Journal of Artificial Intelligence Research 74 (2022) 1739-1774 Submitted 11/2021; published 07/2022

Synthesis and Properties of Optimally Value-Aligned Normative Systems

Nieves Montes NMONTES@IIIA.CSIC.ES
Carles Sierra SIERRA@IIIA.CSIC.ES
Artificial Intelligence Research Institute (IIIA-CSIC)
Campus UAB Carrer de Can Planas, Zona 2
08193 Bellaterra, Barcelona

Abstract

The value alignment problem is concerned with the design of systems that provably abide by our human values. One approach to this challenge is through the leverage of prescriptive norms that, if carefully designed, are able to steer a multiagent system away from harmful outcomes and towards more beneficial ones. In this work, we first present a general methodology for the automated synthesis of value aligned normative systems, based on a consequentialist view of values. In the second part, we provide analytical tools to examine such value aligned normative systems, namely the Shapley value of individual norms and the compatibility of several values under a fixed set of norms. We illustrate all of our contributions with a running example of a society of agents where taxes are collected and redistributed according to a set of parametrised norms.

1. Introduction

In recent years, the term *value alignment* has been used to refer to the challenge of building artificial intelligence (AI) systems that comply, uphold, and respect the moral values that our societies care most about. This concern is rooted in the increasing power, autonomy and ubiquity of these technologies. The complexity of some of the algorithms that power AI systems, coupled with the sensitivity of the areas where they are deployed, entail the risk that us, the humans, might lose control over the systems whose primary purpose is precisely to serve us.

Within the multiagent systems (MAS) community, the value alignment problem translates into ensuring that the interactions at the heart of a society of agents are ethically appropriate. At least a subset of these agents are assumed to be software-enabled. Some approaches to value alignment in MAS introduce values as central elements in the reasoning schemes of agents' architectures (Atkinson & Bench-Capon, 2016). However, these methods assume complete access to the inner workings of the agents, a perk that does not apply in situations where the host of the interaction platform is not in charge of developing the agents.

To overcome this limitation, we turn to a widely studied element within the MAS literature: *prescriptive norms* (Savarimuthu & Cranefield, 2011). Prescriptive norms consist of regulations, constraints and directives on the behaviour of agents, possibly accompanied by monitoring and sanctioning provisions for detected violations. Usually, prescriptive norms are imposed by a system designer or central authority. When prescriptive norms are ap-

plied to software-enabled agents, they are often referred to as *technical norms* (van de Poel, 2020).

In this work, we claim that prescriptive norms have the potential to act as the main value-promoting mechanism within a MAS and that they should be leveraged to ensure that the moral values we deem most relevant are upheld by the system. If suitably designed, prescriptive norms are able to steer a MAS towards more ethically compliant outcomes, and are therefore a splendid avenue to engineer moral values into a society of agents. As we elaborate further in the text, prescriptive norms promote or demote end-states, which must then be assessed in terms of their degree of compliance with respect to some value. For this reason, we conceive *value alignment* to be a property of the implemented norms with respect to the values we intend to embed.

Even under the assumption that norms have the potential and should be aligned with respect to values, the challenge remains in finding *which* norms are actually the most aligned. The work presented in this paper tackles this problem by defining and deploying a novel and general methodology to automatically synthesise normative systems based on their degree of value promotion. Furthermore, we also provide an analytical toolbox to help the MAS designer to extract insights about the optimal normative systems that are returned by the value-guided search. Hence, the contributions of this paper are two-fold:

1. First, we present a systematic methodology to synthesise normative systems based on maximal value alignment. Formally, we seek to solve the following optimisation problem:

$$N^* = \arg\max_{N \in \mathcal{N}} \mathsf{Algn}_{N,V} \tag{1}$$

   where $\mathcal{N}$ is the space of possible normative systems, V is the set of values of interest and $\mathsf{Algn}_{N,V}$ is the degree of alignment of normative system N (an element in $\mathcal{N}$) with respect to the values in V.

   Our approach is differentiated from previous ones in that we take an explicitly consequentialist view of the relationship between norms and values. We present and discuss the underlying assumptions of our methodology, to help the reader understand the engineering choices that are made later on. This conceptual commitment allows us to quantify the value alignment of a set of norms with respect to a value (or set of values) through state features that are designated as proxies for the value in question.

   Despite unequivocally adopting a consequentialist position, the methodology we put forward is general enough to be applied to a wide range of MAS, from very simple ones with blindly obedient agents, to others where participants are endowed with complex decision-making models. Our automated synthesis can be applied to either situation, as long as the norms governing the transitions between consecutive states are linked to a set of optimisable parameters. Moreover, no restrictions are placed on the values that such parameters may take: they can be either continuous or categorical, bounded within a domain or not.

   The main steps of our synthesis approach are:

(a) Define the set of variables that completely specify the state of the MAS. These variables define the state space $\mathcal{S}$ that the system will visit as it transitions due to the actions taken by its populating agents and the norms in place.

(b) Define the norms n_i in the normative system N that regulate the transitions between consecutive states of the MAS. Every norm targets a particular aspect of the transition. All norms are parametric on a set of values, whose domains and constraints have to be specified. This parametrisation provides the search space $\mathcal{N}$ over which the maximising search in eq. (1) is performed.

(c) Define the set of values of interest V and the functions that capture their meaning in the context of the MAS. This step provides the *alignment* target function for the search in eq. (1).

(d) Choose a suitable optimisation algorithm and perform one maximising search for every value of interest. Use the search space defined in step (b) and the alignment established in step (c) as the objective function for the search.

2. In the second part of the paper, we provide analytical tools to examine the resulting optimal normative systems. These tools are intended to shine light on *how* are the optimal normative systems previously obtained operating in order to achieve their (hopefully) large degree of alignment, and whether they represent a good compromise between several competing values. Such insights should help the system designer reflect on the choices made prior to running the normative system search, and adapt accordingly if any of the information provided by these tools is deemed as unacceptable.

   The analytical toolbox we contribute to enrich the synthesis methodology is made up of two complementary metrics:

   (a) The concept of *Shapley value of an individual norm* within a given normative system is made possible by adopting the notion that any optimal normative system is a coalition of individual norms working together to promote some value. Our notion of Shapley value is imported straight from the cooperative game theory literature. It examines the interactions among the individual norms, given an alignment function with respect to some value. It helps establish whether an individual norm is critical or unimportant when it comes to promoting a specific value.

   (b) The notion of *value compatibility* quantifies how successful or neglectful of other values are norms that have been optimised for different goals. In other words, how much of a compromise is a normative system able to achieve between competing interests. This metric is complementary to the Shapley value. Just as the Shapley values examine the interaction among individual norms for a fixed value, compatibility looks at interactions between values under a fixed set of norms. Additionally, we also present and discuss the *compatibility maximising normative system*, which is intended to achieve the largest degree of harmony among values.

This paper is organised as follows. In Section 2 we review other works from the literature related to ours. Then, in Section 3 we discuss the philosophical assumptions that our value

alignment formal model is built upon. The formal model itself is presented in Section 4. Next, the optimisation that searches the most value-aligned normative systems is addressed in Section 5. The second part of this paper is comprised of Sections 6 and 7, which examine the Shapley values of individual norms and the compatibility among values, respectively. Finally, the main take-away points, conclusions and future work are presented in Section 8. All of our contributions are illustrated with a running example of a toy social model.

2. Related Work

In the AI literature, the most straightforward approach to incorporate values into autonomous agents comes from the practical reasoning community (Atkinson & Bench-Capon, 2016). There, values are explicitly incorporated into the reasoning schemes of agents for action and plan selection (van der Weide et al., 2010; Visser et al., 2015; Teze et al., 2019) and decisions on rule compliance (Szabo et al., 2020; Bench-Capon & Modgil, 2017). The upside of these strategies for value embedding is their easy explainability and transparency. On the downside, they require a lot of explicitly encoded information, as well as complete access to the inner architecture of the agents. This perk might not be available in some cases, particularly when the host of the platform where agents interact is not in charge for the development of the participating agents.

To address this shortcoming, an alternative approach focuses on the design and implementation of adequate norms for value promotion. Norms in multiagent systems are viewed from one of two perspectives: *conventions* and *prescriptions* (Conte & Castelfranchi, 1999; Grossi et al., 2012). Conventions are patterns of behaviour that spread through a population and emerge as the dominant agent strategy (Morris-Martin et al., 2019), often following an evolutionary process (Sandholm, 2009). Hence, conventions are part of the agents' internal constructs that emerge as the result of an adaptation process, without a central authority involved in the adoption of the convention.

In contrast, prescriptive norms are obligations, prohibitions and permissions that provide guidance on the behaviour of agents. This guidance may be either regimented, where forbidden behaviours are rendered unavailable when the prescription is implemented; or non-regimented, where agents have the ability to disobey a rule, though they might be encouraged to abide by it through some monitoring and sanctioning mechanism (Morris-Martin et al., 2019). Although work on prescriptive norms usually sticks to one of two models, regimented norms can be viewed as an extreme case of non-regimented norms. For example, the representation used by Fagundes et al. (2016) includes a detection probability for every norm. A norm can be made regimented by imposing perfect enforcement, i.e. probability of detection equal to 1. Overwhelmingly, prescriptive norms are synthesised and imposed on the agents forming a MAS by a central authority or mechanism (see e.g. all works cited in the following paragraph), according to some notion of optimality.

Originally, the purpose of prescriptive norms (also referred to as *social laws*) was to ensure conflict-free, coordinated operation of a team of robots (Shoham & Tennenholtz, 1995; Onn & Tennenholtz, 1997). Subsequent more contemporary solutions to achieving coordination through rules include online design that modifies and refines the norms in place at run-time in an open MAS (Morales et al., 2013), and guarantees on the evolutionary stability of the resulting normative system (Morales et al., 2018).

Despite their popularity as coordination mechanisms, the leverage of prescriptive norms as an avenue to embed moral values into a MAS is only recently being explored. Most notably, work by Serramià et al. has tackled this problem both from a qualitative (Serramià et al., 2020) and a quantitative utility-based (Serramià et al., 2018) perspective. Theirs is the work most similar to the one we present here. A key difference should be noted between their approach and ours. Serramià et al. implicitly assume a *deontological* view of the norm-value relationship. The values supported by every candidate norm are encoded in a value support function, without further justification. This function is then fed as an input to the problem-solving algorithm, which is responsible for finding the most value-aligned set of norms under some consistency requirements. This deontological view is also implicitly adopted by Ajmeri et al. (2020). In their work, personal assistant agents reason about the norms in place, and the values and goals of the user (including preferences over values) to select ethically appropriate actions.

In contrast to their approach, in this work we take an explicitly *consequentialist* view of the norm-value relationship. We claim that the support that a norm has for a value has to be empirically assessed by the outcomes that are brought about by it. Hence, computation of value alignment is based on features of the MAS state that serve as proxies for the value under examination.

In this paper, we extend our previous work on the synthesis of value-aligned normative systems (Montes & Sierra, 2021) on several fronts. First, we fully develop the philosophical foundations on values and norms (Section 3) that underlie many of the technical decisions made later on. Second, we exemplify value alignment with respect to an aggregation of two values, in addition to alignment with respect to both values separately (Section 4.2). Third, we expand the discussion on the Shapley value of individual norms (Section 6) by examining its properties from the cooperative game theory literature, and assessing which of these are relevant in our value alignment context. Finally, we also expand the discussion on value compatibility (Section 7) by introducing the Compatibility Maximising Normative System (CMNS) and running an optimisation search to find it in the context of our running example.

3. Underlying Assumptions

Before jumping to the technical part of the paper, we present the philosophical foundations that underlie our formal model of value alignment. With this exercise, we intend to build robust foundations for our computational model and provide sound justifications for the choices we will make during its formulation (Section 4). Our formalisation of values is built on two main points. The first relates to the concept of values as formal objects and their function within a society of agents. The second concerns the concept of norms, their function and their relationship to values.

3.1 The Nature of Values

First, we present our assumptions on *values*. Values are very abstract concepts that have been the object of intense study in philosophy for centuries (Macintyre, 1998). Currently, one of the most widely accepted theories of moral values in psychology and sociology is Schwartz's theory of basic human values (Schwartz, 1992, 2012). The main success of this

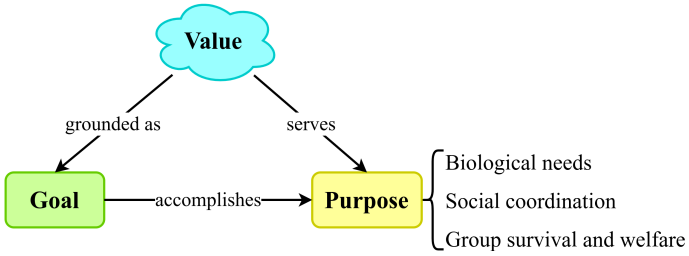

Figure 1: The three components of Schwartz's theory of basic human values and their relationships.

framework has been the identification of a spectrum of moral values that is universally reproduced across cultures. Nevertheless, the conceptualisation of values that it works with is also quite standard across the social sciences and humanities, and explicitly states the characteristics of values that are often implicit in other theories of morality (Rokeach, 1972; Feather, 1995; Spates, 1983). Schwartz's theory establishes the main features of *values* in relation to the *goals* that instantiate their meaning in a particular context and the ultimate *purpose* that they serve. The three concepts are deeply interrelated and their connections are displayed in schematic form in Figure 1.

The features of values that Schwartz's theory outlines and that are relevant to our formalisation are: values (1) are concepts or beliefs; (2) transcend specific situations; (3) refer to desirable goals, end states or behaviours; and (4) serve as evaluation criteria. Features (1) and (2) establish the nature of values and their relationship to physical (or virtual) reality. Values are general, abstract guiding principles (note the cloud-shaped box in Figure 1) that are not linked to any particular social context. Values are omnipresent constructs, that may or may not be relevant to the current context. When a value is relevant to a specific situation, it instantiates an explicit *goal* (bottom left in Figure 1) whose attainment helps further that value. For example, value "equality" in a tax policy context can refer to the effective redistribution of wealth, while in a domestic context it might relate to the even split of house chores between partners. Although the value itself remains unchanged and it is relevant to both this scenarios, its "content" is tailored to the scenario it applies to. To express the relationship between an abstract value and the meaning it takes in a particular context, we say that *an explicit goal g grounds the semantics of value v in context C*.

Features (3) and (4) state the usage of values. Essentially, values serve as moral measuring standards to make judgments about the outcome of a plan, the current state of the world and/or the actions that lead to it. All of these judgments evaluate adherence to a moral stance: whether they respect, uphold and promote the value of interest. Of course, the specific criteria that evaluates whether an action or a situation complies with a value depends on the context where the judgment is made. Since values are instantiated as situation-dependent goals, in order to assess whether the current state of the world or a particular strategy adhere to a value, we should examine how close they are or lead to the goal that is grounding the meaning of the value in that context.

In summary, values are abstract concepts that become operational in the form of explicit goals in a particular context to enable judgments on the world around us. However, beyond values and their grounding goals, there is another fundamental aspect of values left: their *purpose*. According to Schwartz's theory, values are socially desirable constructs to help us cope with three requirements of human existence: the biological needs of individuals, requisites of coordinated social interactions, and survival and welfare needs of groups. Luckily, many of us do not have to think about sheer survival when making everyday decisions. Rather, we justify our actions in terms of our moral values, even if by acting ethically we are, down the line, engaging in beneficial behaviour from an evolutionary perspective.

We argue that, of the three concepts discussed thus far, the *goals* that capture the meaning of values are the best avenue to derive a mathematical formalisation to be embedded into an artificial MAS, for three reasons. First, although values are very abstract, their grounding goals are not, and their fulfilment can be empirically evaluated. Second, we choose goals over purpose since, for technically enabled agents, concerns about evolutionary survival do not really apply. Third, by modelling the goals that ground values, we grant complete control on the *meaning* of values over to the system designer. Agents do not learn how to gradually de-abstract values depending on the context they are in. Rather, the designer is responsible for deciding how does a particular value manifest in the context where the MAS will be operating.

3.2 The Role of Norms

Now that we have established the characteristics of values and how we intend to represent them, we turn to the other main protagonist of this work: norms. As introduced early on, we approach norms from the prescriptive perspective: rules and regulations handed out by a central authority or system designer, that dictate or provide guidance the behaviour of agents, and that affect the outcomes they are able to achieve. This view of norms as prescriptions is differentiated to the view of norms as conventions, where they refer to socially acceptable and expected behaviour emanating from the agents themselves. From here on, whenever we use the word "norms", we are talking about prescriptions.

In this work, we accept the possibility of norms to be *regimented*, i.e. perfectly enforceable. Although regimented norms are not very representative of real-life interactions, in virtual environments they are often technically possible to achieve. In fact, in our running example we will assume norm regimentation, so that we do not need to add extra degrees of freedom related to the agents' decision-making models.

Norms play a central role in our value alignment model because, in our view, they have the potential to ensure ethically compliant outcomes in a society of agents. By modifying incentives and providing opportunities, norms have the ability to steer a community of agents towards particular outcomes. When norms are implemented, the incentive structure and constraints they impose on the agents causes some states to be promoted over others, by making them obligatory or more desirable (e.g. by assigning a penalty to alternative states). Hence, it is expected that some states will be more likely and easily achievable than others. Consequently, the resulting outcome (i.e. the last state visited by the MAS) is strongly dependent on the norms currently implemented. If norms are carefully designed, they facilitate the achievement of outcomes where the goals that ground the meaning of

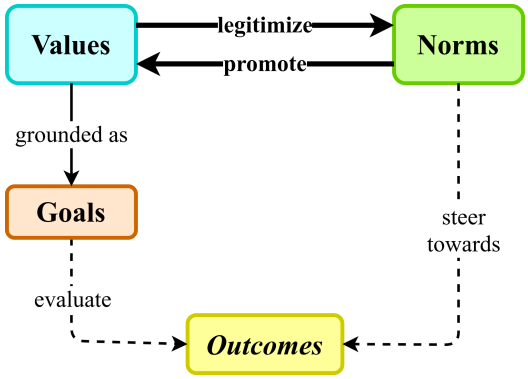

Figure 2: The relationship between *norms* and *values* goes through the grounding *goals* and the achieved *outcomes*.

values are fulfilled to a greater extent than if no regulations were imposed. When a set of norms are successful in this endeavour, we say that they are *aligned with respect to the values* they sought to promote. Hence, we conceive alignment as being a property of norms with respect to values.

Figure 2 illustrates the relationship between values and norms as a diagram. At a shallow level, the two form a feedback loop: norms promote values, and values legitimise the enforcement of norms. At a deeper level, however, the relationship between norms and values is held together by the outcomes that norms steer the system towards and that are favourably evaluated by values. Note that, while norms are context-dependent (they are crafted with a particular domain, e.g. online trade, in mind), values are not. Therefore, values should first be de-abstracted into their grounding goals. Then, these goals can be employed as the standards for which the eventual outcomes (and consequently the norms that lead to them) are held against.

4. Formal Model of Value Alignment

Our computational model to quantify the alignment of a set of norms (or normative system) with respect to some value is inspired by Sierra et al. (2019). In the present framework, a MAS consists of a set of agents G that interact with one another and their environment. The world is modelled as a labelled transition system (Gorrieri, 2017):

Definition 1. The *world* is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T})$, where $\mathcal{S}$ is a set of *states*, $\mathcal{A}$ is a set of *actions* and $\mathcal{T} \subseteq \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ is a set of *transitions* between states labelled by an action.

At any point in time, the MAS is in state $\mathbf{s} \in \mathcal{S}$. This is understood as a *global state*, encoding all the information there is to know about the system. Among $\mathcal{S}$, the initial state is denoted by $\mathbf{s}_0$. Changes in the global state of the system are brought about by the actions $\mathcal{A}$ that agents take. Just as $\mathbf{s}$ is understood to be the global state of the system, action $a \in \mathcal{A}$ refers to the *joint* action profile that is executed by the whole of the agents. Executing joint action a in state $\mathbf{s}$ moves the system to a new state $\mathbf{s}'$. We denote transition $(\mathbf{s}, a, \mathbf{s}')$ with the notation $\mathbf{s} \xrightarrow{a} \mathbf{s}'$.

To simulate long-term evolution of the system, single transitions are not enough. Rather, several transitions are concatenated forming a *path*:

Definition 2. A *path* $p = \mathbf{s}_0 \xrightarrow{a_1} \mathbf{s}_1 \xrightarrow{a_2} ... \xrightarrow{a_n} \mathbf{s}_n$ over the world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$ is a sequence of transitions in $\mathcal{T}$ between states in $\mathcal{S}$, starting at initial state $\mathbf{s}_0$.

We denote the set of paths over the world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$ of length n (visiting $n+1$ states) as $\mathcal{P}_n$. The set of all paths (of any length) over the world is denoted as $\mathcal{P}$. Given a path $p \in \mathcal{P}$, the function $\mathsf{out} : \mathcal{P} \to \mathcal{S}$ returns the last state (or the *outcome*) of path p, $\mathsf{out}(p) = \mathbf{s}_n$.

Running Example

We illustrate all of our contributions with a running example of a toy social model, reminiscent of the public goods game. In this society, a set of technologically-enabled agents are endowed with some initial wealth. To facilitate the exchange of resources, a common fund is set up. Agents contribute to the common fund with a tax amount that is dependent on their economic status. However, a subset of evader agents try to skip the payment. They might get caught and obliged to pay the original taxes plus an additional fine. Finally, the common fund is invested, grown by a fix amount and redistributed back to the agents, regardless of their evading tendencies, in a way that is also dependent on the economic status of the individuals.

In our running example, the MAS is composed of a set of 200 agents $G = \{1, 2, ..., 200\}$. Of these 200, we set 10 to be "evader" agents, which will try to skip payment when the tax collection and redistribution scheme is imposed. This composition of the society (10 out of 200 evader agents) is an arbitrary choice made for exemplification purposes.

All agents have some wealth, which is initialised randomly according to a uniform distribution between 0 and 100. At any given time-step, every agent is characterised by its current wealth x_i as well as an integer that denotes the wealth segment it belongs to. To find the wealth segments, all agents are ranked from highest to lowest wealth and then split into 5 equally populated groups. The agents that belong to the poorest group are assigned to segment #1, while those in the wealthiest group are assigned segment #5. Therefore, an agent at any state is characterised by the tuple $(x_i, seg_i) \in (\mathbb{R}^+ \times \{1, 2, 3, 4, 5\})$, where x_i is agent i's wealth (which is always non-negative) and seg_i is the wealth segment or group it is assigned to. Consequently, the global state space is $\mathcal{S} = (\mathbb{R}^+ \times \{1, 2, 3, 4, 5\})^{|G|}$, where $|G|$ is the number of agents (200). In principle, in the unregulated situation agents can exchange money as they see fit at any transition.

4.1 Norms

In an unregulated world, actions may lead to undesired outcomes. As we have already argued, to avoid this we introduce normative systems as a way to steer the system away from harmful results and in the direction of valued outcomes.

Normative systems are composed of individual norms. An individual norm n_i is a regulation that targets a specific aspect of a state transition. The distinction between different aspects of a state transition will be clarified when we go through the norms in the running example. For the time being, we simply consider a norm to be a regulation that constraints how transitions $\mathcal{T}$ between states in $\mathcal{S}$ take place. The complete transition is

determined by the norms plus the decision-making models of the agents, who choose among the actions available to them.

In this paper, we work with parametric norms. Every n_i is linked to a set of *normative parameters* P_i, alongside with their domains and any possible constraints. When we talk about norms, we refer to the uninstantiated parameters, with no quantities assigned. As an instance closely related to our running example, consider income taxes, which partially determine your wealth increment from one month to the next. In many countries, income tax rates are regulated by a norm that taxes in an incremental way with respect to salary. To instantiate it in a concrete case, numerical parameters have to be set to, for example, decide how the income range is divided into several groups and which rate is applied to every group.

A normative system N is a set of individual norms $\{n_i\}$ with all of their parameters instantiated to some quantity, respecting any domain-dependent constraints. Every norm in N is responsible for regulating an aspect of a transition between two states. Together, the application of the norms in N to the world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$ restricts the original set of transitions to a subset of those, $\mathcal{T}^N \subseteq \mathcal{T}$. Consequently, the set of possible paths $\mathcal{P}$ is also restricted to a subset of the original, $\mathcal{P}^N \subseteq \mathcal{P}$. We use notation $\mathbf{s} \xrightarrow{N} \mathbf{s}'$ to denote the transition between two states as regulated by normative system N.

Note that individual norms by themselves do not determine transitions. Rather, a coalition of (in general) several norms is necessary. Also note that the individual norms that make up a normative system are fully instantiated, as all the parameters they depend on are assigned a quantity. Any normative system N, then, belongs to the family $\mathcal{N}$ of all normative systems composed of the same set of norms but whose parameters take different quantities. The dimensionality of $\mathcal{N}$ depends on the number of parameters needed to specify all norms of a normative system. Essentially, the family of normative systems $\mathcal{N}$ defines a search space consisting of the domains of the normative parameters of its individual norms.

Our methodology does not indicate how should the various aspects of state transitions be itemised into several norms. This is an engineering choice that must be made by the system designer. There is, nonetheless, a guideline that we believe should be respected when formulating the normative system. If there are constraints involving several parameters, these parameters should all be related to the same individual norm. An example of this restriction being respected can be found next, in norm n_2 of our running example.

Running Example

In our social model, we introduce a set of norms (1) to regulate the collection and redistribution of taxes and (2) to randomly detect evader agents and impose a fine on top of their taxes. Thus, within this model, transitions happen under the regulation of normative system $N = \{n_1, n_2, n_3, n_4\}$, where:

$\mathbf{n_1}$ is the norm specifying the tax rate for every wealth segment. It is parametric on the set $P_1 = \{collect_j\}_{j=1,\ldots,5}$, where $collect_j$ corresponds to the fraction of their current wealth that every member of wealth segment j must contribute to the common fund at every transition. All $collect_j$ components have their values bounded in the range $[0, 1]$.

n_2 is the norm specifying how should the invested funds (which grow by a fixed 5% rate) be redistributed back to the agents. It is parametric on the set $P_2 = \{ redistribute_j\}_{j=1,\ldots,5}$, where $redistribute_j$ corresponds to the fraction of the invested common fund that is allocated to the j-th wealth segment. This reimbursement is then shared equally among all agents in the group. Again, all $redistribute_j$ components are bounded in the range $[0, 1]$. Also, the following linear constraint holds:

$$\sum_{j=1}^{5} redistribute_j = 1 \tag{2}$$

implying that the totality of the common fund is reimbursed back to the agents.

n_3 is the norm that determines the probability of detecting evaders, who at every transition attempt to skip payment. It is parametric on a single value $P_3 = \{catch\}$, which corresponds to the probability that any evader agent will be detected and be made to pay its corresponding taxes plus an additional fine on any state transition. To emulate the struggle of fiscal authorities, the range of *catch* is bounded in $[0, 1/2]$, despite corresponding to a probability that could in principle take values in the range $[0, 1]$.

n_4 is the norm specifying how harsh should the punishment be on the detected evaders. Similarly to n_3, it is parametric on a single value, $P_4 = \{fine\}$. Whenever an evader is caught, the amount that it is obliged to contribute to the common fund is equal to the taxes it was trying to evade in the first place (that are dependent on its wealth segment), plus the additional fraction given by *fine*. If the total payment would result in the agent having to contribute with an amount greater than its total current wealth, then the payment equals to the totality of the current wealth of the agent. The value of *fine* is bounded in the range $[0, 1]$.

This example shows how to define a parametric normative system as a set of norms, each of them targeting a concrete aspect of the system's transitions. The family of normative systems $\mathcal{N}$ we work with is the one with components $\{n_1, n_2, n_3, n_4\}$. Our search space, then, is determined by the domains of the normative parameters $P_N = \{collect_j, redistribute_j, catch, fine\}_{j=1,\ldots,5}$ plus the constraint in eq. (2). It should be noted that, despite bearing a remote resemblance with real-life tax codes, the example model is not intended to be a reliable reflection of actual tax policy. It is just a simple example to illustrate our methodology in action.

Second, in our example we do not consider any reasoning schemes by the agents. Evader agents always attempt to evade their payment, while non-evader agents are always compliant and pay their part. This choice is not a feature of our general methodology, but of this example only. We have made this choice in order to reduce the degrees of freedom of our example and focus on the main topic of this work, the synthesis of value-aligned norms.

4.2 Value Alignment

So far, we have modelled a society as a set of agents that transition between states by executing actions. Also, we have introduced norms as restrictions on the feasible transitions.

Now, we want to quantify how effective are those norms at promoting the values we want to embed in the system, i.e. how well *aligned* they are with respect to some values of interest.

In the underlying assumptions of our model (Section 3), we established that values are grounded as goals that evaluate, among other objects, the outcomes that the system achieves. Mathematically, we model these grounding goals as functions over the states of the system that ought to be maximised. As the state space $\mathcal{S}$ is different depending on the MAS in question, the goal that encapsulates the meaning of any value has to be defined in terms of features of that state space, and will not, in general, be applicable to other contexts.

Definition 3. Given a world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$ and a set of values V, the *semantics* of value $v \in V$ in the world is a function $f_v : \mathcal{S} \to [-1, 1]$ that evaluates the states of the world, where $f_v(\mathbf{s}) \sim -1, 0, +1$ indicates that state $\mathbf{s}$ strongly opposes, is neutral or strongly promotes value v, respectively.

Note that Definition 3 entails, as anticipated in Section 3.1, that values need first to be de-abstracted into a function that captures its meaning for the particular domain at hand. Our approach demands this step, and is unable to work with abstract values that have not been grounded first.

As the semantics of any value are given by a function that grades the states of the world, the goal capturing the meaning of the value, then, aims at maximising the corresponding function, by achieving a state that is as compliant towards the value in question as possible. Often, however, one is not just interested in promoting a single value, but rather would like to achieve compliance with respect to several values. Given a set of values $V = \{v_1, ..., v_m\}$ and their semantics functions $f_1, ..., f_m$, we propose that the semantics of the set of values in V should be grounded by an *aggregation* function $F_V : [-1, 1]^m \to [-1, 1]$. F_V takes in the compliance with respect to every individual value and merges them into a single metric. To shorten notation, we denote $F_V(f_1(\mathbf{s}), ..., f_m(\mathbf{s}))$ as $F_V(\mathbf{s})$.

We set the range of all value semantics functions to be bounded in order to facilitate comparison between values. This will become particularly relevant in Section 7, when we look into incompatibilities between values. If the range was unbounded, it would be difficult to establish which value is being actually more aggressively pursued. Therefore, it is highly convenient to grade states in a continuous bounded domain. The choice of the bounds at ± 1 is made out of convenience for the ease of working with unit quantities.

In summary, values evaluate states. How to extend such an assessment to the norms in place? Figure 2 gives a clear hint. Norms steer the system towards (hopefully) beneficial outcomes, that are assessed with the semantics function that capture the meaning of a value in that particular world. Mathematically, we understand an *outcome* to be the final state of a path p in the world, i.e. $\mathsf{out}(p)$. Since norms restrict the available paths, they also limit the final states where they can end. And, just like any other state, the outcome of a path can be graded according to the semantics function of a value.

We put all of these ideas together to define the alignment of a normative system with respect to a value:

Definition 4. Given a world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$, a normative system N that applies to it, and a set of values $V = \{v_1, ..., v_n\}$ with semantics functions $f_1, ..., f_n$, the *alignment* $\mathsf{Algn}_{N,v}$ *of*

normative system N with respect to value $v \in V$ is computed as:

$$\mathsf{Align}_{N,v} = \mathbb{E}\left[f_v\left(\mathsf{out}(\mathcal{P}^N)\right)\right] \tag{3}$$

where $\mathcal{P}^N$ is the random variable of the subset of paths restricted under the normative system N.

Equation (3) states that, in order to compute the alignment of a set of norms with respect to some value, the evolution of the system under the norms in N has to be simulated and let to achieve some outcome. Then, this final state is assessed in terms of the meaning that value v takes in the world, the semantics function f_v. In order to have a statistically significant quantity, the expected value should be computed over a sufficiently large random sample of norm-regulated paths using, e.g. Monte Carlo sampling (Lemieux, 2009). For computational convenience, we also propose to restrict the length of the sampled paths to a fixed number, and hence compute the expected value over $\mathcal{P}^N_n$, for some fixed n.

In line with the consequentialist view we present in Section 3, the alignment in eq. (3) is computed by considering the ultimate consequences (i.e. the end-state) that the implementation of a set of norms brings about. However, other alternatives are possible. For example, one may prefer to compute the alignment by considering the value semantics function applied to all the states that are visited during a path. In that case, one would also need to specify how to aggregate the evaluation of f_v over all the visited states (e.g. average, minimum...).

Once the alignment for a normative system has been established in absolute terms, we can compare several norm sets with one another. To do so, we define the *relative alignment* between two normative systems:

Definition 5. Given a world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$, two normative systems N_1 and N_2 that apply to it, and a value v with semantics function f_v, the *relative alignment between N_1 and N_2* $\mathsf{RAlign}_{N_1/N_2,v}$ *with respect to value v* is computed as:

$$\mathsf{RAlign}_{N_1/N_2,v} = \mathsf{Align}_{N_1,v} - \mathsf{Align}_{N_2,v} \tag{4}$$

Equations (3) and (4) can be readily extended to compute the (relative) alignment with respect to a set of values V. Instead of f_v, the evaluation of path outcomes is made with an aggregation function F_V.

Running Example

In our example tax model, we are interested in the two values $V = \{\textit{equality}, \textit{fairness}\}$. We define semantics functions with respect to the two values individually and for their aggregation as well. These two values will exemplify goals that are not correlated and that are achieved through different taxing strategies.

First, we ground the meaning of value *equality* in the context of our model as, of course, economic equality. To do so, we use the well-known Gini index indicator (Gini, 1912), a widespread metric to quantify wealth and income inequality (The World Bank, Development Research Group, 2019). The Gini index is bounded between 0 (for perfect equality) and 1 (for perfect inequality). In our model, we consider that a normative system N is highly

aligned with respect to equality if, by the end of a fixed-length path, the Gini index is as low as possible:

$$f_{eq}(\mathbf{s}) = 1 - 2 \cdot GI(\mathbf{s}) \tag{5}$$

where $GI(\mathbf{s})$ is the Gini index for the wealth distribution at global state $\mathbf{s}$, which is computed as:

$$GI(\mathbf{s}) = \frac{\sum\limits_{i,j \in G^2} |x_i - x_j|}{2 \cdot |G|^2 \cdot \bar{x}} \tag{6}$$

where x_i denotes the wealth of agent $i \in G$ and $\bar{x}$ the average of the distribution. Note that in eq. (5) we introduce an affine transformation in order to map perfect equality ($GI \sim 0$) to maximum alignment ($f_{eq} \sim 1$) and perfect inequality ($GI \sim 1$) to minimum alignment ($f_{eq} \sim -1$).

Second, we ground the semantics of value *fairness* to mean that evader individuals should be punished for their evasion. Hence, we consider that fairness is being highly promoted if, by the end of a fixed-length path, as many evaders as possible belong to the poorest wealth segment:

$$f_{fair}(\mathbf{s}) = 2 \cdot \hat{\mathbb{P}}\left[seg_i = 1 | evader_i\right] - 1 \tag{7}$$

where $\hat{\mathbb{P}}\left[seg_i = 1 | evader_i\right]$ denotes the estimated probability that the wealth segment of an agent i is the lowest one, provided that they are an evader agent. This estimation is computed as the proportion of evader agents in segment 1 at the final global state. Again, an affine transformation is introduced to map the probability range $[0, 1]$ to the alignment range $[-1, 1]$.

Given that, in our virtual society, there are more agents per wealth group at any time-step (40) than evader agents (10), in the best-case scenario all evaders would end up in segment #1. Consequently, the upper bound for function (7) is $+1$. If there were more evaders than agents per wealth segment, the semantics function would need to be modified so that the potential maximum alignment does not fall below 1.

Third, we turn to the alignment for the aggregation of both values *equality* and *fairness*. Here we take a demanding position. We consider that the set of values $V = \{equality, fairness\}$ are being upheld overall when both values are being simultaneously promoted:

$$F_V(\mathbf{s}) = \begin{cases} -f_{eq}(\mathbf{s}) \cdot f_{fair}(\mathbf{s}) & \text{if } f_{eq}(\mathbf{s}) < 0 \text{ and } f_{fair}(\mathbf{s}) < 0 \\ f_{eq}(\mathbf{s}) \cdot f_{fair}(\mathbf{s}) & \text{otherwise} \end{cases} \tag{8}$$

where $f_{eq}(\mathbf{s})$ and $f_{fair}(\mathbf{s})$ correspond to the semantics functions in eqs. (5) and (7) respectively. The piece-wise definition of $F_V(\mathbf{s})$ is necessary in order to avoid that negative alignment with respect to both *equality* and *fairness* would result in positive alignment with respect to their aggregation.

5. Search of Optimal Normative Systems

The purpose of this section is to demonstrate how to find optimally aligned normative systems with respect to some values, as stated in eq. (1). The search space is determined by the normative parameters, with their domains and constraints, and the objective function

to optimise corresponds to the alignment with respect to the value of interest, computed by eq. (3) as the expectation of the value semantics function on a sample of outcome states.

To perform the search, in this paper we use a Genetic Algorithm (GA) as our optimisation method. This choice, however, is not a defining feature of our methodology. Any optimisation strategy that is suitable, given the domains and constraints of the normative parameters, is apt to perform the task.

Genetic Algorithms (Luke, 2013) are a family of versatile search methods where a population of candidate solutions is maintained. Given the nature of our problem, a candidate in this population consists of a fully instantiated normative system, with a numerical quantity assigned to all of its parameters. Candidates are selected for breeding based on their fitness, i.e. how well aligned is the N instance with respect to the value for which we are optimising. A crossover operation is performed over highly aligned normative systems, in hopes of generating two even better instances. When enough offspring are generated, they substitute the original population. The process is repeated iteratively until some stopping criteria is met.

We initialise our population of candidate normative systems randomly within the bounds allowed for every parameter. For the selection with replacement step, we employ the 1 vs. 1 tournament technique (Miller & Goldberg, 1995). Two normative systems are drawn at random from the population, and the fittest of the two is selected for crossover. This step is repeated once more, to select the other parent to breed with the winner of the previous tournament.

Typically, the crossover step is performed with bit-wise operations, since GAs are mostly employed in optimisation tasks over discrete search spaces. In the case concerning this work, however, we are dealing with a continuous search space defined by the bounds and constraints of the normative parameters as described in Section 4.1. To handle this, we turn to a crossover technique suitable for continuous spaces, intermediate recombination (Mühlenbein & Schlierkamp-Voosen, 1993). This method is controlled by hyperparameter $p \geq 0$, which determines the explorability of the search. The larger p is, the more exploratory the search is. Additional tweaks are introduced in order to ensure that the linear constraint (2) is satisfied.

To enhance the exploitability of the search, we introduce elitism (Baluja & Caruana, 1995) into the algorithm. This technique consists of replacing a small number of the worst newly generated candidates with the same number of the best aligned candidates from their parent generation. This popular variant of genetic optimisation guarantees that the alignment of the most promising normative system will not decrease from one generation to the next. We denote by k the number of "elite" candidates that are carried over from one generation to the next.

For the stopping criteria, we take advantage of the fact that the target alignment functions are bounded up to 1, and hence we set the search to stop when a very promising normative system instance with alignment over some high threshold is found. In order to avoid excessively long searches, the algorithm is halted after a large number of total iterations, and a moderate number of partial iterations, i.e. rounds of the search for which the most promising candidate is not updated. An overview of the hyperparameters of our GA implementation is presented in Table 5 in Appendix A.

Table 1: Optimisation results for the running example with respect to the two values of interest plus their aggregation: optimal normative parameters defining the normative system, and their associated optimal alignment.

Value and semantics function	Optimal normative parameters	Optimal alignment $\text{Algn}^*_{\text{N,v}}$
Equality, eq. (5)	$collect = [20\%, 29\%, 26\%, 35\%, 27\%]$	0.95
	$redistribute = [20\%, 22\%, 19\%, 26\%, 13\%]$	
	$catch = 44\%$	
	$fine = 61\%$	
Fairness, eq. (7)	$collect = [1\%, 30\%, 37\%, 72\%, 66\%]$	0.93
	$redistribute = [2\%, 23\%, 42\%, 24\%, 9\%]$	
	$catch = 45\%$	
	$fine = 56\%$	
Aggregation, eq. (8)	$collect = [2\%, 79\%, 56\%, 65\%, 59\%]$	0.66
	$redistribute = [2\%, 28\%, 25\%, 35\%, 10\%]$	
	$catch = 31\%$	
	$fine = 77\%$	

Other than their versatility (for example in adapting the search to a continuous search space like ours), GAs are particularly attractive for the task at hand because they do not require the analytical formulae of the gradient of the target function with respect to their arguments, i.e. the normative parameters in our case. In other search methods, particularly those based on gradient ascent/descent (Ruder, 2016), disposing of the gradient function is extremely desirable, if not outright necessary. Yet, given the nature of the problem we are tackling, obtaining such an expression is an unnecessarily demanding task, as it would require deriving an expression of the alignment explicitly as a function of the normative parameters.

Although our running example is relatively small, we resort to inexact methods for the optimisation task. In fact, we would encourage readers interested in implementing this methodology for larger, more complex scenarios to stick with meta-heuristics methods (like the GA implemented here or simulated annealing), precisely because they do not impose any requirements of the optimisation target concerning continuity or differentiability. Regardless of the domain on is working with, the optimisation target that this methodology needs to maximise is the alignment, which is empirically estimated by generating a sample of simulation runs and evaluating their outcomes. Hence, this optimisation target is not differentiable. However, note that the computational requirements for the optimisation search are expected to grow with the size of the social model one is working with. Deriving this growth in resources for the optimisation search as a function of the size of the MAS is object for future work.

Running Example: Optimisation Results

In the context of our running example, searching the optimally aligned normative systems means finding which taxing policies maximise the promotion of values *equality* and *fairness* (and their aggregation), according to the meaning we have imbued in these values in eqs. (5) and (7).

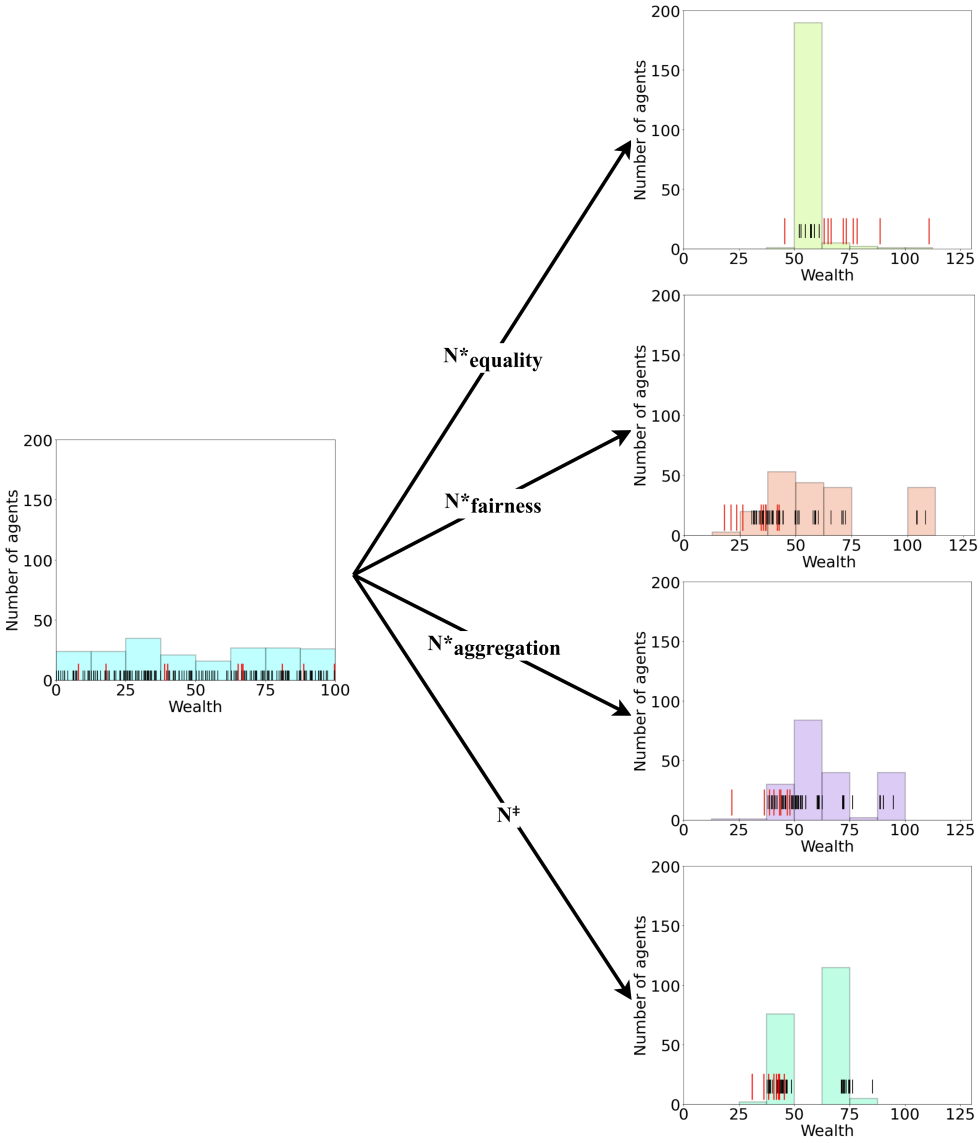

Figure 3: Wealth distribution and rug plot indicating the location of law-abiding agents (regular black marks) and evaders (longer red marks), at the initial state (left) and after a sample path of 10 transitions under the optimal normative system for equality (right first), for fairness (second), for their aggregation (third) and for the normative system that preserves the maximum compatibility between the two values (fourth, see Section 7).

Table 1 presents the optimisation results for the two values we have modelled (*equality* and *fairness*) plus their aggregation. The optimal alignment for equality and fairness separately are very satisfactory, with large positive values > 0.9. For the aggregation of values, the optimal alignment is still fairly good, over 0.6. This decrease in the optimal alignment

for the aggregation with respect to the values individually speaks to how demanding the aggregation of values in eq. (8) is, as it is necessary that both equality and fairness be promoted to a large degree simultaneously in order for the alignment with respect to their aggregation be also high.

We provide an intuitive interpretation of the optimal normative parameters obtained. For equality, the differences between the components of *collect* are small across wealth groups. In practice, this means that wealthier agents contribute to the common fund with more resources in absolute terms, as their wealth is larger. The even redistribution rates across groups then ensure that all agents receive a similar portion of the invested funds. The moderate values for the *collect* and *redistribute* parameters in the optimal model with respect to equality correspond to a compromise between funnelling enough resources from rich to poor agents in order to shrink the wealth distribution, but not channelling too many as to swap them, which would be detrimental towards lowering the Gini index.

For fairness, the normative parameters indicate that another mechanism is in place in order to push evaders towards the poorest group. It is worth noting that neither the probability of catching evaders nor the fine they are imposed are particularly large, they are similar or even smaller than those found for the optimal normative system with respect to equality. Rather, it appears that evaders are pushed towards group #1 by retrieving a lot of resources from the upper wealth groups, where undetected evaders manage to sneak, and then redirecting them towards the middle class. This middle segment is vastly composed of law-abiding citizens, since detected evaders belong to the lower groups and undetected ones belong to the upper ones. Hence, the norms enforce fairness by identifying the wealth groups most likely to include cheaters and directing their wealth elsewhere. It does not collect many taxes from group #1, but the norms keep the cash flow in and out of that group very limited, so that already poor evaders do not have any avenue to enrich themselves.

The optimal parameters with respect to the aggregation of the two values are the most difficult of the three to interpret, since they have to achieve a compromise between effectively punishing evaders and keeping wealth inequality under control. On one hand, the cash flow in and out of wealth group #1 is extremely limited (both $collect_1$ and $redistribute_1$ are ~ 0.02). Since this feature is common to the optimal parameters with respect to fairness, we suspect that its function is similar to the one it played in the optimal norms for fairness only: keep evaders in the lowest wealth segment once they have been detected and pushed there. Additionally, the fine rate to achieve the aggregation of both values is the highest across all optimal normative systems, close to 80%. This indicates that in order to achieve fairness without hurting equality too much, the norms promoting the aggregation of the two values rely more on fines that exclusively target evaders in order to punish them, while the normative system optimised for fairness alone did not consider the collateral harm it could inflict on non-evader agents.

On the other hand, the trends of the *collect* and *redistribute* parameters for the wealth segments other than #1 are somewhat parallel between the optimal normative system for equality alone and the aggregated values. This observation seems to indicate that equality is mostly promoted through a similar taxing strategy to the case when it was only that value being considered.

Figure 3 provides a visual representation of the evolution of the society under each of the three optimal normative systems.[1] These plots visually transmit the fact that the different norm sets lead the system towards different outcomes. Clearly, the final distribution under the optimal normative system for equality is very narrow, but does allow most evaders to become the richest individuals in the population. On the contrary, the final wealth distribution under the optimal norms for fairness is much broader but does push evaders towards the lower positions. The final distribution under the optimal norms for the aggregation of the two values is somewhat in the middle, with a compromise between a moderately narrow wealth distribution and pushing evaders towards the lowest positions in the wealth ranking.

6. Shapley Values of Individual Norms

So far, we have been able to synthesise, in an automated fashion, normative systems that are optimally aligned with respect to values. We have illustrated the methodology in the context of our running example. This synthesis is always contingent on the understanding we have of those values in the context where the MAS operates. Now, the second part of this work begins, where we provide an analytical toolbox to examine the optimal normative systems we have attained more closely. These tools are aimed at providing insights to the system designer on the output of the synthesis process. In particular, we quantify the contribution of every individual norm in an optimal normative system to the overall alignment through their *Shapley values* in this section. In Section 7 we examine *value compatibility*, i.e. how successful an overall set of norms is at compromising between competing values. These metrics should help the system designer reflect on the engineering choices made prior to the automated search and inform any changes to the choice of normative parameters and/or alignment function, before iterating the synthesis-analysis process until satisfactory results are obtained, both regarding a high degree of alignment of the norms with respect to the values of interest and acceptable metrics regarding their Shapley values and compatibility measurements.

In this section, we look at the interaction between individual norms through their Shapley values. We are interested in quantifying how much is a particular norm contributing towards the overall alignment of a normative system with respect to some value. For example, are the rate of evader detection and the fine imposed relevant when it comes to achieving equality, or are they not?

In order to quantify the importance of individual norms, we take the view of any normative system N as a grand coalition of individual norms $\{n_i\}$ working together to achieve high alignment with respect to some value. In order to allocate the credit to the different norms for achieving such promotion, we import the notion of Shapley value (Shapley, 1951) from cooperative game theory, and adapt it to our context:

Definition 6. Given a normative system $N = \{n_i\}$, a value v for which the semantics function f_v in world $(\mathcal{S}, \mathcal{A}, \mathcal{T})$ has been defined, the *Shapley value of norm n_i with respect*

1. Short videos displaying all the intermediate states between the initial and the final global states of a sample path are available, for the four normative systems in Figure 3, at `https://github.com/nmontesg/aamas21/tree/main/videos` and at the online appendix.

to value v is given by:

$$\begin{aligned}\phi_i(v) &= \sum_{N' \subseteq N \setminus \{n_i\}} \frac{|N'|!\,(|N| - |N'| - 1)!}{|N|!} \cdot \left(\mathsf{Align}_{N' \cup \{n_i\}, v} - \mathsf{Align}_{N', v}\right) = \\ &= \sum_{N' \subseteq N \setminus \{n_i\}} \frac{|N'|!\,(|N| - |N'| - 1)!}{|N|!} \cdot \mathsf{RAlign}_{N' \cup \{n_i\}/N', v}\end{aligned} \tag{9}$$

Definition 6 can be readily extended to a set of values V for which an aggregation function F_V has been defined.

The sum in eq. (9) is taken over all normative systems N' from which at least individual norm n_i is not included. Then, the Shapley value for n_i is computed through the relative alignment between the introduction of norm n_i ($\mathsf{Align}_{N' \cup \{n_i\}, v}$) and its absence ($\mathsf{Align}_{N', v}$). Not that if other norms besides n_i are absent from N', they are not to be reintroduced in $N' \cup \{n_i\}$.

Two issues need to be addressed to clarify how the computation ought to be performed: (i) what does it mean for an individual norm n_i to be absent from a normative system N'; and (ii) which normative systems should the sum in eq. (9) include. We address (i) first to be able to answer (ii) later.

Consider an arbitrary normative system instance N, from which we wish to remove some subset of individual norms $\{n_i, n_j, n_k ...\}$ to obtain a new normative system $N' = N \backslash \{n_i, n_j, n_k ...\}$. We denote the numerical quantities upon which norm n_i is parametric in normative system N as $P_i^{(N)}$. To proceed with the removal, we first need to introduce another normative system instance, N_{bsl}, which we refer to as the *baseline* normative system. N_{bsl} belongs to the same family of normative systems as N, meaning that it has the same set of individual norms $\{n_i\}$ related to the same parameters and subject to the same bounds and constraints. However, the numerical quantities for the parameters linked to N_{bsl} are (manually) set in a way as to reflect the lack of evolution. That is, when N_{bsl} is implemented on the MAS, the initial state remains unchanged after an arbitrary number of transitions.

Once N_{bsl} is defined, to remove norm subset $\{n_i, n_j, n_k ...\}$ from N we substitute the values of the parameters of all the norms in the removal set, $P_i^{(N)}, P_j^{(N)}, P_k^{(N)} ...$, by their baseline counterparts, $P_i^{(bsl)}, P_j^{(bsl)}, P_k^{(bsl)} ...$. Consequently, the normative parameters of system $N' = N \setminus \{n_i, n_j, n_k ...\}$ is composed of the original parameter quantities for the non-removed norms, $P_l^{(N)}$ for $n_l \notin \{n_i, n_j, n_k ...\}$, plus the baseline parameters for the removed norms, $P_m^{(bsl)}$ for $n_m \in \{n_i, n_j, n_k ...\}$.

Now that we know how to remove subsets of norms from a normative system, we have answered question (i). Essentially, a norm is absent when the quantities of the parameters it depends upon have been substituted by their baseline counterparts. Now we are in a position to provide a straightforward answer to issue (ii). The sum in eq. (9) is taken over $N' \subseteq N \setminus \{n_i\}$, i.e. all normative systems similar to the input normative system N from which *at least* the individual norm n_i has been removed. Hence, to obtain all terms in the summation, one must first remove n_i from N by substituting its parameters by their baseline. Then, substitute the parameters linked to other norms according to all possible combinations of the remaining ones $\{n_j\}_{j \neq i}$. Finally, $N' \cup \{n_i\}$ is obtained by setting $P_i^{(bsl)}$ back to the original parameter values $P_i^{(N)}$. This final step is only performed for n_i, not for

any other norms $n_j \neq n_i$ that may also be absent from N'. Therefore, the sum in eq. (9) contains $2^{|N|-1}$ terms.

Finally, we clarify that the Shapley value includes factorial terms on the size of the original normative system $|N|$ (which is a fixed quantity across all normative systems in the same family) and to the trimmed one $|N'|$. It should be noted that $|N'|$ only counts the individual norms that have *not* been substituted by baselines, as those that have been are considered as absent.

Concerning the baseline normative system, for the time being we do not provide a systematic method to find an adequate baseline given an arbitrary normative system family. The only restriction we impose is that the baseline normative system, just as all other normative systems in the same family, has to respect the domain-dependent constraints. For our running example, luckily, the simplicity of the scenario allows to define N_{bsl} from mere intuition. However, we assert that a good choice for a baseline normative system is one such that, for any sampled path, the initial global state is kept unchanged after a sequence of transitions of arbitrary length:

$$\mathbf{s_0} \xrightarrow{N_{bsl}} \mathbf{s_0} \xrightarrow{N_{bsl}} ... \xrightarrow{N_{bsl}} \mathbf{s_0} \tag{10}$$

This is a fairly demanding requirement, since we are not referring to the expectation over a sample of random paths, but to deterministic equality over every single possible path. Provided we are able to design a baseline where the above property holds, then the alignment of the baseline normative system simply corresponds to the assessment of the initial state according to the semantics function of value v:

$$\mathsf{Algn}_{N_{bsl},v} = f_v(\mathbf{s}_0) \tag{11}$$

Running Example

In our example model, setting the baseline parameters can be manually made thanks to its simplicity:

$$N_{baseline} = \left\{ \begin{array}{l} n_1 \sim collect = [0, ..., 0] \\ n_2 \sim redistribute = [\frac{1}{5}, ..., \frac{1}{5}] \\ n_3 \sim catch = 0 \\ n_4 \sim fine = 0 \end{array} \right\} \tag{12}$$

Note that the choice of the *redistribute* list respects constraint (2). We have experimentally checked that this choice of parameters does indeed leave the initial global state of the system unchanged.

6.1 Properties of the Shapley Value

In the context of traditional cooperative game theory, it can be proven that the Shapley value is the only payoff distribution scheme that fulfils all of the following properties (Peters, 2008):

1. *Efficiency*: All of the payoff obtained by the coalition is allocated to some player.
2. *Null player*: If a player is *null*, then his Shapley value equals to zero.

3. *Symmetry*: If two players are symmetrical, then their Shapley values are equal.

4. *Additivity*: For any player, the Shapley value in the additive cooperative game equals to the sum of the Shapley values of the separate games.

Next, we review these properties in the context of normative systems, and provide definitions for these concepts within our value alignment context. Some of these properties can be readily integrated with our norm synthesis methodology and provide valuable insights. Others are not so interesting. All results are expressed in terms of alignment with respect to a single value, however they are all extensible to the alignment with respect to the aggregation of values.

First, we start with the *efficiency* property (1). A payoff distribution is efficient if all the reward achieved by the coalition is distributed back to its members. In our value alignment context, it is not material reward that we are allocating but recognition of a norm as crucial to effectively promote some value.

As formulated in Definition 6, the Shapley value for an individual norm is actually *not* efficient with respect to the absolute alignment of the normative system it belongs to:

$$\sum_{n_i \in N} \phi_i(v) \neq \mathsf{Algn}_{N,v} \tag{13}$$

However, the Shapley value of individual norms *does* maintain the efficiency property with respect to the *relative* alignment between the normative system being examined and the baseline:

Proposition 1. Given a normative system instance N, a baseline N_{bsl} for the family of normative systems to which N belongs and a value v (with semantics function f_v), the Shapley values of individual norms $n_i \in N$ are *efficient* with respect to the relative alignment between N and N_{bsl}:

$$\sum_{n_i \in N} \phi_i(v) = \mathsf{Algn}_{N,v} - \mathsf{Algn}_{N_{bsl},v} = \mathsf{RAlgn}_{N/N_{bsl},v} \tag{14}$$

The proof is provided in Appendix A.

Given the result in Proposition 1, the interpretation of the Shapley value in the normative systems context is slightly different from the one made in classical cooperative game theory. In that field, it is routinely assumed that the empty coalition does not achieve any utility. In our context, the "norm-less" situation (where all the normative parameters are set to their baseline values) may, in general, have non-zero alignment.

Therefore, the Shapley values of individual norms are not allocating credit for the absolute alignment that the normative system achieves. Instead, the Shapley value is allocating credit for alignment *relative to the baseline*. As explained previously, the baseline normative system should be set in a way as to halt the progress of the MAS under study (i.e. the initial state is kept unchanged after an arbitrary number of consecutive transitions). Consequently, the Shapley values are allocating the credit for the progress, from an ethical standpoint, from the initial state to the outcome achieved by the norms in place.

In case we wanted to have the Shapley values of individual norms be efficient with respect to the *absolute* alignment (i.e. have eq. (13) be fulfilled with an equality symbol), it

would be necessary to design a baseline such that $\mathsf{Algn}_{N_{bsl},v} = 0$. Computationally, it is an open question whether such a baseline exists given an arbitrary value semantics function, and whether its uniqueness is guaranteed. Additionally, from an interpretation perspective, we expect that the resulting baseline values for the normative parameters might be difficult to interpret and justify, since it would not reflect the lack of progress in the system, but rather the introduction of just enough regulation as to shift the system towards an ethically neutral outcome. Therefore, despite one might think that it is more desirable to have the Shapley values be efficient with respect to the absolute alignment, we believe that it is actually more informative to have them be efficient with respect to the relative alignment between N and N_{bsl}.

The second desirable property of the Shapley value that we examine is the role of *null norms* (2). The result from classical cooperative game theory is directly importable into our normative systems context. First, we define what a null norm is, in analogy to the concept of a null player:

Definition 7. A norm n_i is a *null norm* within normative system N with respect to value v if, for any $N' \subseteq N \setminus \{n_i\}$, it holds that $\mathsf{Algn}_{N' \cup \{n_i\},v} = \mathsf{Algn}_{N',v}$.

It is straightforward to prove from the above definition and eq. (9) that any null norm has zero Shapley value, $\phi_i(v) = 0$, meaning that it should not be given any credit for steering the system towards an ethically compliant state. However, the opposite is not necessarily true. A Shapley value of zero is only indicative of a null norm if the normative system it belongs to is *monotone*:

Definition 8. A normative system N is *monotone with respect to value* v (with semantics function f_v) if $\mathsf{Algn}_{N_2,v} \geq \mathsf{Algn}_{N_1,v}$, $\forall N_1, N_2 \subseteq N$ such that $N_1 \subset N_2$. (It is *strictly monotone* if $\mathsf{Algn}_{N_2,v} > \mathsf{Algn}_{N_1,v}$).

In a monotone normative system, switching any normative parameter from the baseline back to the original quantities necessarily increases the alignment. It can be interpreted as a normative system where all of its individual norms, to a greater or lesser extent, are contributing towards the alignment whenever they are reintroduced into the coalition.

In a monotone normative systems, the following holds:

Proposition 2. If normative system N is monotone with respect to value v and norm n_i has zero Shapley value with respect to v, $\phi_i(v) = 0$, then n_i is a null norm with respect to v.

The proof is provided in Appendix A.

Property (3) of the traditional Shapley value states that symmetrical players have identical Shapley values. In the context of normative systems, we have:

Definition 9. Two different norms $n_i, n_j \in N$ are *symmetric* with respect to value v if, for any $N' \subseteq N \setminus \{n_i, n_j\}$, it holds that $\mathsf{Algn}_{N' \cup \{n_i\},v} = \mathsf{Algn}_{N' \cup \{n_j\},v}$.

Indeed, norms that achieve the same level of promotion after they are separately introduced should be allocated the same amount of credit for the alignment realised:

Table 2: Shapley values for all the individual norms conforming the optimal normative systems with respect to the value for which they are optimised.

Value	Norm	Shapley value
Equality	n_1	0.50
	n_2	0.03
	n_3	0.07
	n_4	0.01
Fairness	n_1	0.19
	n_2	0.45
	n_3	0.46
	n_4	0.42
Aggregation	n_1	0.00
	n_2	0.27
	n_3	0.25
	n_4	0.31

Proposition 3. If $n_i, n_j \in N$ are two *symmetric norms* with respect to value v, then $\phi_i(v) = \phi_j(v)$.

Again, Proposition 3 is proven in Appendix A.

Finally, property (4) of the Shapley value that sets it apart from other payoff allocation schemes is the additivity property. However, this property is not as interesting as the other ones in our context, as its applicability is very limited. It could be applied in cases where aggregation functions over sets of values are defined as linear combinations of the individual values. To illustrate the limited scope of this property, we point to the running example in this paper, where we have defined an aggregation function for two values, however not as a linear combination. For this reason, we will not analyse this property in detail.

Running Example

Table 2 presents the Shapley values of every individual norm in the optimal normative systems in Table 1, with respect to the value for which they have been optimised. Additionally, for the three cases it has been experimentally checked that the efficiency property in eq. (14) holds. Those results are presented in Table 6 in Appendix A.

For value *equality*, the norm with the highest Shapley value is by far n_1, which is related to the collection of taxes. All other norms have Shapley values ~ 0, including the other norm of economic nature, n_2, linked to the redistribution of the common fund. These results reinforce our explanation for the optimal normative parameters with respect to equality in Table 1, where we conjectured that the wealth distribution is shrunk right after taxes are collected.

In contrast, for value *fairness*, the situation is the opposite, with norms n_2, n_3 and n_4 all having similar and large Shapley values, significantly above that of norm n_1. This would indicate that to punish evaders, it is most important to detect them (n_3), as unde-

tected evaders would automatically rise as the wealthiest members in the society. Then, the common fund needs to be very unevenly redistributed (see the optimal parameters in Table 1) towards the middle class, which is mostly composed by law-abiding citizens, as we have argued in Section 5. Note that imposing a fine on evaders is important, yet it is on approximately the same level as the redistribution of taxes. This indicates that, when it comes to punishing evaders, the norms that exclusively target them (n_3 and n_4) are just as relevant as the rule n_2 that directs their resources elsewhere, even at the expense of harming non-evader agents.

Last of all, the Shapley values for the optimal norms with respect to the aggregated values appear to be closer to those for fairness, with similar quantities for n_2, n_3 and n_4. Surprisingly, the norms related to tax recollection n_1, which stood out when it came to value *equality*, has Shapley value of zero for the aggregation.

6.2 Monotonicity in Low Dimensional Normative Systems

In order to assert whether the norms with zero Shapley values are actually null or not, we check the monotonic behaviour of every optimal normative system with respect to the value for which it has been optimised. To do so, we use a rather inelegant, brute-force approach. Due to the reduced number of parameters of our running example, however, this approach is still feasible, although not recommended in general.

The pseudo-code to check whether the optimal normative systems N^* are monotonic appears in Algorithm 1. Note that we only check pairs of subsets N_1, N_2 (being N_1 included within N_2, $N_1 \subset N_2$) for which N_2 includes only one more individual norm than N_1. It can be very easily proven that the check in Algorithm 1 is equivalent to a monotonic normative system:

Proposition 4. Given a normative system N and a value v (with semantics function f_v), if $\forall N' \subset N$, $\forall n \in N \setminus N'$ it holds that $\mathsf{Algn}_{N' \cup \{n\},v} \geq \mathsf{Algn}_{N',v}$ (i.e. Algorithm 1 returns `True`) iff N is monotone with respect to v.

Proof. The forward implication (Algorithm 1 returns `True` $\implies$ N monotone) can be proven by considering that consecutively adding new individual norms must at least maintain or improve the alignment, for any $N_2, N_1 \subseteq N$ such that $|N_2| > |N_1|$ it must hold that $\mathsf{Algn}_{N_2,v} \geq \mathsf{Algn}_{N_1,v}$.

Algorithm 1: Brute-force approach to check the monotonic behaviour of an optimal normative system.

```
foreach N_1 ⊂ N do
    compute Algn_{N_1,v}
    foreach n_i ∈ N \ N_1 do
        N_2 ← N_1 ∪ {n_i}
        compute Algn_{N_2,v}
        if Algn_{N_1,v} > Algn_{N_2,v} then
            return False
return True
```

The reverse implication (N monotone $\implies$ Algorithm 1 returns `True`) follows from the fact that, because N is monotone, $\forall N_1, N_2 \subset N$ such that $N_1 \subseteq N_2$ then $\mathsf{Algn}_{N_2,v} \geq \mathsf{Algn}_{N_1,v}$, this is in particular true for cases such that $|N_2| = |N_1| + 1$. □

One may think that monotonic normative systems are a promising subclass of normative systems for which optimisation with the target of maximum value alignment can be simplified. That is to say, instead of optimising for all normative parameters at once (like we do with the GA in our running example), one could start at the baseline normative system and optimise for one normative parameter at a time.

However, in order for this procedure to be correct, it would need to hold that the optimal normative system in a family with respect to the value of interest is indeed monotonic, even before such an optimal normative system has been computed. At this point, we are unable to provide such a guarantee, and hence recommend users of this methodology to stick with meta-heuristic methods. In fact, the following results concerning our running example prove that of the three optimal set of norms computed, only one is monotonic.

Running Example

Running Algorithm 1 over all the optimal normative systems (using the alignment function for the values for which they have been optimised) returns the following results: $N^*_{equality}$ is indeed monotone, while $N^*_{fairness}$ and $N^*_{aggregation}$ are not.

By the results reported in Table 2, we can assert that for value *equality*, all individual norms except n_1 (related to tax collection) are *null norms* (or close to null). This observation reinforces the dominant role that the *collect* rates have when it comes to promoting equality. By the optimal parameters in Table 1, we already hypothesised that the wealth distribution is shrunk right after taxes are collected. This conjecture seems to be supported by the finding that n_1 is the only non-null norm when it comes to equality (under its optimal normative system).

7. Value Compatibility

The Shapley values have allowed us to examine, given a value and its semantics, the relationships between the constituent norms in a normative system. In this section, we study an analogous issue: given a normative system, what is the relationship among several values that may (or not) be supported by it.

Schwartz's theory of basic human values establishes that actions executed in pursue of some value may have consequences that are either congruent or in conflict with other values (Schwartz, 2012). Such thinking can be extended to our policy design context. Implementing normative systems that aggressively promote some value might have collateral consequences, either positive or negative, for the promotion of other values. To quantify such relationships, we introduce the concept of *value compatibility* as an extension to the framework presented in Section 4.

Definition 10. Given a normative system N, a set of values $V = \{v_1, v_2, ..., v_k\}$ with semantics functions $f_1, f_2, ..., f_k$ are *compatible to degree d* (or d-compatible) under N if, for all values in V, it holds that $\mathsf{Algn}_{N,v} \geq d$.

It immediately follows that, if some set of values are compatible to degree d, they are also compatible to any degree $d' \leq d$. If a normative system is highly specialised towards some value at the expense of others, the compatibility will be low, possibly negative. Normative systems that reach a compromise will presumably maintain much higher compatibility between the values, even if not aligned with any of them to the maximum possible amount ~ 1.

The concept of compatibility works with alignment functions with respect to single values, $\mathsf{Algn}_{N,v}$. Nonetheless, we can easily expand it to aggregations over sets of values:

Definition 11. Given a set V of d-compatible values under normative system N, an aggregation function $F_V : [-1, 1]^n \to [-1, 1]$ is said to *preserve the compatibility* if $\mathsf{Algn}_{N,V} \geq d$.

Trivially, aggregation functions based on linear combinations of the alignment for the individual values do preserve d-compatibility by setting $d = \min_{v \in V} \mathsf{Algn}_{N,v}$.

In our framework, the concept of value compatibility makes sense and can potentially provide a lot of insight into the optimal normative systems obtained. However, an analogous concept for norms (i.e. norm compatibility, see e.g. the relationship among exclusive norms in Serramià et al., 2020) is not applicable in our approach. By construction, norms in a normative system control different aspects of a state transition. Intuitively, all the individual norms that make up a normative system are compatible with one another by construction.

Running Example

The compatibility computations for the optimal normative systems obtained in our tax policy model are presented in Table 3. It clearly manifests that the optimal normative system for equality is very neglectful of the fairness of the system, as it has negative alignment with respect to it. Unsurprisingly, its alignment with respect to the aggregation of values is also very poor. Hence, despite being very effective at reducing wealth inequality, under the optimal norms for equality the values *equality* and *fairness* are very incompatible.

Meanwhile, the norms optimised for fairness do uphold equality to a much larger extent than the other way around. Under this optimal normative system, the two values are compatible to degree ~ 0.6. This is a purely collateral effect, since the semantics function for fairness in eq. (7) does not encode any information related to the width of the wealth distribution. This result is visually confirmed by the second row in Figure 3. In its quest

Table 3: The optimal normative system with respect to value $\mathbf{v_i}$ (see Table 1) has its alignment computed with respect to values $\mathbf{v_j}$. For example, in the first row the optimal normative system with respect to equality is examined from the perspective of fairness and the aggregation of both values.

		$\mathbf{v_j}$		
		Equality	**Fairness**	**Aggregation**
$\mathbf{v_i}$	**Equality**	-	-0.28	-0.26
	Fairness	0.60	-	0.56
	Aggregation	0.71	0.88	-

to treat evader agents harshly, the optimal normative system for fairness also impoverishes many law-abiding citizens, most of which end up in the lower half of the wealth range. This happens because the retrieval of resources is, at least in part, done through a very uneven redistribution of taxes, a policy that affect all agents equally regardless of their evader status. Hence, the initial wealth distribution is narrowed as a consequence and the Gini index is decreased, although this is not the primary objective of the optimal norms for fairness.

Last row in Table 3 displays the alignment under the optimal norms for our aggregation function of the two values. It stands out as the one with the highest quantities: actually, the optimisation for the aggregation of the two values leads to a much higher compatibility degree than the optimisation for any of the two values separately, over 0.7. This is a consequence of our demanding aggregation function in eq. (8). The downside of this ambitious approach is that it does not preserve the compatibility as established in Definition 11, except under the optimal normative system for equality (for which the values are actually *incompatible*).

Also, the last row in Table 3 shows that the optimal normative system for the aggregation of values is more aligned with respect to fairness than equality. Similarly, the Shapley values in Table 2 for the aggregation of values are most similar to those of the optimal norms for fairness. Hence, we can conclude that the aggregation of values is definitely relying more on the promotion of fairness than equality.

7.1 Compatibility Maximising Normative System

Of course, we are most interested in finding the *maximum* compatibility degree for an arbitrary set of values. In fact, given a normative system N and a set of values V, their maximum compatibility degree is given by:

$$d_{max} = \min_{v \in V} \mathsf{Algn}_{N,v} \tag{15}$$

Equation (15) can, just as any other function, serve as an optimisation target. By finding the normative system that maximises the right-hand side of eq. (15), we obtain the normative system for which the set of values in question are the most compatible. Then, we can define the *compatibility maximising normative system* (CMNS) $N^{\ddagger}$ as:

$$N^{\ddagger} = \arg\max_{N \in \mathcal{N}} \min_{v \in V} \mathsf{Algn}_{N,v} \tag{16}$$

Note that eq. (16) is actually not an optimisation for an alignment function, as defined by the formal model in Section 4. One could think of taking a set of values $V = \{v_1, ..., v_n\}$ with semantics functions $f_1, ..., f_n$, setting their aggregation function as the minimum $F_V(\mathsf{s}) = \min_{v \in V} f_v(\mathsf{s})$, and then performing an optimisation of the type we have presented in Section 5. However, this would not be equivalent to searching the CMNS, since the expected value and the minimum operators are not in general commutable.

The notion of the CMNS echoes that of Pareto optimality (Lockwood, 2008). In the game theory literature, an outcome is *Pareto optimal* if there is no other outcome in which all participants are at least as well off, and at least one participant is better off. In this work, the CMNS is the normative system such that no other normative system can simultaneously

have larger alignment with respect to all values of interest. Therefore, Pareto optimality and compatibility maximisation are analogous, although not identical, concepts. While Pareto optimality fails if at least one participant in a game is better off (while all others retain the same utility), the CMNS fails if an alternative normative systems is encountered such that alignment with respect to *all* values is improved.

Running Example

To find the CMNS of our running example, we run the optimisation search using the same version of a GA as in Section 5, but taking eq. (16) as the target function. The set of values for which the alignment is computed (and then the minimum taken) is $V = \{equality, fairness\}$. To compute the expected values for the alignment, we again perform Monte-Carlo sampling with a sample of 500 paths of 10 transitions each.

Table 4: Optimisation results for the CMNS: optimal normative parameters and maximum compatibility degree, and alignments with respect to values *equality*, *fairness* and their aggregation.

$\mathbf{N}^{\ddagger}$	$collect = [4\%, 60\%, 74\%, 33\%, 58\%]$	$\mathbf{d}^{\ddagger}_{\mathbf{max}}$	0.74
	$redistribute = [3\%, 37\%, 35\%, 16\%, 9\%]$	$\mathsf{Algn}_{N^{\ddagger},\text{eq}}$	0.74
	$catch = 45\%$	$\mathsf{Algn}_{N^{\ddagger},\text{fair}}$	0.74
	$fine = 85\%$	$\mathsf{Algn}_{N^{\ddagger},\text{aggr}}$	0.56

The results of the optimisation for maximum alignment appear in Table 4. First, the optimal compatibility degree is at 0.74, slightly higher than the compatibility achieved under the optimal normative system for the aggregated values. Also, this compatibility degree is attained by promoting both values equally, although this requirement was not originally encoded in the objective function of the search. In summary, the CMNS does keep the compatibility degree between the two values (equality and fairness) to the largest degree among all optimal normative systems found in this work. However, the compatibility degree achieved by the optimal norms for the aggregation is only slightly below, where promotion of fairness is significantly boosted at the expense of a minor misalignment with respect to equality.

Second, we review the optimal parameters for the CMNS. The trends of the *collect* and *redistribute* parameters with the wealth segments are most similar to those of the optimal norms for the aggregation of values (see Table 1). Surprisingly, $N^{\ddagger}$ is the normative system that treats evaders the harshest. It has one of the highest *catch* rates (tied with that from the optimal normative system for fairness), and the largest *fine* across all normative systems analysed. However, despite being the most punitive norm set, its alignment for value *fairness* is actually lower than that of $N^{*}_{\text{aggregation}}$. This finding reinforces our observation that evaders are not actually punished by exclusively targeting them through n_3 and n_4, as we hypothesised in Section 5.

In Figure 3, bottom right row, the outcome that is reached under the CMNS is shown. The comparison with the optimal normative system for the aggregation of values shows that, despite some qualitative similarities in the normative parameters, the outcomes they lead to

are definitely different. Instead of reaching a wealth distribution halfway between N^*_{equality} and N^*_{fairness}, $N^\ddagger$ splits agents into two wealth groups by their wealth (above and below 50 units). Evaders all fall into the first group, hence ensuring fairness is at least moderately promoted. Meanwhile, because the two peaks, although separate, are still narrow, equality is also maintained.

8. Conclusions

In this work, we have proposed a solution to the problem of automated synthesis of normative systems based on value promotion. To do so, we have committed to a consequentialist position regarding the relationship between norms and values, and have delegated the responsibility to provide the meaning of values to the system designer. The methodology we have presented to tackle the task is fairly general and allows the designer to tailor it to the MAS at hand: choices need to be made regarding the norms in place, the semantics function of values and the optimisation strategy to perform the search.

The running example to illustrate our methodology shows how to apply each step of the methodology to a very simple model. In it, we have not introduced any reasoning schemes by the agents that would allow them to decide whether or not to abide by the norms, and which action to take among those that are designated as legal. We have omitted the introduction of individual reasoning schemes to keep the work focused on the automated synthesis of norms. However, we anticipate that the optimal normative system will depend on the composition of the society, namely the concrete reasoning schemes that are implemented and the proportion among them, analogously to the results obtained by Fagundes et al. (2016).

The results obtained in our running example for our alignment-maximising search have been very satisfactory, although the model has a reduced number of normative parameters and hence the search space is relatively small. Although it is possible to hypothesise about the role of the obtained normative parameters (i.e. with the aid of visual representations), the interpretability of the resulting normative parameters quantities is very much aided by the Shapley values of individual norms. The role that this indicator plays in our normative systems context is similar to the explainable AI literature, particularly in the estimation of feature importance in supervised machine learning models (Štrumbelj & Kononenko, 2013; Lundberg & Lee, 2017). Additionally, some of the desirable properties that give the Shapley value a privileged position in cooperative game theory have been proven to apply to the area of this work.

The last contribution of this paper is a numerical quantification in the compatibility between values. We have defined this concept formally and then performed a search for the normative system that maximises it in our running example. However, one could argue that the results for the CMNS are not much of an improvement over those obtained with a demanding aggregation function, and a "regular" optimisation using the derived alignment function as the search target.

Overall, this paper contains a substantial contribution to the field of automated synthesis of value-aligned normative systems, a field where there is not a load of work to build upon. We foresee that extensions of this work will study the scalability of our methodology to more complex models dependent upon more normative parameters, and the application

to real-life problems of policy design and analysis. Interesting future work should also attempt to integrate the synthesis methodology presented here with a framework that, prior to running the optimising search, automatically de-abstracts the value of interest into a semantics function, taking into account the domain at hand. Such an integration would allow practitioners to work with values directly as abstract entities.

Code Availability

All the code to go along with this work has been integrally developed in Python 3. It is available under an MIT license at `https://github.com/nmontesg/aamas21` and as an online appendix.

Acknowledgments

This work has been supported by the AppPhil project (RecerCaixa 2017), the CIMBVAL project (funded by the Spanish government, project #TIN2017-89758-R), the EU WeNet project (H2020 FET Proactive project #823783) and the EU TAILOR project (H2020 #952215).

Appendix A. Supplementary Information

Table 5: Hyperparameters of the Genetic Algorithm to find the optimally value-aligned normative systems. The first six refer to hyperparameters of the GA itself, and are covered in Section 5. The latter two refer to the Monte Carlo sampling: number of state transitions per path and total amount of paths sampled to compute the alignment.

Hyperparameter	**Value**
Population size	100
p (intermediate recombination)	0.25
k (elitism)	5
Maximum total iterations	500
Maximum partial iterations	50
Fitness threshold	0.9
Path length	10
Path sample size	500

Proofs of Propositions 1 to 3

In order to prove Propositions 1 to 3, we will use an equivalent definition of the Shapley value from the one in eq. (9). For a complete formulation of the Shapley value solution

Table 6: Optimal alignment found for every value, alignment of the baseline normative system and sum over the Shapley values with respect to that value. It can be checked that the sum over the Shapley values is efficient with respect to the relative alignment between the optimal normative system and the baseline, see eq. (14).

Value	$\mathsf{Algn}^{*}_{\mathbf{N},\mathbf{v}}$	$\mathsf{Algn}_{\mathbf{N_{bsl}},\mathbf{v}}$	$\sum_{\mathbf{n_i}\in\mathbf{N}^{*}} \phi_{\mathbf{i}}(\mathbf{v})$
Equality	0.95	0.34	0.61
Fairness	0.93	-0.59	1.52
Aggregation	0.66	-0.22	0.88

concept from the perspective of classical cooperative game theory, the reader is directed to Chalkiadakis et al. (2011, Ch. 2).

Given a normative system N, we denote by Π_N the set of all permutations over N. There are $|N|!$ permutations in total. For a permutation $\pi \in \Pi_N$, we denote by $S_\pi(n_i)$ the set of predecessors of norm n_i in permutation π. For example, in the social model considered throughout the paper, we have $N = \{n_1, n_2, n_3, n_4\}$. A permutation is $\pi = (n_1, n_3, n_4, n_2)$. Then, $S_\pi(n_1) = \{\}$, $S_\pi(n_2) = \{n_1, n_3, n_4\}$, $S_\pi(n_3) = \{n_1\}$ and $S_\pi(n_4) = \{n_1, n_3\}$.

The marginal contribution of norm n_i to the alignment for value v in permutation π is defined as:

$$\Delta^{v}_{\pi}(n_i) = \mathsf{Algn}_{S_\pi(n_i)\cup\{n_i\},v} - \mathsf{Algn}_{S_\pi(n_i),v} = \mathsf{RAlgn}_{S_\pi(n_i)\cup\{n_i\}/S_\pi(n_i),v} \tag{17}$$

Δ^{v}_{π} measures the improvement in the alignment (with respect to value v) when norm n_i joins its predecessors in permutation π.

Then, the Shapley value for norm n_i can be defined as the average over all permutation of its marginal contribution to the alignment:

$$\phi_i(v) = \frac{1}{|N|!} \sum_{\pi\in\Pi_N} \Delta^{v}_{\pi}(n_i) \tag{18}$$

Proof of Proposition 1. π_j denotes the norm in the j-th position in permutation π.

$$\begin{aligned}
\sum_{n_i \in N} \phi_i(v) &= \sum_{n_i \in N} \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \Delta_\pi^v(n_i) = \\
&= \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \sum_{n_i \in N} \Delta_\pi^v(n_i) = \\
&= \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \mathsf{Algn}_{\{\pi_1\},v} - \mathsf{Algn}_{\{\},v} + \mathsf{Algn}_{\{\pi_1,\pi_2\},v} - \mathsf{Algn}_{\{\pi_1\},v} + \\
&\qquad + \ldots + \mathsf{Algn}_{\{\pi_1,\ldots,\pi_n\},v} - \mathsf{Algn}_{\{\pi_1,\ldots,\pi_{n-1}\},v} = \\
&= \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \mathsf{Algn}_{\{\pi_1,\ldots,\pi_n\},v} - \mathsf{Algn}_{\{\},v} = \\
&= \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \mathsf{Algn}_{N,v} - \mathsf{Algn}_{N_{bsl},v} = \\
&= \frac{1}{|N|!} |N|! \left(\mathsf{Algn}_{N,v} - \mathsf{Algn}_{N_{bsl},v}\right) = \mathsf{RAlgn}_{N/N_{bsl},v}
\end{aligned}$$

□

Proof of Proposition 2. Because N is monotone (see Definition 8), it must be that $\Delta_\pi^v(n_i) \geq 0,\ \forall \pi \in \Pi_n, \forall n_i \in N$.

Then, if a norm has $\phi_i(v) = \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \Delta_\pi^v(n_i) = 0$, it must be that every term equals zero, $\Delta_\pi^v(n_i) = 0,\ \forall \pi \in \Pi_N$. This is equivalent to having $\mathsf{Algn}_{N' \cup \{n_i\},v} = \mathsf{Algn}_{N',v},\ \forall N' \subseteq N \backslash \{n_i\}$, as every $N' \subseteq N \backslash \{n_i\}$ can be identified with $S_\pi(n_i)$ for some $\pi \in \Pi_N$. This derivation corresponds to n_i being a null norm (see Definition 7). □

Proof of Proposition 3. Suppose $n_i, n_j \in N$ are symmetric norms (see Definition 9). Given a permutation π, we denote by π' the permutation that is obtained from π by swapping n_i and n_j. First, we prove that $\Delta_\pi^v(n_i) = \Delta_{\pi'}^v(n_j)$.

Suppose n_i precedes n_j in π. Then, $S_\pi(n_i) = S_{\pi'}(n_j) = N'$:

$$\begin{aligned}
\Delta_\pi^v(n_i) &= \mathsf{Algn}_{N' \cup \{n_i\},v} - \mathsf{Algn}_{N',v} \\
\Delta_{\pi'}^v(n_j) &= \mathsf{Algn}_{N' \cup \{n_j\},v} - \mathsf{Algn}_{N',v}
\end{aligned}$$

Because n_i and n_j are symmetric, it holds that $\mathsf{Algn}_{N' \cup \{n_i\},v} = \mathsf{Algn}_{N' \cup \{n_j\},v}$. Consequently, $\Delta_\pi^v(n_i) = \Delta_{\pi'}^v(n_j)$.

Now suppose that n_i goes after n_j in π. Now, $N' = S_\pi(n_i) \backslash \{n_j\}$:

$$\begin{aligned}
\Delta_\pi^v(n_i) &= \mathsf{Algn}_{N' \cup \{n_j\} \cup \{n_i\},v} - \mathsf{Algn}_{N' \cup \{n_j\},v} \\
\Delta_{\pi'}^v(n_j) &= \mathsf{Algn}_{N' \cup \{n_i\} \cup \{n_j\},v} - \mathsf{Algn}_{N' \cup \{n_i\},v}
\end{aligned}$$

Again, because n_i and n_j are symmetric, the second terms of the right-hand sides are equal, and $\Delta_\pi^v(n_i) = \Delta_{\pi'}^v(n_j)$.

Then, because the map between π and its swapped permutation π' is one-to-one, we have:

$$\phi_i(v) = \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \Delta^v_\pi(n_i) = \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \Delta^v_{\pi'}(n_v) = \phi_j(v)$$

□

References

Ajmeri, N., Guo, H., Murukannaiah, P. K., & Singh, M. P. (2020). Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, p. 16–24, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Atkinson, K., & Bench-Capon, T. (2016). States, goals and values: Revisiting practical reasoning. *Argument & Computation*, *7*(2-3), 135–154.

Baluja, S., & Caruana, R. (1995). Removing the genetics from the standard genetic algorithm. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, p. 38–46, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Bench-Capon, T., & Modgil, S. (2017). Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law*, *25*(1), 29–64.

Chalkiadakis, G., Elkind, E., & Wooldridge, M. (2011). Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *5*(6), 1–168.

Conte, R., & Castelfranchi, C. (1999). From conventions to prescription. towards an integrated view of norms. *Artificial Intelligence and Law*, *7*(4), 323–340.

Fagundes, M. S., Ossowski, S., Cerquides, J., & Noriega, P. (2016). Design and evaluation of norm-aware agents based on normative markov decision processes. *International Journal of Approximate Reasoning*, *78*, 33–61.

Feather, N. T. (1995). Values, valences, and choice: The influences of values on the perceived attractiveness and choice of alternatives.. *Journal of Personality and Social Psychology*, *68*(6), 1135–1151.

Gini, C. (1912). *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. Facoltà di Giurisprudenza della R. Università di Cagliari.

Gorrieri, R. (2017). Labeled transition systems. In *Monographs in Theoretical Computer Science. An EATCS Series*, pp. 15–34. Springer International Publishing.

Grossi, D., Tummolini, L., & Turrini, P. (2012). *Norms in Game Theory*, chap. Chapter 12, pp. 191–197. No. 8 in Law, Governance and Technology. Springer, Dordrecht.

Lemieux, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer New York.

Lockwood, B. (2008). Pareto efficiency. In *The New Palgrave Dictionary of Economics*, pp. 1–5. Palgrave Macmillan UK.

Luke, S. (2013). *Essentials of Metaheuristics* (second edition). Lulu.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Macintyre, A. (1998). *A Short History of Ethics: A History of Moral Philosophy from the Homeric Age to the Twentieth Century*. University of Notre Dame Press.

Miller, B. L., & Goldberg, D. E. (1995). Genetic algorithms, tournament selection, and the effects of noise. *Complex Syst.*, *9*.

Montes, N., & Sierra, C. (2021). Value-guided synthesis of parametric normative systems. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '21, p. 907–915, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems. (Best paper award finalist).

Morales, J., López-Sanchez, M., Rodríguez-Aguilar, J. A., Wooldridge, M., & Vasconcelos, W. (2013). Automated synthesis of normative systems. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '13, p. 483–490, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Morales, J., Wooldridge, M., Rodríguez-Aguilar, J. A., & López-Sánchez, M. (2018). Off-line synthesis of evolutionarily stable normative systems. *Autonomous Agents and Multi-Agent Systems*, *32*(5), 635–671.

Morris-Martin, A., Vos, M. D., & Padget, J. (2019). Norm emergence in multiagent systems: a viewpoint paper. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 706–749.

Mühlenbein, H., & Schlierkamp-Voosen, D. (1993). Predictive models for the breeder genetic algorithm i. continuous parameter optimization. *Evolutionary Computation*, *1*(1), 25–49.

Onn, S., & Tennenholtz, M. (1997). Determination of social laws for multi-agent mobilization. *Artificial Intelligence*, *95*(1), 155–167.

Peters, H. (2008). The shapley value. In *Game Theory*, pp. 241–258. Springer Berlin Heidelberg.

Rokeach, M. (1972). *The nature of human values*. Free Press.

Ruder, S. (2016). An overview of gradient descent optimization algorithms..

Sandholm, W. H. (2009). Evolutionary game theory. In *Encyclopedia of Complexity and Systems Science*, pp. 3176–3205. Springer New York.

Savarimuthu, B. T. R., & Cranefield, S. (2011). Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems*, *7*(1), 21–54.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*, pp. 1–65. Elsevier.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*(1).

Serramià, M., López-Sánchez, M., & Rodríguez-Aguilar, J. A. (2020). A qualitative approach to composing value-aligned norm systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, p. 1233–1241, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Serramià, M., López-Sánchez, M., Rodríguez-Aguilar, J. A., Morales, J., Wooldridge, M., & Ansotegui, C. (2018). Exploiting moral values to choose the right norms. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Shapley, L. S. (1951). *Notes on the N-Person Game - II: The Value of an N-Person Game*. RAND Corporation, Santa Monica, CA.

Shoham, Y., & Tennenholtz, M. (1995). On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, *73*(1-2), 231–252.

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

Spates, J. L. (1983). The sociology of values. *Annual Review of Sociology*, *9*(1), 27–49.

Štrumbelj, E., & Kononenko, I. (2013). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665.

Szabo, J., Such, J. M., & Criado, N. (2020). Understanding the role of values and norms in practical reasoning. In Bassiliades, N., Chalkiadakis, G., & de Jonge, D. (Eds.), *Multi-Agent Systems and Agreement Technologies*, pp. 431–439, Cham. Springer International Publishing.

Teze, J. C. L., Perelló-Moragues, A., Godo, L., & Noriega, P. (2019). Practical reasoning using values: an argumentative approach based on a hierarchy of values. *Annals of Mathematics and Artificial Intelligence*, *87*(3), 293–319.

The World Bank, Development Research Group (2019). Gini index (world bank estimate, 1967-2019).. Accessed 7th June 2021, `http://data.worldbank.org/indicator/SI.POV.GINI`.

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385–409.

van der Weide, T. L., Dignum, F., Meyer, J. J. C., Prakken, H., & Vreeswijk, G. A. W. (2010). Practical reasoning using values. In *Lecture Notes in Computer Science*, pp. 79–93. Springer Berlin Heidelberg.

Visser, S., Thangarajah, J., Harland, J., & Dignum, F. (2015). Preference-based reasoning in BDI agent systems. *Autonomous Agents and Multi-Agent Systems*, *30*(2), 291–330.