Value alignment: a formal approach

Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perello-Moragues

Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Catalonia

Abstract. Value alignment in AI has emerged as one of the basic principles that should govern autonomous AI systems. It essentially states that a system's goals and behaviour should be aligned with human values. But how to ensure value alignment? In this paper we first provide a formal model to represent values through preferences and ways to compute value aggregations; i.e. preferences with respect to a group of agents and/or preferences with respect to sets of values. Value alignment is then defined, and computed, for a given norm with respect to a given value through the increase/decrease that it results in the preferences of future states of the world. We focus on norms as it is norms that govern behaviour, and as such, the alignment of a given system with a given value will be dictated by the norms the system follows.

Keywords: Responsible AI · Value-alignment problem · Norms

1 Introduction

The aim of this paper is to explore the alignment of a given system with a given value. To achieve this, we explore the relationships between values, actions and norms in order to propose a precise way of expressing when a norm fosters behaviour in accordance to a value. This is a first step towards a formal foundation for a theory of value-driven behaviour of autonomous entities, that is good enough to engineer value-imbued socio-cognitive technical systems.

Our proposal is based on four main assumptions: First, we adopt a *cognitive view of values*. That is, we understand values as a cognitive construct that is involved in the rational behaviour of an agent [11, 18, 13, 20]. In line with Schwartz theory of motivational values [20] we will assume that values serve as standards, refer to desirable goals and transcend specific actions.

Our second assumption is a *consequentalist* view of values [22] by which the "worthiness" of a value, and therefore, its social meaning, is given by the outcomes of the actions that are aligned with it.

These two views allow us to postulate that values serve two decision-making purposes: (i) to assess the "worthiness" of a state of the world (and thus compare the "worthiness" of two states of the world), and (ii) to decide which is the "better" of two actions.

Next, we assume the usual *teleological view of norms* that conceives norms as a social means to promote desired behaviours and to discourage undesired ones.

We follow the normative notions that are prevalent in the field of multiagent systems (f.e. [2]). More specifically, in this paper we will assume that norms modify the outcomes of actions, by inducing agent behaviour through the use of positive and negative incentives. The fourth assumption is to characterise the space where actions take place to be a situated, online, regulated open multiagent system, in fact a socio-cognitive technical system (SCTS) [12].

Based on these assumptions, we make precise what it means for a norm to be aligned with a value. This is the basis for determining whether a system fosters behaviour that is in accordance with a value or not.

To achieve that characterisation of value-alignment we first establish a relationship between values and preferences and apply this relationship to the evolving states of the world. After that, we discuss how norms are linked with actions and how to determine the value-related effects of performing a norm compliant action. With these elements we then define the key notion of norm-alignment. We illustrates these ideas by instantiating and programming them for a version of the prisoner's dilemma. The last section suggests some open problems.

2 Background

2.1 The Value-Alignment Problem (VAP)

The VAP is motivated by the recognition that autonomous entities (robots, software agents) and intelligent systems in general exhibit increasingly complex behaviour and thus become difficult to regulate. One way to address this concern is to imbue values in intelligent systems. There are currently three main approaches to towards this aim. First the design of guidelines, standards and certifications [17]. Second, "value-based design" where values are brought into the systems from the very start as design requirements [15, 14, 4]. The third approach postulates that autonomous entities should be provably made to comply with values [19]. Our proposal is framed in this analytic approach. We follow a version of this approach framing the compliance problem within hybrid on-line social coordination systems (or *socio-cognitive technical systems*) whereas autonomous rational entities interact within a norm-regulated shared social space [1, 12]. More specifically, we frame the problem of imbuing values in such systems by having norms that are aligned with some values and autonomous agents that, subject to the norms of a social coordination environment, may act in accordance to their own —possibly different— values.

2.2 Values

We want a notion of value that is linked to agent value-guided behaviour that presumes an empirically-grounded, motivational understanding of values [11, 13, 16, 10, 20]. Drawing on Schwartz' theory of basic human values [20], we presume that each society adopts a finite set of basic human values, orders them by importance, and designs norms that foster behaviour that aligns with those values.

Likewise, individuals also adopt and order a finite set of basic values that align with their own behavioural profile, and are influenced by social values and the corresponding norms. Individuals' goal setting depends on the context (which includes applicable norms) and the individuals' mind-frames (which include the *individuals' values*, needs, personality, emotions and beliefs among other constructs). A full description of our proposal is beyond the scope of this paper but Figure 1 sketches our understanding of value-guided behaviour.

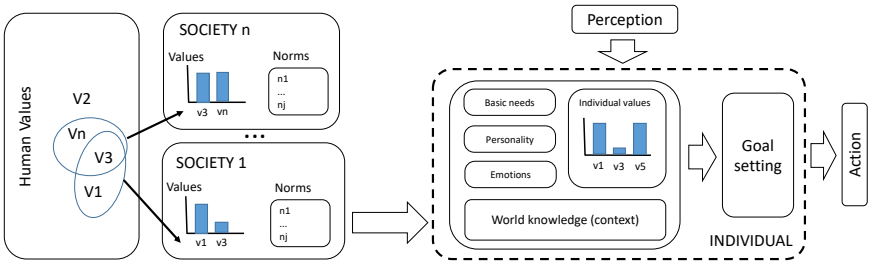


Fig. 1: General view of an agent's value-guided behaviour

With this view we make values serve two main functions: to assess the "goodness" of the states of the world —thus determining preferences between two states— and to decide whether one action is preferable to another. In order to make these intuitions operational it is convenient to take a *consequentialist* view of values by which value is reified by its consequences [22, 15]. By so doing, we may use a representation of the relevant aspects of the world to measure the level of "goodness" of a given state of world. By the same token we may measure the goodness of a given state of the world and compare it with the goodness of the new state that is reached after a given action is taken.[1]

2.3 The Space of Action

We follow the postulates of [12]: (i) there are two primitive (different) components in a SCTS: social world and agents; (ii) the social world has a fixed ontology; (iii) at any moment in time there is an explicit *state of the world* that is unique and the same for all agents; and (iv) the state of the world changes only through admissible events and those agent actions that comply with the regulations of the space. Furthermore, because of consequentialism, we assume (i) that the state of the world can be assessed —and thus we can compare two states and hence prefer one over the other— and (ii) that since the world changes

[1] Consequentalism eventually commits to commensurable values, however one needs not commit to any particular aggregation function. For a single system there may be different ways of computing the aggregation, different observable variables involved in a value and different forms to "score" those variables.

when an agent executes an action, one may assess whether the effect of an action produces a better state of the world or not.

2.4 Related Work

In [7], the authors use values to select among different possible plans. They use the same approach as [23], where the agent has a desired level of satisfaction for each specific value. Different actions have different effects on the current satisfaction levels. For a given goal, the plan composed by actions that "best" modify the current value levels in the direction of the desired ones is the one that is preferred. Unlike these authors, in this paper we are not concerned about the actual motivation of agents, nor about their evaluation process. However we share with them the notion that actions promote or demote values and that a hierarchy of values —thus a notion of preference— is one of the elements involved in the decision of which action to take.

Also close to our main assumptions are value-based argumentation proposals where the main idea is that an argumentation move will promote, demote or be indifferent towards a value [6]. In particular, [3] proposes an argument scheme (PRAS) where a transition between states of the world is labelled for each value according to the promotion/demotion effect induced by an action; sort of like we do but with the purpose that plans, goals and the promotion of values can be explicitly reasoned about. Likewise, in [24], the authors propose an argumentation mechanism (to deal with disagreement about the meaning of a value), where values are defined as preference orders, similar to our own, and the authors discuss the problem of arguing about the values promoted or demoted by an action. In [5], arguments for and against actions are used to help agents choose between actions based on their preferences over these values, and the proposed approach aims at justifying norms in general as well as reasoning about when norms should be violated.

In [21], the authors are interested in the problem of choosing a set of norms that best fit the moral values of a society. In their approach, however, the authors neither provide a formal definition of what values are, nor do they provide a formal model for assessing how much a norm supports (or is aligned with) a value, which they assume is given. While they use preferences to specify which value is preferred to which other value, we use preferences over the states of the world to help us define a single value. As a result, we then assess the alignment of a norm to a value based on whether this norm allows us to move to preferred states with respect to the value in question.

3 A Formal Approach to Values

3.1 Values as Preferences

In this section, we propose a formal model to represent values and ways to compute value aggregation. Values are formally understood as preferences over

behaviour or preferences over states of the world; this is in line with the opinion of many philosophers [8] and coherent with the model introduced in the background. These preferences usually reflect one's sense of right and wrong, good or bad, and hence, help decide what course of action might be "better." For example, if equality between men and women is one of your values, then you would prefer a state of the world where women and men are paid equally as opposed to one where they are not, and you would support actions leading towards that state of the world.

We adopt here the traditional view of the world as a labelled transition system [9]: that is, the world is described as a set of states and the different actions allow us to move from one state of the world to another.

Definition 1. *The world is defined as a labelled transition system* $(\mathcal{S}, \mathcal{A}, T)$*, where* $\mathcal{S}$ *is a set of states,* $\mathcal{A}$ *is a set of actions, and* $\mathcal{T}$ *is a set of labelled transitions (*$T \subseteq \mathcal{S} \times \mathcal{A} \times \mathcal{S}$*). For simplification, we use the notation* $s \xrightarrow{a} s'$ *to describe the transition* $(s, a, s') \in T$*.*

Values, in a given world, then specify which states of the world are preferred to which other states and to what degree. Note that in our model values are individual and thus have to be associated to a particular agent, human or software.

Definition 2. *A value-based preference* Prf *over pairs of world states describes how much preferred is one state of the world over another for a given agent with respect to a given value,* $\mathsf{Prf} : S \times S \times G \times V \rightarrow [-1, 1]$*, where* G *is the set of agents and* V *is the set of values. We use the notation* $\mathsf{Prf}^{\alpha}_{v}(s, s')$ *to describe how much does* $\alpha \in G$ *prefer the state of the world* $s' \in S$ *over* $s \in S$ *with respect to value* $v \in V$*.*

We set the range of preferences to be $[-1, 1]$, where a positive number illustrates that s' is more preferred to s, a negative number illustrates that it is less preferred, and 0 illustrates that they are equally preferred. The larger (smaller) the number is, then the more (less) preferred the latter state is to the former.

3.2 Aggregation of Value-Based Preferences

Although an agent might be capable of assessing what states of the world it prefers with respect to a particular value, trade-offs among values and cumulative effects make that it establishes overall preferences over states of the world for *groups of values.*[2] Similarly, from a social perspective, the determination of the joint preferences of a group of agents over the states of the world is key to enable their joint planning.[3]. Thus a number of aggregation functions can be defined:

[2] For instance, given $\mathsf{Prf}^{\alpha}_{Security}(s, s') > \mathsf{Prf}^{\alpha}_{Privacy}(s, s')$ what should be $\mathsf{Prf}^{\alpha}_{\{Security, Privacy\}}(s, s')$

[3] Given $\mathsf{Prf}^{\alpha}_{Security}(s, s')$ and $\mathsf{Prf}^{\beta}_{Security}(s, s')$ what should be the value of $\mathsf{Prf}^{\{\alpha,\beta\}}_{Security}(s, s')$

– *One's preference with respect to a set of values:* calculated by aggregating one's preference over each value in that set.

$$\mathsf{Prf}^{\alpha}_{V} = p(\{\mathsf{Prf}^{\alpha}_{v}\}_{v \in V}) \tag{1}$$

– *A group of people's preference with respect to a given value:* calculated by aggregating each person's preference over that given value.

$$\mathsf{Prf}^{G}_{v} = q(\{\mathsf{Prf}^{\alpha}_{v}\}_{\alpha \in G}) \tag{2}$$

– *A group of people's preference with respect to a set of values:* calculated by aggregating each person's preference over each value in the set.

$$\mathsf{Prf}^{G}_{V} = f(\{\mathsf{Prf}^{\alpha}_{V}\}_{\alpha \in G}) \tag{3}$$

$$\mathsf{Prf}^{G}_{V} = g(\{\mathsf{Prf}^{G}_{v}\}_{v \in V}) \tag{4}$$

Figure 2 illustrates the relationships between the different aggregation functions. There may be cases that aggregation functions chosen by all agents in a system are coherent with respect to these relationships, for instance a trivial function that works well is the arithmetic average:[4]

$$\mathsf{Prf}^{\alpha}_{V}(s, s') = \frac{\sum_{v \in V} \mathsf{Prf}^{\alpha}_{v}(s, s')}{|V|}$$

$$\mathsf{Prf}^{G}_{v}(s, s') = \frac{\sum_{\alpha \in G} \mathsf{Prf}^{\alpha}_{v}(s, s')}{|G|}$$

as

$$\mathsf{Prf}^{G}_{V}(s, s') = \frac{\sum_{\alpha \in G} \mathsf{Prf}^{\alpha}_{V}(s, s')}{|G|} = \frac{\sum_{v \in V} \mathsf{Prf}^{G}_{v}(s, s')}{|V|}$$

However, in general each agent may choose to combine its preferences over values using a different function/method and thus socially agreeing of a preference over a set of values will depend on the order in which the aggregations are made.

3.3 Value-Based Preferences based on State Properties

As illustrated earlier, values specify our preferences over the states of the world. For example, if one values equality between men and women then s/he will most probably prefer a state of the world where men and women are equally paid to another where women are underpaid.

What distinguishes one state of the world from another are the properties that hold in that state of the world. As such, it is these state properties that

[4] Other average functions would also yield coherence.

$$\begin{array}{ccc} \{\mathsf{Prf}_v^\alpha\}_{\alpha\in G, v\in V} & \xrightarrow{p} & \{\mathsf{Prf}_V^\alpha\}_{\alpha\in G} \\ \downarrow q & & \downarrow f \\ \{\mathsf{Prf}_v^G\}_{v\in V} & \xrightarrow{g} & \mathsf{Prf}_V^G \end{array}$$

Fig. 2: The different value-based preferences and the different aggregation functions

influence the preferences between the states of the world. For that, we say values must be related to state properties.

For example, if one values equality between men and women then the value-based preferences should be influenced by the satisfaction of properties, such as: 1) women and men receiving same salaries, 2) maternity and paternity leaves being equal, etc. Though value-based preferences with respect to equality between men and women should not be influenced by, say, the property of engineer's salaries being in the range [€40,000 – €50,000].

Let Φ_v be the set of properties relevant to value $v \in V$. We then say that any value based preference $\mathsf{Prf}_v(s, s')$ must be dependent on the satisfaction of Φ_v at states s and s'; that is,

$$\mathsf{Prf}_v(s, s') = f(\mathrm{P}(s \models \Phi_v), \mathrm{P}(s' \models \Phi_v)) \tag{5}$$

where $\mathrm{P}(s \models \Phi_v)$ describes the probability of the satisfaction of the set of properties Φ_v at state s, i.e., the degree of satisfaction of Φ_v at state s.

Defining the probability $\mathrm{P}(s \models \Phi_v)$, as well as defining the function f, is outside the scope of this paper and is left for future work; though the example at the end of this paper illustrates how preferences can be based on state properties.

4 The Value-Alignment Problem

The value alignment problem is described, informally, as how much aligned are agents' decisions, and hence actions, with the values that the agents hold dear to them. And since behaviour (decisions and actions) is governed by norms, we describe this alignment as an alignment between the norms that govern behaviour and the values that are held in high regard.

We understand norms as rules that govern behaviour. We say a norm $n \in N$ (where N is a set of norms) is a logical formula that describes the conditions under which a certain action can/cannot be performed along with the post-conditions of that action. When a set of norms N is applied to a world $(\mathcal{S}, \mathcal{A}, T)$, the world is modified by he norms in N, resulting in a new world $(\mathcal{S}, \mathcal{A}, N, T_N)$, which we refer to as a normative world. For example, in a world where people do not get taxed, your money increases by the amount of your salary when your salary is paid (see the action 'salary_received' and the new state s' of Figure 3a, where $Money$ describes how much money one has at a given state). However, the

norm of another country that introduces a 20% tax on your income essentially modifies the action of receiving your salary ('salary_received') by applying the tax, and hence, resulting in a transition to a new state where your income is deducted by 20% (state s'' in Figure 3b).

(a) A world with no taxes (b) A world with 20% taxes

Fig. 3: Applying a norm to a given world alters the transitions and their resulting states

Definition 3. *A normative world* $(\mathcal{S}_N, \mathcal{A}, N, T_N)$ *describes the world* $(\mathcal{S}, \mathcal{A}, T)$ *where the set of norms* N *have been applied to the transitions in* T*, resulting in possibly new transitions and states.*

How much a given norm $n \in N$ is aligned to a given value $v \in V$ with respect to a world $(\mathcal{S}, \mathcal{A}, T)$ then depends on whether applying norm n would result in new transitions (T_N) that would move us to preferred (and possibly new) states or not. To be able to calculate this, we will need to have a list of the different paths in a given world, which we define accordingly.

Definition 4. *A path* p *in a world* $(\mathcal{S}, \mathcal{A}, T)$ *is a sequence of transitions in* T*:* $[s \xrightarrow{\alpha} s', \ldots, s'' \xrightarrow{\beta} s''']$*, such that* $p_F[i] = p_I[i+1]$*, where* $p_I[i]$ *represents the initial state of the* i*th transition in* p *and* $p_F[i]$ *represents the final state of the* i*th transition in* p*. In other words, the final state of every transition equals the initial state of the following transition.*

Alignment is then defined as follows.

Definition 5. *The degree of alignment of a norm* $n \in N$ *with a value* $v \in V$ *with respect to a world* $(\mathcal{S}, \mathcal{A}, T)$ *for a given agent* α *is defined through the accumulated preferences in the resulting normative world that applies norm* n *(that is, the world* $(\mathcal{S}_N, \mathcal{A}, N, T_N)$*), which is specified as:*

$$\mathsf{Algn}^{\alpha}_{n,v}(\mathcal{S}, \mathcal{A}, T) = \frac{\sum_{p \in paths} \sum_{d \in [1, length(p)]} \mathsf{Prf}^{\alpha}_{v}(p_I[d], p_F[d])}{\sum_{p \in paths} length(p)}$$

where $paths$ *is the set of all paths in world* $(\mathcal{S}_{\{n\}}, \mathcal{A}, \{n\}, T_{\{n\}})$*, and* $length(p)$ *describes the length of a path* $p \in paths$*.*

In other words, considering all possible paths in the normative world that applies norm n, $(\mathcal{S}_{\{n\}}, \mathcal{A}, \{n\}, T_{\{n\}})$, we calculate the average change in preferences for each transition of those paths. Of course, our proposal for calculating alignment is an initial proposal that gives equal weight to all paths and all transitions of a path. Alternative approaches may also be considered, which we leave for future work. For example, the variance in the preferences might be indicative, say if one prefers steady increase in preferences over fluctuating preferences. Furthermore, if more knowledge about this world is available, such as the probability of transitions, then one can take this knowledge into consideration and can give less probable paths (or transitions) less weight than others. Similarly, if knowledge about which states are more important (regardless of whether they are less or more preferable), the preference of these states can be given more weight. One, for example, may want to give more weight to states that are in the distant future than those that are in the immediate future as the distant future might be more important.

Note that as preferences are subjective with respect to a given agent α, the alignment of a norm to a value is then also subjective with respect to the same agent α. Also note that alignment is described by positive numbers whereas misalignment by negative numbers, and the higher the number, then the more aligned is the norm with the value in question, and vice versa.

Of course, the definition above requires calculating the preferences between states for all possible transitions in a given world $(\mathcal{S}_{\{n\}}, \mathcal{A}, \{n\}, T_{\{n\}})$. This is not an efficient approach. As such, we propose to use the Monte Carlo sampling method to randomly select some of the paths in this world. Furthermore, we also suggest to restrict the length of these paths, which is useful especially in infinite state spaces. We say let l describe the length of paths, and x the number of sampled paths. Then, alignment can be calculated as follows:

$$\mathsf{Algn}^{\alpha}_{n,v}(\mathcal{S}, \mathcal{A}, T) = \frac{\sum_{p \in paths'} \sum_{d \in [1,l]} \mathsf{Prf}^{\alpha}_{v}(p_I[d], p_F[d])}{x * l} \tag{6}$$

where $paths'$ is a set of x randomly selected paths of length l in the normative world $(\mathcal{S}_{\{n\}}, \mathcal{A}, \{n\}, T_{\{n\}})$.

As with preferences, alignment can be calculated for sets of values, sets of norms, and/or sets of agents, so we can calculate $\mathsf{Algn}^{\alpha}_{n,V}(\mathcal{S}, \mathcal{A}, T)$, $\mathsf{Algn}^{\alpha}_{N,V}(\mathcal{S}, \mathcal{A}, T)$, or $\mathsf{Algn}^{G}_{N,V}(\mathcal{S}, \mathcal{A}, T)$, and so on.

In addition to alignment, we also define the relative alignment of norm n_1 with respect to n_2 for a given value v accordingly.

Definition 6. *The relative alignment of norm n_1 with respect to n_2 for a given value v in a given world $(\mathcal{S}, \mathcal{A}, T)$ describes how much more n_1 is aligned with v than n_2 is aligned with v, and it is specified as:*

$$\mathsf{RAlgn}^{\alpha}_{n_1/n_2,v}(\mathcal{S}, \mathcal{A}, T) = \mathsf{Algn}^{\alpha}_{n_1,V}(\mathcal{S}, \mathcal{A}, T) - \mathsf{Algn}^{\alpha}_{n_2,V}(\mathcal{S}, \mathcal{A}, T)$$

Where positive numbers imply n_1 is more aligned than n_2 with respect to v, and negative numbers imply the opposite holds.

Again, as above, the relative alignment can be calculated for sets of values, sets of norms, and/or sets of agents, as needed.

Last, but not least, we note that computing alignment is based on the assumption that all transitions in a given world are given the same weight. In other words, we assume that transitions are equiprobable to occur. Of course, in reality this is not true due to a couple of reasons. First and foremost, the probability of reaching a given state is not the same for all states, as the norms (or the rules that govern behaviour) might result in having one state more (or less) probable to reach than others. Second, the probability of agents choosing one action over another cannot be predicted and decisions are usually not equiprobable. However, for simplicity, this paper assumes all transitions are equiprobable.

5 Example

Let's illustrate the concept of value alignment with a simple example: the traditional Prisoners' Dilemma whose payouts are presented in Table 1. The game is played repeatedly. Although traditionally this example assumes self interest and rationality of the players, we'll see that within our framework we can tweak it (via norms) so that values other than selfishness can be accommodated.

Table 1: The Prisoners Dilemma.

	β **Co-operates**	β **Defects**
α **Co-operates**	6,6	0,9
α **Defects**	9,0	3,3

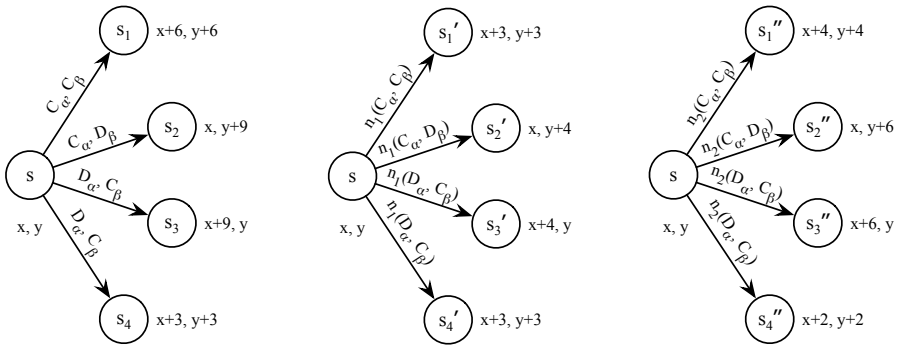
The states of the world are described through the accumulated gain of each agent, which we will represent as (x, y), where x stands for α's accumulated gain and y for β's. Every time a game is played, extra gains are accumulated to make the state change from one state (s) to another (s'), with their properties changing accordingly from (x, y) to (x', y'), with $x' \geq x$ and $y' \geq y$.

5.1 Value-Based Preferences

In this example, we will consider the value *equality*. Equality might mean different things for different people, and it is usually valued differently by different people. We present four different functions that could be used to define preferences with respect to the value 'equality'. Note that preferences are defined in terms of the properties specifying the accumulated gains at each state.

– States with higher equality in accumulated gain are preferred:

$$\mathsf{Prf}(s, s') = \frac{|x - y|}{\max\{x, y\}} - \frac{|x' - y'|}{\max\{x', y'\}} \tag{7}$$



(a) n_0 applied (b) n_1 applied (c) n_2 applied

Fig. 4: Applying norms alters the world

– States with higher equality in accumulated gain are preferred only if my personal gain is not lower:

$$\mathsf{Prf}(s, s') = \left(1 - \frac{|y' - x'|}{\max\{x', y'\}}\right) \cdot \frac{x' - x}{\max\{x', x\}} \tag{8}$$

– States with higher personal gain are preferred only if equality is not lower:

$$\mathsf{Prf}(s, s') = \frac{x' - x}{2(\max\{x', x\})} - \frac{y' - y}{2(\max\{y', y\})} \tag{9}$$

– States with higher personal gain are preferred, regardless of equality:

$$\mathsf{Prf}(s, s') = \frac{x' - x}{\max\{x', x\}} \tag{10}$$

5.2 Norms

We will define three examples of norms that introduce taxes over the gains of agents playing the prisoner's dilemma:

No taxing. n_0: No taxes are to be payed.
Incremental taxing. n_1: Taxes will be paid as follows: no taxes to be paid when the gain is 0 or 3, 3 to be paid as taxes when the gain is 6, and 5 to be paid as taxes when the gain is 9.
Fixed taxing. n_2: 1/3 of the gains of each game is to be paid as taxes.

Norms n_1 and n_2 modify the world that applies n_0 (Figure 4a), as illustrated by Figures 4b and 4c, respectively.

5.3 Value Alignment

The question now is which norms are better aligned with an agent's interpretation of the value 'equality' (where different interpretations are possible, following the different Equations 7–10). The iterated prisoner's dilemma outcome depends on the strategies played by the agents. As this aspect is not the focus of this paper, we will assume that agents will choose their actions randomly. In columns two and three of Table 2, you can find the set of actions from which each agent chooses their actions. In column one, the preference modeling value equality is found, and in the last column, the relative alignment of the three norms for that preference are presented.

Table 2: The relative alignment of norms n_0, n_1 and n_2 with respect to different definitions of the value equality that α may adopt and different sets of actions that α and β may choose from, randomly. Sampling is set to 20,000 and the game is played 10 times (that is, $x = 20,000$ and $l = 10$ in Equation 6).

	$\mathsf{Prf}^{\alpha}_{eq}$	α's actions	β's actions	Relative Alignment
1	Eq. 7	{c}	{c,d}	$n_1 \succ n_0 \sim n_2$
2	Eq. 8	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
3	Eq. 9	{c}	{c,d}	$n_0 \sim n_1 \sim n_2$
4	Eq. 10	{c}	{c,d}	$n_0 \succ n_2 \succ n_1$
5	Eq. 7	{d}	{c,d}	$n_1 \succ n_0 \sim n_2$
6	Eq. 8	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
7	Eq. 9	{d}	{c,d}	$n_0 \sim n_1 \sim n_2$
8	Eq. 10	{d}	{c,d}	$n_0 \sim n_1 \sim\succ n_2$
9	Eq. 7	{c,d}	{c}	$n_1 \succ n_0 \sim n_2$
10	Eq. 8	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
11	Eq. 9	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
12	Eq. 10	{c,d}	{c}	$n_0 \sim n_1 \sim n_2$
13	Eq. 7	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
14	Eq. 8	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
15	Eq. 9	{c,d}	{d}	$n_1 \succ n_0 \sim n_2$
16	Eq. 10	{c,d}	{d}	$n_0 \sim n_1 \succ n_2$
17	Any	{c,d}	{c,d}	$n_0 \sim n_1 \sim n_2$

The results of Table 2 illustrate the following. No matter the actions chosen by α or β, the norm better aligned with a strong support to equality (specified through Equation 7, see lines 1, 5, 9, 13) is incremental taxing (n_1). Moderate supporters of equality (specified through Equations 8 and 9) have no norm specially well aligned (lines 2, 3, 6, 7, 10, and 11), except when the gains of β are higher (by always choosing to defect: choosing action d) in which case they consider incremental taxing better aligned (n_1) (lines 14, 15). Finally, when there is a random selection over $[c, d]$ by both players (line 17) leading then to an egal-

itarian society, there is no preferred norm, as none of them increase inequality over an egalitarian society.

6 Conclusions and Suggested Work

This paper has provided a formal model that defines values as preferences over states of the world, and value-alignment through the increase/decrease of preferences in a given world. A computational model has been presented for calculating the degree of value-based preferences, the degree of the alignment of a norm to a value in a given world, as well as the relative alignment of one norm with respect to another for a given value in a given world.

Future work should help define the different aggregation functions for values (Equations 1–4). For example, how do social values arise from individual values? Future work should also help define the f function and the probability $\mathrm{P}(s \models \Phi_v)$ of Equation 5, which describe how preferences are generated. Last, but not least, future work should study the impact of the assumption made that all transitions are considered equiprobable.

Nevertheless, with this initial formal model for values and value-alignment, we can now formalise questions that can help us study agent societies. Given a set of norms N, a set of values V and set of agents G such that $\mathsf{Algn}^{\alpha}_{n,v}$ and $\mathsf{Prf}^{\alpha}_{v}$ are known for all $n \in N$, $v \in V$, and $\alpha \in G$, then we can formalise the following questions:

Question 1. What is the subset of norms $N^* \subseteq N$ with optimal alignment for group G? That is, how to compute:

$$N^* = \arg\max_{N' \subseteq N} \mathsf{Algn}^{G}_{N',V}$$

Question 2. What is the subset of agents $G^* \subseteq G$ better aligned with norms N? That is, how to compute:

$$G^* = \arg\max_{G' \subseteq G} \mathsf{Algn}^{G'}_{N,V}$$

Question 3. What is the optimal social preference aggregation function? That is, how to compute:

$$f^* = \arg\max_{f \in F} \mathsf{Algn}^{G'}_{N,V}(f\{\mathsf{Prf}^{\alpha}_{V}\}_{\alpha \in G})$$

7 Acknowledgments

This work has been supported by the Catalan funded AppPhil project (funded by RecerCaixa 2017), the Spanish funded CIMBVAL project (funded by the Spanish government, project # TIN2017-89758-R), and the EU funded WeNet project (funded under the H2020 FET Proactive 2018 call, project # 823783).

References

1. Aldewereld, H., Boissier, O., Dignum, V., Noriega, P., Padget, J. (eds.): Social Coordination Frameworks for Social Technical Systems, Law, Governance and Technology Series, vol. 30. Springer International Publishing (2016)
2. Andrighetto, G., Governatori, G., Noriega, P., van der Torre, L.W.N. (eds.): Normative Multi-Agent Systems, vol. 4. Dagstuhl Publishing (2013)
3. Atkinson, K., Bench-Capon, T.J.M.: States, goals and values: Revisiting practical reasoning. Argument & Computation **7**(2-3), 135–154 (2016)
4. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I.: The moral machine experiment. Nature p. 1 (2018)
5. Bench-Capon, T., Modgil, S.: Norms and value based reasoning: justifying compliance and violation. Artificial Intelligence and Law **25**(1), 29–64 (Mar 2017). https://doi.org/10.1007/s10506-017-9194-9, https://doi.org/10.1007/s10506-017-9194-9
6. Bench-Capon, T.J.M.: Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation **13**(3), 429–448 (2003). https://doi.org/10.1093/logcom/13.3.429, http://dx.doi.org/10.1093/logcom/13.3.429
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, IJCAI-17. pp. 178–184 (2017). https://doi.org/10.24963/ijcai.2017/26, https://doi.org/10.24963/ijcai.2017/26
8. Ferrater-Mora, J.: Diccionario de filosofía. Ariel (revised by J-M Terricabras) (1994)
9. Gorrieri, R.: Labeled Transition Systems, pp. 15–34. Monographs in Theoretical Computer Science. An EATCS Series. Springer (2017)
10. Mercuur, R., Dignum, V., Jonker, C.: The use of values for modeling social agents. In: Quan Bai, Fenghui Ren, M.Z.T.I. (ed.) Proceedings of the 3nd International Workshop on Smart Simulation and Modelling for Complex Systems (2017)
11. Miceli, M., Castelfranchi, C.: A cognitive approach to values. Journal for the Theory of Social Behaviour **19**(2), 169–193 (1989)
12. Noriega, P., Verhagen, H., d'Inverno, M., Padget, J.: A manifesto for conscientious design of hybrid online social systems. In: Cranefield, S., Mahmoud, S., Padget, J., Rocha, A.P. (eds.) Coordination, Organizations, Institutions, and Norms in Agent Systems XII - COIN 2016 International Workshops, COIN@AAMAS, Singapore, Singapore, May 9, 2016, COIN@ECAI, The Hague, The Netherlands, August 30, 2016, Revised Selected Papers. Lecture Notes in Computer Science, vol. 10315, pp. 60–78. Springer (2016). https://doi.org/10.1007/978-3-319-66595-5, https://doi.org/10.1007/978-3-319-66595-5
13. Parks, L., Guay, R.P.: Personality, values, and motivation. Personality and Individual Differences **47**(7), 675–684 (2009)
14. Van de Poel, I.: Values in engineering design. In: Meijers, A.W.M. (ed.) Handbook of the Philosophy of Science, pp. 973–1006. Elsevier (2009)
15. van de Poel, I.: Translating values into design requirements. In: Michelfelder, D.P., McCarthy, N., Goldberg, D.E. (eds.) Philosophy and Engineering: Reflections on Practice, Principles and Process, pp. 253–266. Springer Netherlands, Dordrecht (2013)
16. Reiss, S.: Who Am I?: The 16 Basic Desires That Motivate Our Actions and Define Our Personalities. Berkley Pub. (2002)

17. Rizzo, A.: Ethically Aligned Design, Version 2 (Dec 2017), https://standards.ieee.org/industry-connections/ec/ead-v1.html
18. Rohan, M.J.: A rose by any name? the values construct. Personality and Social Psychology Review **4**(3), 255–277 (2000)
19. Russell, S.: Provably beneficial artificial intelligence. The Next Step: Exponential Life, BBVA-Open Mind (2017)
20. Schwartz, S.H.: Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Advances in experimental social psychology, vol. 25, pp. 1–65. Elsevier (1992)
21. Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A., Rodríguez, M., Wooldridge, M., Morales, J., Ansótegui, C.: Moral values in norm decision making. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018. pp. 1294–1302 (2018), http://dl.acm.org/citation.cfm?id=3237891
22. Sinnott-Armstrong, W.: Consequentialism. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2015 edn. (2015)
23. di Tosto, G., Dignum, F.: Simulating social behaviour implementing agents endowed with values and drives. In: Multi-Agent-Based Simulation XIII - International Workshop, MABS 2012, Valencia, Spain, June 4-8, 2012, Revised Selected Papers. pp. 1–12 (2012). https://doi.org/10.1007/978-3-642-38859-0_1, https://doi.org/10.1007/978-3-642-38859-0_1
24. van der Weide, T.L., Dignum, F., Meyer, J.C., Prakken, H., Vreeswijk, G.A.W.: Practical reasoning using values: Giving meaning to values. In: Proceedings of the 6th International Conference on Argumentation in Multi-Agent Systems. pp. 79–93. ArgMAS'09, Springer-Verlag, Berlin, Heidelberg (2010)