reward without damaging someone else's.

In a similar fashion to our extension from the classical Nash equilibrium to the alignment equilibrium, we can adapt Pareto optimality to our value alignment framework. Again, we need not consider actions for a single transition, but rather strategies adopted during the whole duration of the game:

Definition 5. *Given a normative system N and an alignment function with respect to the value of choice for agent i Algn_v^i, a tuple of individual strategies $\left(e_1^*, ..., e_{|G|}^*\right)$ is said to lead to a* Pareto optimal alignment *if there is no other tuple of individual strategies $\left(e_1, ..., e_{|G|}\right)$ such that:*

$$\mathsf{Algn}_{\left(e_1,...,e_{|G|}\right),v}^j > \mathsf{Algn}_{\left(e_1^*,...,e_{|G|}^*\right),v}^j \quad \text{for at least one agent } j \in G, \text{ and} \quad (7)$$

$$\mathsf{Algn}_{\left(e_1,...,e_{|G|}\right),v}^i \geq \mathsf{Algn}_{\left(e_1^*,...,e_{|G|}^*\right),v}^i \quad \text{for all agents } i \in G \quad (8)$$

Therefore, a Pareto optimal alignment corresponds to a situation where no agent can improve its alignment without hurting one or several other agent alignments. The novel concept of Pareto optimal alignment allows us to make the same generalisations than those when moving from the classical Nash equilibrium to the alignment equilibrium. They have been explicitly described previously.

3   Use-Case: Two-Agent Iterated Prisoner's Dilemma

Now, we consider a very simple two-agent system modelled after the benchmark Iterated Prisoner's Dilemma (IPD) game [3], and we use it to illustrate the previously introduced concepts.

**Table 1.** Outcome matrix for the Prisoner's Dilemma game.

		β	
		Cooperate	Defect
α	Cooperate	$(6, 6)$	$(0, 9)$
	Defect	$(9, 0)$	$(3, 3)$

In formal terms, this model consists of an agent set with two members, $G = \{\alpha, \beta\}$; the set of individual actions available to each of them is $\mathcal{A}_i = \{\text{cooperate}, \text{defect}\}$. The transitions between states are characterised by the joint actions that α and β take, hence the set of actions is $\mathcal{A} = \mathcal{A}_i^2 = \{(a_\alpha, a_\beta)\}$. The tuple of individual actions (a_α, a_β) determines the outcome of the transition. Such outcomes are tied to rewards agents receive, $r_i(a_\alpha, a_\beta), \forall i \in G$, via the outcome matrix displayed in Table 1.

In this model, the game is played in an iterated way. We characterise states by the agents' wealth, which is defined as the accumulated rewards received since the game started. The states of the world correspond to the tuple of agents' wealth, noted by (x_α, x_β). Initially, both agents start with $(x_\alpha, x_\beta) = (0, 0)$. After agents have chosen their actions (a_α, a_β) and the rewards (r_α, r_β) are determined, the transition $s \to s'$ is completed by updating the values for agents' wealth.

We formally lay out the main features of our model in the following terms:

Definition 6. *We define the* Two-Agent Iterative Prisoner's Dilemma (2A-IPD) *model as a tuple $(G, \mathcal{S}, \mathcal{A}, \mathcal{T})$, where:*
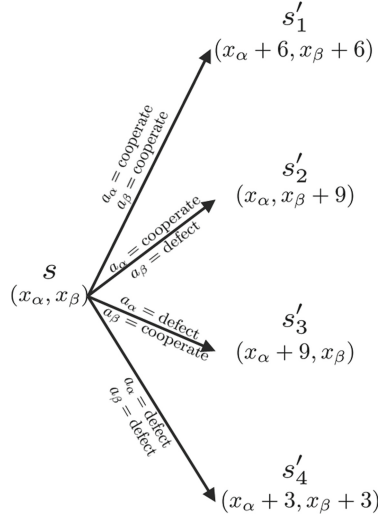
Fig. 1. Labelled transition system representing a single transition of the 2A-IPD model.

- $G = \{\alpha, \beta\}$ *is a set of two agents.*
- $S = \{(x_\alpha, x_\beta)\}$ *is the set of states, composed by the tuples of all possible combinations of agents' wealth.*
- $A = \{(a_\alpha, a_\beta)\}$, *where* $a_i \in \{cooperate, defect\}$, *is the set of joint actions.*
- $\mathcal{T} : S \times A \times S$ *is the set of transitions which relate the current state* $s = (x_\alpha, x_\beta)$ *and joint actions* (a_α, a_β) *to the next state* $s' = (x'_\alpha, x'_\beta)$:

$$\forall t \in \mathcal{T}, \quad t = (s, (a_\alpha, a_\beta), s') \quad such \ that \quad \begin{cases} x'_\alpha = & x_\alpha + r_\alpha(a_\alpha, a_\beta) \\ x'_\beta = & x_\beta + r_\beta(a_\alpha, a_\beta) \end{cases} \quad (9)$$

where $r_i(a_\alpha, a_\beta)$ *are the rewards given by the outcome matrix (Table 1).*

A labelled transition system representation for one transition of this model is displayed in Fig. 1.

3.1 Preference Functions

Within this model, we wish to quantify the alignment of different agent behaviours with respect to two apparently opposed values: *equality* and *personal gain*. In [19] some preferences related to these values were already formulated. However, we introduce new ones here that are less sensitive to the numerical choice for the rewards in Table 1.

In order to quantify alignment with respect to equality, we will assess it in each state by making use of the well-known Gini Index (GI) [6]. In our two-agent model, the GI is computed for state s by taking the values of agents' wealth at

that particular point in the repeated game, (x_α, x_β). For $|G| = 2$, the Gini Index for our system becomes:

$$\text{GI}(s) = \frac{|x_\alpha - x_\beta|}{2(x_\alpha + x_\beta)} \tag{10}$$

The lower bound for the GI is always 0, indicating perfect equality among all participants. In the case of two agents, the maximum possible value is $\frac{1}{2}$ [4]. Then, in order to map perfect equality (GI = 0) to maximum preference (+1) and perfect inequality (GI = $\frac{1}{2}$) to minimum preference (−1), the interval for the Gini Index $[0, \frac{1}{2}]$ is linearly mapped to the interval of definition of preferences $[−1, 1]$. This transformation results in the following definition for the preference function over the value *equality*:

$$\text{Prf}^i_{\text{equality}}(s, s') = 1 - 4 \cdot \text{GI}(s') = 1 - 2 \cdot \frac{|x'_\alpha - x'_\beta|}{x'_\alpha + x'_\beta} \tag{11}$$

where $i = \alpha, \beta$.

There are two important points to be noted about this preference function. First, it is numerically equivalent for both agents, since x'_α and x'_β are interchangeable. It is an intuitive property of the preference with respect to the value *equality* that it should be indeed identical for both agents. Second, $\text{Prf}^i_{\text{equality}}(s, s')$ is a function of only the properties of the system in the post-transition state s'. Hence, states with high (low) equality are (not) preferred regardless of the parity in the previous state s.

The latter property is not a requirement of preference functions. Other formulae could be devised that depend on the increase/decrease of the Gini Index or some other indicator. Since our model starts off from a very peculiar position of perfect equality, we have considered that it is more helpful to monitor the eventual disparity that may arise as the game proceeds, rather than comparing consecutive states.

The other value under consideration, personal gain, is quantified through the following preference function:

$$\text{Prf}^i_{\text{gain}}(s, s') = \begin{cases} -1 & \text{if } x'_i - x_i = 0 \\ -\frac{1}{3} & \text{if } x'_i - x_i = 3 \\ \frac{1}{3} & \text{if } x'_i - x_i = 6 \\ 1 & \text{if } x'_i - x_i = 9 \end{cases} \tag{12}$$

According to this definition, states are preferred with respect to personal gain by ranking the possible rewards in any given transition and mapping them to equally spaced points in the preference interval $[−1, 1]$. This choice is made to reflect the greediness of this value, since the preference is only dependent on immediate gains, regardless of how well off the agent may already be or the circumstances of her peer.

Now, we have the preference functions to evaluate the alignment of the system for the two agents. In order to compute the alignment from preferences, Eq. (2) is employed, with $x = 10,000$ sampled paths of length $l = 10$ (the number of game iterations).

3.2   Individual Strategies

Now that we have encoded values into two different preference functions, our purpose, then, is to find which agent strategies in the 2A-IPD model result in alignment equilibrium and Pareto optimal alignment positions with respect to equality and personal gain. The set of all possible strategies for a single agent is a vast space of logical formulae, dictating whether past actions should be taken into account to decide the future action, how far back to look and many other considerations. In order to work with the alignment equilibrium without the burden of considering *all* possible strategies, we restrict ourselves to two subspaces of possible strategy profiles, from which we will determine the alignment equilibria and Pareto optimal alignments:

1. *Random-action profiles:* Both α and β choose at each round the action to take randomly, according to independent and fixed probabilities of cooperation. These are analogous to mixed strategies in game theory, where actions are taken based on a fixed probability.
2. *Heterogeneous profiles:* β cooperates randomly following a fixed probability. α, in contrast, either cooperates or defects with probability 0.5 in the first round. In subsequent rounds, it follows one of these strategies:
   - *Tit-for-tat:* α's action in the current round is β's action in the previous round.
   - *Mostly cooperate:* α defects if in the previous round both agents defected. Otherwise, it cooperates.
   - *Mostly defect:* α cooperates if in the previous round both players cooperated. Otherwise, it defects.

   These are some of the most common strategies in non-cooperative game theory. We set that only one agent follows them in order not to condition the outcome of all transitions on the result of the first one.

Note that all of the strategies presented here are included in the set of reactive strategies [13], where behaviour is only dependent on the opponent's immediate past action.

4   Results

First, the results for the alignment under random-action profiles with respect to equality and personal gain are presented in Figs. 2 and 3 respectively. They are plotted as a function of the cooperation probabilities of both agents, which specify concrete instances of the individual random strategies e_α and e_β.

According to the definition of alignment equilibrium in Eq. (4), in Fig. 2 there is an infinite number of alignment equilibria corresponding to strategies satisfying $e_\alpha = e_\beta$ (here we only plot 10×10 strategy pairs). To see this, if we consider that the cooperation probability of agent β is e_β, then α's alignment is maximal when e_α is equal to e_β, and α has no incentive to deviate. Similarly if
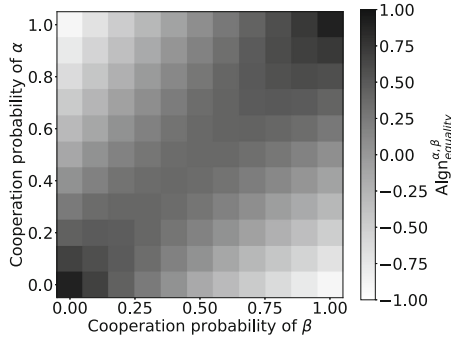
Fig. 2. Alignment of agents α and β with respect to value *equality*, Eq. (11), under random-action profiles.

we fix the cooperation probability of agent α, given the symmetry of Eq. (11). Therefore, an infinite number of alignment equilibria are found at $e_\alpha^* = e_\beta^*$.

It is unsurprising that the *status quo* with respect to equality corresponds to both agents having the same cooperation probability. It is worth noting that this result does not encourage any agent to increase its cooperation probability to enhance equality, but rather to act similarly to her peer.

The Pareto optimal alignment with respect to equality is found when either both agents always cooperate or always defect, since they always receive identical rewards and the Gini Index is kept to zero. So, actually, alignment equilibrium strategies with respect to equality do include Pareto optimal strategies.

As for alignment with respect to gain in Fig. 3, the maximum alignment for any agent is obtained when she defects and the other player follows a complete cooperation strategy. The alignment equilibrium, again considering only random-action profiles, is found at the position where both agents never cooperate. This is a consequence of the single Nash equilibrium of the stage game at the position (defect, defect), in combination with the definition of preference with respect to personal gain in Eq. (12) being directly related to the individual gains for a single round.

Differently from the results obtained for value *equality* where equilibrium positions were Pareto optimal, for value *personal gain* this is not the case. Starting from constant mutual defection, both agents could improve their alignment by turning to constant mutual cooperation instead, since they would attain larger gains at each round. This result is also due to the direct relationship between preferences with respect to personal gain and the actual rewards in each round, see Eq. (12).

This pair of strategies (both agents always defecting for all repetitions of the game) is the alignment equilibrium when both agents' alignments are computed with respect to value *personal gain*. An interesting observation is that this is also the alignment equilibrium when one player's alignment is computed with respect to equality, and the other with respect to gain. Let's consider the case where
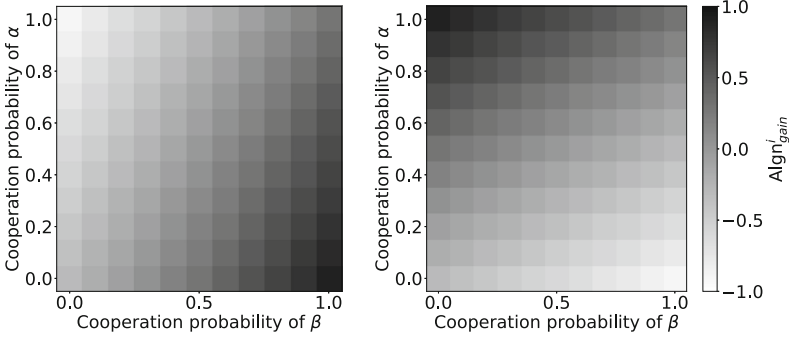
Fig. 3. Alignment of agent α (left) and β (right) with respect to value *personal gain*, Eq. (12), under random-action profiles. Note that $\mathsf{Algn}^\alpha_{\text{gain}}$ and $\mathsf{Algn}^\beta_{\text{gain}}$ are mutually transposed.

α's alignment is computed with respect to gain, Fig. 3 left, and β's alignment is computed with respect to equality, Fig. 2.

Regardless of β's actions, α's alignment is always maximised under null cooperation probability, so the preferred strategy to take is to defect in every round. Then, β maximises her alignment with respect to equality by imitating α, that is, always defecting as well. The result is that the only alignment equilibrium satisfying Eq. (4) for random-actions profiles is found at $e^*_\alpha = e^*_\beta = 0$. It is worth noting that the calculation for the alignment is specific to each agent in order to reflect their different priority values. Again, this alignment is not Pareto optimal. If both agents switched to $e_\alpha = e_\beta = 1$, α could enhance her alignment with respect to personal gain while maintaining β's alignment with respect to equality at 1.

A summary of the results presented so far for random-action profiles is displayed in Table 2.
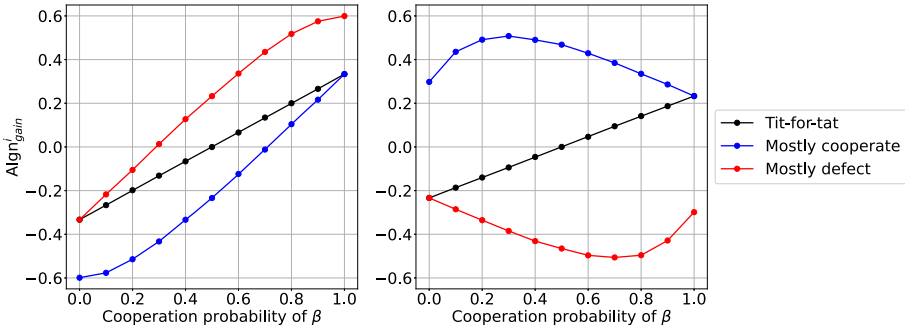
Now, we look into the alignment with respect to a single value under heterogeneous profiles. The alignment for both players is plotted as a function of the cooperation probability of β, under the various possible strategies for α, for value *equality* in Fig. 4 and for value *personal gain* in Fig. 5.

Table 2. Summary of the results obtained for two randomly-behaving players, highlighting the position of alignment equilibrium strategies and their Pareto optimality.

Value for α	Value for β	Alignment equilibrium	Pareto optimal
Equality	Equality	$e^*_\alpha = e^*_\beta$	Included
Personal gain	Personal gain	$e^*_\alpha = e^*_\beta = 0$	No
Personal gain	Equality	$e^*_\alpha = e^*_\beta = 0$	No
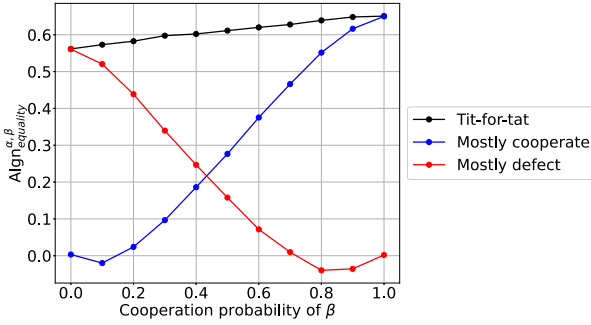
Fig. 4. Alignment of agents α and β with respect to value *equality*, Eq. (11), under heterogeneous profiles, depending on the strategy followed by player α.

Fig. 5. Alignment of agent α (left) and β (right) with respect to value *personal gain*, Eq. (12), under heterogeneous profiles, according to the strategy followed by player α.

First, we focus our attention on the alignment with respect to equality in Fig. 4. In this case, the three strategies result in very distinct trends. *Tit-for-tat*, equivalent to imitating the opponent, leads to a very stable alignment with respect to equality, independently of the extend of collaboration of β. The other two strategies, *mostly cooperate* and *mostly defect*, result in alignments that are strongly dependent on β's cooperation probability. When β defects often (cooperation $\sim$0), *mostly defect* is the preferred individual strategy for α. In contrast, when β cooperates often (cooperation $\sim$1), *mostly cooperate* is the most suitable strategy for α. These results are in line with those obtained for random-action profiles with respect to equality (Fig. 2), where we observed that the alignment equilibrium was reached when both payers behaved similarly.

In order to find the alignment equilibrium from results presented as in Fig. 4, we must look for the position(s) such that: (a) For constant cooperation probability of β, *i.e.* by fixing the position along the x axis, changing α's strategy by switching line colour leads to a decrease in alignment; and (b) for fixed α's strategy, *i.e.* maintaining the line colour, the cooperation probability of β (the coordinate along the x axis) corresponds to a maximum along that line.

For heterogeneous profiles, the alignment equilibrium with respect to equality is found when β's strategy along the iterated game settles in cooperation (probability of cooperation equals 1) and α applies the *tit-for-tat* or *mostly cooperate* strategies, both of which result in α cooperating for all rounds of the game, except possibly in the first one. Neither β nor α would then have any incentive to unilaterally deviate.

Differently to the case of random strategies, where there was an infinite number of alignment equilibria (the positive diagonal), in this case there is a single alignment equilibrium that corresponds to the persistent collaboration of both agents along the iterated game, which is something clearly desirable from a social perspective. Then, the introduction of strategic directives for α has shifted the equilibrium with respect to equality towards cooperation. It is also worth noting that, in this case, the equilibrium strategy corresponds exactly with the only Pareto optimal position.

Second, we concentrate on alignment with respect to personal gain, Fig. 5. In this case, α's alignment is strongly dependent on the cooperation probability of β. The three strategies yield alignments for α that are monotonically increasing with β's collaboration. Also, the three strategies are always equally ranked. This means that, given a fixed cooperation ratio for β, the most aligned strategy is always *mostly defect*, followed by *tit-for-tat* and finally *mostly cooperate*. The strategies are ordered from least to most cooperative.

As for β's alignment, it increases linearly with its cooperation probability when α follows *tit-for-tat*. It displays a peak at low collaboration rates when α deploys *mostly cooperates*, and a valley at high cooperation probabilities when α follows *mostly defect*.

Again, considering only the heterogeneous profiles that have generated these results, there are two alignment equilibria with respect to personal gain corresponding to β not cooperating at all and α following either *mostly defect* or *tit-for-tat*. To achieve this conclusion, we first note that, for any cooperation probability of β, α always enhances her alignment by following *mostly defect*. Then, once α has settled for this strategy, the best choice for β is to always defect. These two observations lead to α following either *mostly defect* or *tit-for-tat* and β never cooperating. Given that β always defects, both these strategies converge to α always defecting as well, except maybe at the first round. It is worth noting that these equilibria are actually far from the maximum possible alignment for either agent. Nor do they result in a Pareto optimal alignment, since both agents could improve their alignment by having α follow *tit-for-tat* and β increase her probability of cooperation.

Finally, in an exercise similar to that performed for random-action profiles, we find the alignment equilibrium position when agents prioritise different values. Since players in this strategy profiles are not equivalent (α behaves in a conscious way while β behaves completely randomly), we must examine two possibilities: α prioritises *personal gain* while β prioritises *equality*, and vice-versa.

Table 3. Summary of the results obtained under heterogeneous profiles, highlighting the position of alignment equilibrium strategies and their Pareto optimality (Nomenclature for strategies: TfT: *tit-for-tat*; MC: *mostly cooperate*; MD: *mostly defect*).

Value for α	Value for β	Alignment equilibrium	Pareto optimal
Equality	Equality	$e_\alpha^* = \text{TfT/MC}$; $e_\beta^* = 1$	Yes
Personal gain	Personal gain	$e_\alpha^* = \text{TfT/MD}$; $e_\beta^* = 0$	No
Personal gain	Equality	$e_\alpha^* = \text{TfT/MD}$; $e_\beta^* = 0$	No
Equality	Personal gain	$e_\alpha^* = \text{TfT/MC}$; $e_\beta^* = 1$	Yes

For the first possibility, α will always need to follow *mostly defect* to ensure that her alignment with respect to personal gain is maximised, regardless of β's probability of cooperation. Then, in order to attain maximum alignment with respect to equality, β will settle on constant defection. Hence, the alignment equilibrium when α prioritises personal gain and β prioritises equality is the same as when both prioritised personal gain. Yet again, these strategies do not lead to a Pareto optimal alignment, since both agents could improve the alignment with respect to their prioritised values by having α follow *tit-for-tat* and β increase her probability of cooperation.

For the second possibility, α will follow *tit-for-tat*, since this strategy dominates the two others when it comes to equality, regardless of the cooperation probability of β. Given this observation, β will then enhance her alignment with respect to *personal gain* by always cooperating. At this position, α can resort to *tit-for-tat* or *mostly cooperate* indistinctly. This strategy profile is equal to the equilibrium found when both agents prioritised equality. In this case, it also corresponds to a Pareto optimal alignment, since α has achieved the maximum possible alignment with respect to equality.

It is worth pointing out that the alignment equilibrium positions under agents prioritising different values are driven by the player following the more conscious strategy, α in this case. That is to say that when α focuses on personal gain, the solution concepts are identical regardless of the value that β (the randomly behaving agent) holds in high regard. The same result is found when α focuses on equality instead.

A summary with the results analysed for the model under heterogeneous profiles is provided in Table 3.

5   Conclusions and Future Work

In this work, we have reviewed a formal framework that establishes preferences over the states in the world, and we have specified the computation of value alignment through the increase or decrease in preferences over states in a normative world. Some further study needs to be done in the theoretical front, but for the time being we have been able to implement this framework in the Iterated Prisoner's Dilemma model.

Inspired by the classical Nash equilibrium and Pareto optimality in game theory, we have introduced the novel notions of *alignment equilibrium* and *Pareto optimal alignment*. These solution concepts extend the existing definitions to account for different values beyond individual rewards and generalise to cases where agents may have different priorities. We have been able to identify both equilibria and Pareto optimal alignments in our Two-Agent Iterated Prisoner's Dilemma model, with respect to the values *equality* and *personal gain*. An interesting finding is that, under both strategy subsets under consideration, alignment equilibria positions with respect to equality include Pareto optimal outcomes, while equilibrium positions with respect to value gain do not.

This work intends to exemplify an application of the proposed model for the value alignment problem. Further work, built on it, remains to be done. First, on the analytical side, formal properties of the alignment equilibrium and its relationship with Pareto optimal alignments should be explored. Second, the model can be naturally extended to account for the introduction of norms, such as taxes, fines or the banning/enforcement of behaviour. An interesting outcome from such research should be the shift, if any, in the alignment equilibria positions. Finally, a third line of work should be focused on the development of methodologies to synthesise norms with optimal alignment with respect to values of choice.

References

1. Andrighetto, G., Governatori, G., Noriega, P.: Normative Multi-agent Systems, vol. 4. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, April 2013. https://doi.org/10.4230/DFU.Vol4.12111.i

2. Atkinson, K., Bench-Capon, T.: States, goals and values: revisiting practical reasoning. Argument Comput. **7**(2–3), 135–154 (2016). https://doi.org/10.3233/AAC-160011

3. Axelrod, R.M.: The Evolution of Cooperation. Basic Books, New York (1984)

4. Bellù, L., Liberati, P.: Inequality Analysis: The Gini Index. Food and Agriculture Organization of the United Nations (2006). http://www.fao.org/policy-support/resources/resources-details/en/c/446282/

5. Chinchuluun, A., Pardalos, P., Migdalas, A., Pitsoulis, L.: Pareto Optimality Game Theory And Equilibria, vol. 17. Springer, New York (2008). https://doi.org/10.1007/978-0-387-77247-9

6. Cowell, F.A.: Measuring Inequality. LSE Perspectives in Economic Analysis. Oxford University Press, Oxford (2009)

7. Cranefield, S., Winikoff, M., Dignum, V., Dignum, F.: No pizza for you: value-based plan selection in BDI agents. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 178–184. International Joint Conferences on Artificial Intelligence Organization (2017). https://doi.org/10.24963/ijcai.2017/26

8. Gorrieri, R.: Labeled transition systems. Process Algebras for Petri Nets. MTC-SAES, pp. 15–34. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55559-1_2

9. Huang, L.: Nash theorem (in game theory). In: Encyclopedia of Mathematics. Springer (2002). http://www.encyclopediaofmath.org/index.php?title=Nash_theorem_(in_game_theory)&oldid=40004

10. Lu, Y.: Artificial intelligence: a survey on evolution, models, applications and future trends. J. Manage. Anal. **6**(1), 1–29 (2019). https://doi.org/10.1080/23270012.2019.1570365

11. Mahmoud, M.A., Ahmad, M.S., Mohd Yusoff, M.Z., Mustapha, A.: A review of norms and normative multiagent systems. Sci. World J. **2014**, 1–23 (2014). https://doi.org/10.1155/2014/684587

12. Miceli, M., Castelfranchi, C.: A cognitive approach to values. J. Theor. Soc. Behav. **19**(2), 169–193 (1989). https://doi.org/10.1111/j.1468-5914.1989.tb00143.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-5914.1989.tb00143.x

13. Nowak, M.A., Sigmund, K.: Tit for tat in heterogeneous populations. Nature **355**(6357), 250–253 (1992). https://doi.org/10.1038/355250a0

14. Osborne, M.J., Rubinstein, A.: A Course in Game Theory. MIT Press, Cambridge (2012)

15. Russell, S.: Provably beneficial artificial intelligence. In: The Next Step: Exponential Life. BBVA-Open Mind (2017)

16. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Penguin LCC, New York (2019)

17. Schwartz, H.M.: Multi-Agent Machine Learning: A Reinforcement Approach. Wiley, New York (2014). https://doi.org/10.1002/9781118884614

18. Schwartz, S.H.: An overview of the Schwartz theory of basic values. Online Read. Psychol. Cult. **2**(1) (2012). https://doi.org/10.9707/2307-0919.1116

19. Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., Perello-Moragues, A.: Value alignment: a formal approach. In: Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS 2019 (2019)

20. Sullivan, S., Philip, P.: Ethical theories (2002). https://www.qcc.cuny.edu/SocialSciences/ppecorino/ETHICS_TEXT/CONTENTS.htm