Encoding Ethics to Compute Value-Aligned Norms.

Minds and Machines (2023) 33:761–790
https://doi.org/10.1007/s11023-023-09649-7

Encoding Ethics to Compute Value-Aligned Norms

Marc Serramia[1] · Manel Rodriguez-Soto[2] · Maite Lopez-Sanchez[3] · Juan A. Rodriguez-Aguilar[2] · Filippo Bistaffa[2] · Paula Boddington[4] · Michael Wooldridge[5] · Carlos Ansotegui[6]

Received: 6 March 2023 / Accepted: 20 September 2023 / Published online: 22 November 2023

Abstract

Norms have been widely enacted in human and agent societies to regulate individuals' actions. However, although legislators may have ethics in mind when establishing norms, moral values are only sometimes explicitly considered. This paper advances the state of the art by providing a method for selecting the norms to enact within a society that best aligns with the moral values of such a society. Our approach to aligning norms and values is grounded in the ethics literature. Specifically, from the literature's study of the relations between norms, actions, and values, we formally define how actions and values relate through the so-called *value judgment function* and how norms and values relate through the so-called *norm promotion function*. We show that both functions provide the means to compute value alignment for a set of norms. Moreover, we detail how to cast our decision-making problem as an optimisation problem: finding the norms that maximise value alignment. We also show how to solve our problem using off-the-shelf optimisation tools. Finally, we illustrate our approach with a specific case study on the European Value Study.

Keywords Ethics and AI · Moral values · Decision support · Norms · Optimisation

✉ Marc Serramia
marc.serramia-amoros@city.ac.uk

[1] Department of Computer Science, City, University of London, Northampton Square, EC1V 0HB London, United Kingdom

[2] Artificial Intelligence Research Institute, Carrer de Can Planas, 08193 Bellaterra, Spain

[3] Department of Mathematics and Computer Science, University of Barcelona, Gran via de les Corts Catalanes, 585, 08007 Barcelona, Spain

[4] Geller Institute of Aging and Memory, School of Biomedical Sciences, University of West London, St Mary's Road, London W5 5RF, UK

[5] Department of Computer Science, University of Oxford, 7 Parks Rd, Oxfrod OX1 3QG, UK

[6] Department of Computing and Industrial Engineering, University of Lleida, Carrer de Jaume II, 69, Lleida 25001, Spain

1 Introduction

Norms have been extensively established in human societies to regulate societies. Some norms take a legal stance (Nadelmann, 1990), whereas others consider their social dimension (Bicchieri, 2005). Norms can be applied to many contexts—such as, e.g., security in organisations (Yazdanmehr & Wang, 2016; Grimes & Marquardson, 2019) or common resource allocation (Sethi & Somanathan, 1996)—as well as to facilitate decision support (Meinard & Cailloux, 2020; Keller & Savarimuthu, 2017). Alternative societies, such as those formed by software agents (i.e., Multi-Agent Systems), also use norms as coordination mechanisms (Fitoussi & Tennenholtz, 2000; Savarimuthu et al., 2013; Campos et al., 2013; Morales et al., 2018), and research findings in this area can be naturally extrapolated to a variety of societies.

When regulating societies –be they human or multi-agent– it is essential to acknowledge that actions carry ethical implications. Along with the Machine Ethics literature (Anderson & Anderson, 2011; Tolmeijer et al., 2021; Bostrom & Yudkowsky, 2011; Svegliato et al., 2021), there are different initiatives considering these ethical implications and advocating for beneficial and trustworthy Artificial Intelligence (Russell et al., 2015; Chatila et al., 2021). For instance, the European Commission has proposed both *Ethics Guidelines for Trustworthy AI* (European Comission, 2019) and the *Artificial Intelligence Act* (European Comission, 2023) Additionally, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE Standards Association, 2016), with a committee devoted to "Embedding Values into Autonomous Intelligent Systems", considers moral values as a first-class criterion. Along this line, in this paper, we take the stance of normative sociotechnical systems (Singh, 2014) and advocate for constraining individuals' behaviour by selecting the norms to enact in a society (/sociotechnical system) that are value-aligned. By following this approach, we adopt the perspective of a policy maker (/system designer). Therefore, computing the value alignment of norms appears as a challenging problem that we must address to assist the policymaker.

Value alignment in autonomous systems has often been addressed from a Reinforcement Learning perspective (Abel et al., 2016; Noothigattu et al., 2019; Rodriguez-Soto et al., 2020). In particular, Inverse Reinforcement Learning (Abbeel & Ng, 2004; Hadfield-Menell et al., 2016) was proposed for learning values by observing human behaviour. Alternatively, some Multi-Agent Systems literature also covers the ethical dimensions of norms . The usual approach is to consider the existence of a value system composed of moral values and a relationship between norms and values, the so-called promotion and demotion functions (Bench-Capon & Atkinson, 2009; Atkinson et al., 2006; Luo et al., 2017; Lopez-Sanchez et al., 2017; Serramia et al., 2018). These functions express whether a given norm promotes or demotes a given value and, eventually, the degree of promotion or demotion. Thus, a norm promotion function encodes the *value alignment* of each norm. Hence, it can be further employed to compute the value alignment of a *normative system* (i.e., the set of norms to enact in a society (Serramia

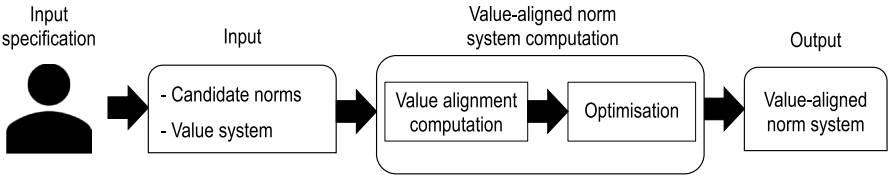

Fig. 1 The value-aligned norm system engineering process

et al., 2018)). Although norm promotion functions are often used , their mathematical formalisation is typically overlooked, hindering their usage in computing value alignment. As an exception, we highlight the proposal in Sierra et al. (2019). Sierra et al. consider values as preferences over states of the world to, later on, assess the value alignment of a norm in terms of the preference increase for those state transitions affected by the norm.

To tackle the computation of value alignment for norms, we can resort to existing literature. Both Sociology and Psychology have long studied human values, how they shape behaviour, as well as their relative importance across individuals and societies. Different value models have been proposed and compared (Cheng & Fleischmann, 2010), rendering Schwartz's value model (Schwartz, 2012) as the most comprehensive one (Hanel et al., 2018). Additionally, the ethics literature has specifically studied the relationship between norms and values.

Indeed, in ethics, typically, a norm is considered to promote a moral value depending on how it regulates an action and how this action is considered concerning the moral value (Urmson, 1958; Hansson & Hendricks, 2018). Therefore, the ethics literature counts on the means to set the foundations for a mathematical definition of such promotion function and, ultimately, of value alignment for a norm system. It is worth noticing that henceforth we use the terms moral and ethical interchangeably (without differentiation) as it is common practice in the Philosophy literature (Frankena, 1973; Audi, 1999; Fieser & Dowden, 2020).

Against this background, we provide tools for a decision-maker to compute the norm system –out of a set of candidate norms– that best aligns with a given value system.

As Fig. 1 shows, this computation process is decomposed into two main steps: firstly, we compute the value alignment of all candidate norms; and secondly, we apply an optimisation process that chooses the subset of norms that are most value-aligned (i.e., the value-aligned norm system). Based on the ethics and AI literature (Dignum, 2017), we propose a mathematical methodology for the value alignment computation that is based on the formal definition of a *norm promotion function*, which computes the value alignment of a norm system; and a *value judgement function*, which characterises values and formalises their relationship to actions.

The paper is structured as follows. Section 2 delves into the literature to learn how norms and values relate.

Subsequently, Sect. 3 studies the fundamental relationships between norms and characterises norm systems. From these basic concepts, Sect. 4 introduces the mathematical definitions of the so-called value judgement function and of value system.

Then, Sect. 5 specifies norm promotion functions, which characterise how norms promote moral values. Thereafter, Sect. 6 defines the problem of computing value-aligned norm systems, and Sect. 7 describes how we solve this problem with optimisation techniques. Next, Sect. 8 illustrates our method in the context of the European Values Study (EVS, 2021) . Finally, Sect. 9 provides related work, and Sect. 10 draws conclusions and sets paths to future research.

2 Analysing the Relationship Between Norms and Values

The relationship between the norms enacted in a society and the values that this society is aligned with is one of the main research subjects in ethics (Hansson, 2001; McNamara, 2011). Within ethics, moral values (also called ethical principles) express the moral objectives *worth striving for* van de Poel & Royakkers (2011, p.72)[1]. Examples of human values[2] include fairness, respect, freedom, security, and prosperity (Cheng & Fleischmann, 2010). Every ethical theory considers one or more moral values that should guide our behaviour (Cooper, 1993). From these considered values, an ethical theory can prescribe a series of norms as means to realise them van de Poel & Royakkers (2011, Sect. 3.4–3.6).

Moreover, since norms regulate actions, we need to judge actions ethically to determine which norms to prescribe. Therefore, it is argued in Cointe et al. (2016); Cooper (1993); Hansson and Hendricks (2018) that the central theme that unites norms and values is the moral consideration (judgement) of actions. Specifically, an action can be judged as either good or bad to perform (or skip) for a given moral value (Chisholm, 1963). This relationship between norms and values being influenced by actions implies that if a society considers an action to be good to perform from the standpoint of a given moral value, then any norm permitting or obligating such an action would be considered a norm that *promotes* that value (Cooper, 1993; Hansson & Hendricks, 2018). Contrarily, a norm prohibiting the same action would *demote* that moral value. Classically, a norm is considered to promote a moral value depending on how it regulates an action and how this action is considered concerning the moral value (Urmson, 1958; Hansson & Hendricks, 2018): (i) Obligation (if the action is good to perform and bad to skip); (ii) Permission (if the action is good to perform); (iii) Prohibition (if the action is good to skip and bad to perform).

It is clear then that to assess the value alignment of a norm system; we must consider not only the relationship between norms and values but also the ethical dimension of the actions being regulated. Figure 2 depicts the relationships that we have identified between norms, values, and actions: (i) Norms *regulate* actions; (ii) Moral values *judge* actions; and (iii) Norms *promote/demote* moral values. Figure 2 offers

[1] Moral values are often very high ideals that can seldom be achieved perfectly, though this does not preclude us from pursuing them.

[2] As previously mentioned in the introduction, Sociology and Psychology have also extensively studied human values, which are often defined as abstract ideals that guide people's behaviour (Schwartz, 2012) or idealized standards with an "ought" character (Maio, 2016).

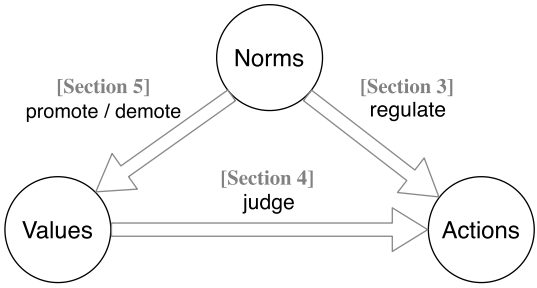

Fig. 2 Relationship between norms, values and actions

a very similar structure to the diagrams in Cooper (1993); Hansson and Hendricks (2018), showing a relationship between values, norms, and actions.

Formalising these relationships (regulation, judgement, and promotion/demotion) will provide the foundations for a mathematical definition of value alignment for a norm system. Thus, action regulations are formalised in Sect. 3.2, action judgements are formalised in Sect. 4, and finally, value promotion is formalised in Sect. 5. Based on that, we will introduce our notion of value alignment for a norm system in Sect. 6.

3 Formalising Actions, Norms, and Norm Systems

In this section, we focus on formalising the notions of *action domain*, the actions to regulate, and *normative domain*, the norm space to consider to compose a norm system. Furthermore, we characterise the desirable features of the norm system resulting from the norm engineering process outlined in Fig. 1.

Figure 2 shows that norms and actions are related through the regulation relation. We formalise the normative domain based on a given action domain. Thus its norms will regulate the actions in the action domain.

3.1 The Action Domain

We consider a finite set of available actions A that an agent can perform. An agent can perform an action in A provided that the conditions of the state of the world uphold. To encode the state of the world, we consider a propositional language $\mathcal{L}$ (with propositions in P and the logical operator "and"). Thus, we refer to a *context* of an action as a subset of the propositions of the language $\varphi \subseteq P$ describing the conditions that must hold for the actions to be feasible. Propositions in a context are connected with *and* semantics, and hence they must all hold.

The norms we consider in this paper will aim at regulating actions considering the context wherein they can be undertaken. This will define the so-called *action domain* for norms:

Definition 1 (Action domain) An action domain is a set $\mathbb{A} \subseteq \mathcal{P}(P) \times A$ of pairs so that each action is related to the set of propositions that make the action feasible.

Example 1 To illustrate the concepts that we introduce in this paper, we use a running example of the public civility game (Rodriguez-Soto et al., 2020). This game considers agents navigating between two points. The agents may encounter garbage blocking their way and have to decide how to deal with it (kick it out of the way or clean it by carrying it to a bin) considering the implications this may have. Thus, in this scenario, we consider a propositional language $\mathcal{L}$ with propositions $\mathcal{P} = \{see\_garbage,\ garbage\_in\_front\}$ and actions $A = \{kick,\ clean\}$. Then, we define an action domain $\mathbb{A} = \{kg, cg, ca\}$, where:

- $kg = (\{garbage\_in\_front\}, kick)$ corresponds to the action of kicking the garbage out of the way if the agent finds garbage in front.
- $cg = (\{garbage\_in\_front\}, clean)$ represents the action of cleaning the garbage if the agent finds garbage in front.
- $ca = (\{see\_garbage\}, clean)$ is the action of cleaning all garbage that the agent can see (both if it blocks its way or not).

3.2 The Normative Domain

We use norms to regulate the actions that must, should, or must not be performed. Here our notion of norm is based on a simplification of the one in Morales et al. (2015). Our notion of norm establishes obligations, permissions, and/or prohibitions (Meyer & Wieringa, 1993) of agent's actions. Formally:

Definition 2 (*Norm*) Given an action domain $\mathbb{A} \subseteq \mathcal{P}(P) \times A$, a norm is a pair $\langle \varphi, \theta(a) \rangle$, where: (i) $(\varphi, a) \in \mathbb{A}$, with a being an action and φ the precondition that must hold for a to be feasible; and (ii) $\theta \in \{Obl, Per, Prh\}$ is a deontic operator.

Example 2 From the action domain in Example 1, we consider the following candidate norms:

- $Per(kg) = \langle \{garbage\_in\_front\}, Per(kick) \rangle$. The norm permitting to kick garbage out of the agent's way.
- $Obl(cg) = \langle \{garbage\_in\_front\}, Obl(clean) \rangle$. The norm obliging agents to clean garbage in their way
- $Obl(ca) = \langle \{see\_garbage\}, Obl(clean) \rangle$. The norm obliging agents to clean all garbage they see.

We rely on the decision maker to provide *candidate* norms to consider for regulation, namely, the space of norms to explore when looking for a norm system. We assume that norms at hand are considered beneficial (for example, because they align with the decision-maker's goal). We aim to select the best norms out of these candidate norms. Furthermore, as argued in Serramia et al. (2018), their relations must also be factored in when selecting norms for regulation. Norm relations have been previously studied in the literature. Thus, for example, Grossi and Dignum (2005) studies the relationship between abstract and concrete norms,

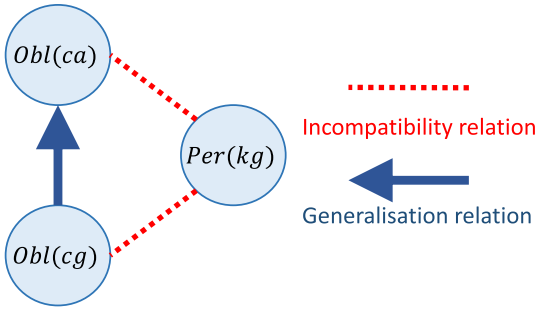

Fig. 3 Example of a normative domain for the public civility game

whereas (Kollingbaum et al., 2006; Vasconcelos et al., 2009) focus on norm conflicts —and solve them based on first-order unification and constraint solving techniques. Here, we borrow the *exclusivity* and *generalisation* norm relations already identified in Serramia et al. (2018). Informally, two norms are mutually exclusive when they are incompatible; and a norm is more general than another one when it subsumes its regulation (has broader regulation scope).

Let N denote a non-empty set of norms. On the one hand, the *exclusivity* relation is an irreflexive, and symmetric binary relation $\mathfrak{R}_i \subseteq N \times N$. If $(n_i, n_j) \in \mathfrak{R}_i$ we say that n_i, n_j are incompatible or mutually exclusive. On the other hand, the *generalisation* relation is an irreflexive, antisymmetric, and transitive binary relation $\mathfrak{R}_g \subseteq N \times N$. If $(n_i, n_j) \in \mathfrak{R}_g$, we say that n_i is generalised by n_j (i.e., it is more specific than n_j).

Once formalised norms and their relations, we define the structure that encodes them, characterising the norm space of a decision maker.

Definition 3 (*Normative domain*) A normative domain is a tuple $\langle \mathbb{A}, N, \mathfrak{R} \rangle$ such that: $\mathbb{A}$ is an action domain; N is a set of candidate norms regulating actions in $\mathbb{A}$; and $\mathfrak{R} = \{\mathfrak{R}_i, \mathfrak{R}_g\}$ is a set of norm relations over N.

Example 3 We build an example of a normative domain (see Fig. 3) with the action domain in Example 1 and the set of candidate norms in Example 2.

3.3 Characterising Norm Systems

The goal of the process depicted in Fig. 1 is to obtain a norm system. We will refer to any subset of the norms in a normative domain as a *norm system*. Since norm systems are just subsets of candidate norms, a norm system can contain incompatible norms or redundant norms (due to exclusivity or generalisation relationships). Thus, following (Serramia et al., 2018), when selecting norms, we require that the resulting norm system is sound, meaning that it does not contain incompatible nor redundant norms.

Definition 4 (*Sound norm system*) Given a normative domain $\langle \mathbb{A}, N, \mathfrak{R} \rangle$, we say that a norm system $\Omega \subseteq N$ is sound iff for each $n_i, n_j \in \Omega$, $(n_i, n_j) \notin \mathfrak{R}_i$, and n_j is not a successor[3] of n_i ($n_j \notin S(n_i)$) or vice versa ($n_i \notin S(n_j)$).

Example 4 Considering the normative domain of Fig. 3, only the norm systems with a single norm ($\{Per(kg)\}$, $\{Obl(cg)\}$, and $\{Obl(ca)\}$) are sound. Notice, for instance, that $\{Per(kg), Obl(cg)\}$ is not sound because norms are exclusive, and $\{Obl(ca), Obl(cg)\}$ is neither sound because, since $Obl(ca)$ generalises $Obl(cg)$, they are redundant.

To be more precise, the goal of the process in Fig. 1 is to yield a particular type of sound norm system, namely one that is aligned with the moral values specified by the decision maker.

4 Value-Based Judgement of Actions

As introduced in Sect. 2, ethics is the branch of philosophy that reflects on what is moral[4], right, or good (Frankena, 1973; Audi, 1999; van de Poel & Royakkers, 2011). The philosophical discipline of ethics is eminently practical because we do not want to know what is good or bad out of mere curiosity but because we want to know what we ought to do Cooper (1993); Wallach and Allen (2008). To guide us in that matter, the field of normative ethics prescribes the correct action to do in any given situation [40]. Of course, we can only provide guidelines on how to do good if we first define what is good, to begin with. That, and other foundational problems, are the subject of the field of metaethics, which attempts to clarify the ethical methodology and terminology (Beauchamp & Childress, 2009).

Within ethics, moral values (also called ethical principles) bridge normative ethics and metaethics. In the AI literature, values are seen as criteria[5] to discern which actions are right from wrong (Charisi et al., 2017; Dignum, 2017). Examples of values include justice, happiness, and autonomy. We formally characterise moral values following these informal definitions. As shown in Fig. 2, values and actions are related. Specifically, a value judges the extent to which actions' performance (or non-performance) is beneficial or detrimental. Thus, we formally characterise moral values through their judgement of actions as follows.

Definition 5 (Moral value) We characterise a moral value through a pair of value judgement functions $v = \langle \alpha_v^+, \alpha_v^- \rangle$. Given a set of actions $\mathbb{A}$, each of these functions takes an action and returns its evaluation $\alpha_v^+, \alpha_v^- : \mathbb{A} \to [-1, 1]$. Function α_v^+

[3] Formally, the successors of n are $S(n) = \{n', (n', n) \in \mathfrak{R}_g\} \cup \{n', \exists n_1, \ldots, n_r \text{ s.t. } (n', n_1), (n_1, n_2), \ldots (n_r, n) \in \mathfrak{R}_g\}$

[4] Morality here refers to the codes of conduct that, given some conditions, would be adopted by all rational people (Gert & Gert, 2020).

[5] Notice that here criteria can be assimilated to ethical principle.

evaluates the praiseworthiness of performing the action, while $\alpha_v^-(a)$ evaluates the praiseworthiness of not performing the action[6]. These evaluations are real numbers in the interval $[-1, 1]$: a positive number stands for moral value promotion, whereas a negative one stands for demotion. We require that an action cannot be praiseworthy (or blameworthy) both to perform and to skip concerning the same moral value. Thus, for a moral value to be well-defined, its value judgement functions have to satisfy the following:

$$\alpha_v^+(a) \cdot \alpha_v^-(a) \leq 0, \forall a \in A \tag{1}$$

Value judgement functions allow us to quantify the moral praiseworthiness of performing/skipping actions. Note that the condition in Equation 1 dictates that if an action is praiseworthy to perform, it must be either blameworthy or neutral to skip. Similarly, if the action is blameworthy to perform, it must be praiseworthy or neutral to skip. Overall, these value judgement functions within our characterisation of moral values allow us to adhere to previous literature (Charisi et al., 2017; Dignum, 2017) and use moral values as criteria for discerning right (praiseworthiness) from wrong (blameworthiness). Values can thus be considered as being more general than norms, as values are abstract criteria that judge actions. In contrast, norms regulate actions within specific contexts through a fixed syntax that uses particular deontic operators.

Example 5 We judge the actions of the action domain in Example 1 with respect to two values: civility $Civ = \langle \alpha_{civ}^+, \alpha_{civ}^- \rangle$; and timeliness[7] $Tim = \langle \alpha_{tim}^+, \alpha_{tim}^- \rangle$. In terms of civility, the action of cleaning garbage is highly praiseworthy to perform, but neutral to skip since the garbage is not the agent's property. In terms of timeliness though, the action is slightly blameworthy to perform as it will take time to clean and slightly praiseworthy to skip. Thus, the judgement functions of both moral values for the action *cg* may, for instance, be as follows:

$$\alpha_{civ}^+(cg) = 0.8 \quad \alpha_{civ}^-(cg) = 0 \quad \alpha_{tim}^+(cg) = -0.5 \quad \alpha_{tim}^-(cg) = 0.5$$

The judgements for action *ca* will be similar to those of *cg* but more extreme, since this action applies to all garbage, not only to the one in front of the agents. Thus:

$$\alpha_{civ}^+(ca) = 1 \quad \alpha_{civ}^-(ca) = 0 \quad \alpha_{tim}^+(ca) = -1 \quad \alpha_{tim}^-(ca) = 1$$

Regarding civility, kicking the garbage aside is blameworthy to perform and praiseworthy to skip as it could spill rubbish everywhere, block other agent's path or even harm them. On the other hand, kicking the garbage aside saves time because it rapidly frees the agent's path towards the target. Thus, this action is highly praiseworthy

[6] Note that, $\alpha_v^+(a)$ and $\alpha_v^-(a)$ are independent so $\alpha_v^-(a)$ is not necessarily equal to $-\alpha_v^+(a)$.

[7] Timeliness can be assimilated into the moral value of achievement, defined by Schwartz (2012) as being related to competence and personal success.

to perform and highly blameworthy to skip in terms of timeliness. Therefore, the judgement functions for *kg* could be defined as:

$$\alpha^{+}_{civ}(kg) = -1 \quad \alpha^{-}_{civ}(kg) = 1 \quad \alpha^{+}_{tim}(kg) = 1 \quad \alpha^{-}_{tim}(kg) = -1$$

Ethical reasoning typically involves not a single moral value but multiple moral values and value preferences (Bench-Capon & Atkinson, 2009; Luo et al., 2017; Serramia et al., 2018) conforming to a *value system*. Value systems can be individual or shared by a society. Although values could be socially agreed/negotiated (Aydoğan et al., 2021) or computationally aggregated (Lera-Leri et al., 2022), i n this work, we assume that we know the society's value system in order to select norms accordingly. Recall that in Fig. 1 we considered a value system as one of the two main inputs of our value-aligned norm system engineering process. As depicted in Fig. 2, values judge actions via their judgement functions. As these value judgement functions characterise moral values, they also implicitly constitute an integral part of the value system, which is explicitly composed of the values and their preferences.

Definition 6 (Value system) A value system is a tuple $\langle V, \succeq \rangle$, where: V stands for a non-empty set of moral values, and $\succeq$ is a ranking[8] over the moral values in V. If $v \succeq v'$ we say that v is more preferred than v', and if also $v' \succeq v$ we say that v and v' are indifferently preferred, and note it as $v \sim v'$.

Notice that, unlike our value system, the definitions in Bench-Capon and Atkinson (2009) and Serramia et al. (2018) do not consider the link between values and actions. Moreover, although (Luo et al., 2017) considers the relation between actions and values, it does not quantify it. Furthermore, in terms of the ordering structure used, we favour rankings as they are more flexible than the total orders used in Luo et al. (2017) and Bench-Capon and Atkinson (2009), though they are stricter than the partial order used in Serramia et al. (2018). We do so because partial orders would require us to make arbitrary assumptions when values are unrelated (in the order).

Example 6 The values of *civility* and *timeliness*, $V = \{Civ, Tim\}$, together with the ranking $Civ \succeq Tim$, constitute a value system.

5 Promotion of Moral Values Through Norms

Once established the relation between actions and values, as well as our formal definition of value system, we now focus on the relation between norms and values. Recall the relationships depicted in Fig. 2. There, norms promote values: we capture

[8] In particular, a ranking is irreflexive, transitive, and total. Note that, by being irreflexive and transitive, this relation disallows the existence of cycles over preferences: $\nexists v_1, \ldots, v_k$, s.t. $v_1 \succeq \cdots \succeq v_k \succeq v_1$ and $v_1 \nsim \cdots \nsim v_k \nsim v_1$.

this relationship by means of the so-called *norm promotion function*. Specifically, this norm promotion function evaluates how much each norm promotes each value, considering the norm's deontic operator and the praiseworthiness of its regulated action. In this section, we first characterise the properties the norm promotion function ought to satisfy and then propose one.

5.1 Characterising Norm Promotion

We require that a norm promotion function satisfies the following essential properties:

Deontic and judgement dependency The promotion function only depends on the deontic operator of the norm and the value judgement of its regulated action for the value.
Deontic coherence Norms that regulate the same action but have different deontic operators should have coherent promotions. In particular

- If permitting an action promotes (demotes) a value, obligating the same action must also promote (demote) that value.
- If prohibiting a norm promotes (demotes) a value, obliging or permitting it must demote (promote) that value.

Coherence (or correlation) with value judgements Norm promotion and value judgement must be aligned. This property is divided into three cases:
Neutrality If an action is neutral to a value, norms regulating the action should also be neutral to the value.
Praiseworthiness pursuit If an action is praiseworthy to a value, permitting or obliging the action should promote the value while prohibiting the action should demote the value.
Blameworthiness avoidance If an action is blameworthy to a value, permitting or obliging the action should demote the value while prohibiting the action should promote the value. Notice also that, in this case, it is worse to oblige the action than to permit it.

Formally, we include these requirements in the promotion function definition:

Definition 7 (*Promotion function*) Let $\langle \mathbb{A}, N, \mathfrak{R} \rangle$ be a normative domain with candidate norms N over the actions in $\mathbb{A}$, $\langle V, \succeq \rangle$ a value system, and $\pi : V \times N \to [-1, 1]$, a function over pairs of values and norms. We say that π is a promotion function (and therefore $\pi(v, n) \in [-1, 1]$ is the degree of promotion/demotion of n to v) if it satisfies the following properties:

Deontic and judgement dependency Suppose $n = \langle \varphi, \theta(a) \rangle$, π is a piecewise function depending on the deontic operator (we note π^{θ} the promotion function for each of the deontic operator cases).

$$\pi(v,n) = \begin{cases} \pi^{Obl}(\alpha_v^+(a), \alpha_v^-(a)) \text{ if } \theta = Obl, \\ \pi^{Per}(\alpha_v^+(a), \alpha_v^-(a)) \text{ if } \theta = Per, \\ \pi^{Prh}(\alpha_v^+(a), \alpha_v^-(a)) \text{ if } \theta = Prh, \end{cases}$$

Deontic coherence

- $\pi(v, \langle \varphi, Per(a) \rangle) \cdot \pi(v, \langle \varphi, Obl(a) \rangle) \geq 0$.
- $\pi(v, \langle \varphi, Obl(a) \rangle) \cdot \pi(v, \langle \varphi, Prh(a) \rangle) \leq 0$.
- $\pi(v, \langle \varphi, Per(a) \rangle) \cdot \pi(v, \langle \varphi, Prh(a) \rangle) \leq 0$.

Neutrality Suppose $n = \langle \varphi, \theta(a) \rangle$, then $\alpha_v^+(a) = \alpha_v^-(a) = 0 \Rightarrow \pi(v,n) = 0$.
Praiseworthiness pursuit If $\alpha_v^+(a) > 0$:

- $\pi(v, \langle \varphi, Per(a) \rangle) \geq 0$
- $\pi(v, \langle \varphi, Obl(a) \rangle) \geq 0$
- $\pi(v, \langle \varphi, Prh(a) \rangle) \leq 0$

Blameworthiness avoidance If $\alpha_v^+(a) < 0$:

- $\pi(v, \langle \varphi, Per(a) \rangle) \leq 0$
- $\pi(v, \langle \varphi, Obl(a) \rangle) \leq 0$
- $\pi(v, \langle \varphi, Prh(a) \rangle) \geq 0$
- $\pi(v, \langle \varphi, Per(a) \rangle) \geq \pi(v, \langle \varphi, Obl(a) \rangle)$

5.2 Defining a Norm Promotion Function

Considering the characterisation of the family of norm promotion functions in Definition 7, this work proposes a linear norm promotion function. The rationale behind its design is that obligations will promote the value proportionally (increasing linearly) to the praiseworthiness to perform—and blameworthiness to skip—their regulated action. Conversely, the more blameworthy a regulated action is to perform – and praiseworthy to skip – the more a prohibition norm will promote the corresponding value.

Finally, the promotion of permissions must be between that of obligations and that of prohibitions while having the same sign as that of obligations (due to deontic coherence). Therefore, we assess the promotion of permitting an action as a fraction ϵ of the promotion of obliging it. Although establishing this fraction remains a task of the decision-maker, it is worth noticing that ϵ values close to 1 will favour the selection of permission norms. In contrast, ϵ values close to 0 will favour obligations. Thus, we define the linear promotion function as follows:

Definition 8 (*Linear promotion function*) Given a normative domain with a set of candidate norms N over the actions in $\mathbb{A}$ and a value v with value judgement functions α_v^+ and α_v^-, we define $\pi_{lin} : V \times N \to [-1, 1]$, such that for a value $v \in V$ and a norm $n = \theta(a) \in N$:

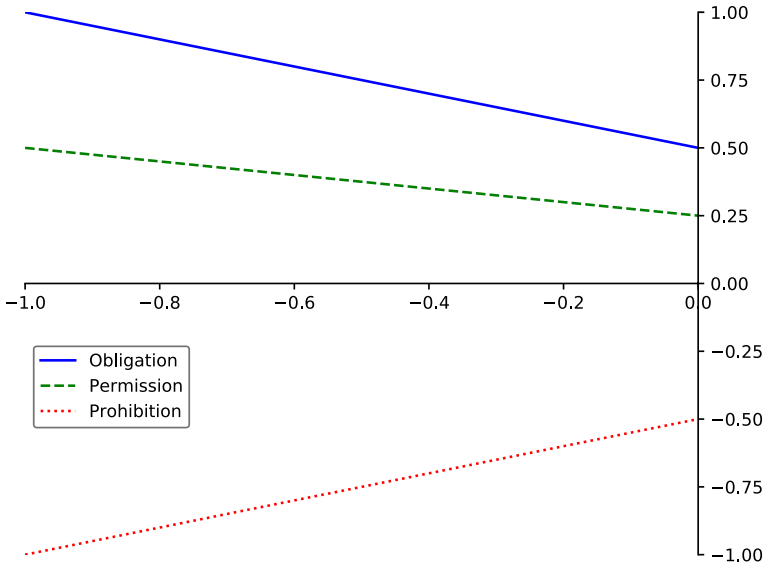

Fig. 4 Linear promotion function π_{lin} for a wholly praiseworthy action $\alpha_v^+(a) = 1$ and $\epsilon = 0.5$

$$\pi_{lin}(v,n) = \begin{cases} \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} & \text{if } \theta = Obl, \\ \epsilon \cdot \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} & \text{if } \theta = Per, \\ \frac{-\alpha_v^+(a) + \alpha_v^-(a)}{2} & \text{if } \theta = Prh, \end{cases} \tag{2}$$

where $\epsilon \in [0, 1]$. Note that ϵ ranges from $\epsilon = 0$, meaning that permissions always have 0 promotion (thus, they are disregarded), to $\epsilon = 1$, meaning that permissions have the same promotion as obligations.

Figure 4 shows an example of the linear promotion function π_{lin} for an action that is totally praiseworthy to perform ($\alpha_v^+(a) = 1$) and $\epsilon = 0.5$. The x-axis represents the range of possible blameworthiness of skipping the action $\alpha_v^-(a)$ and goes from -1 in the left to 0 in the right (being a praiseworthy action, positive values of $\alpha_v^-(a)$ cannot be considered due to Eq. 1). The three lines represent the promotion degrees (in the y-axis from -1 to 1) for the three possible norms regulating the action: obligation, permission, and prohibition. Note that since the action is praiseworthy, prohibiting it always results in a negative promotion, while permitting or obliging it always implies a positive promotion. In particular, obliging this praiseworthy action always has greater promotion value than permitting it independently of how blameworthy it is to skip.

In addition to the 2D representation of the linear norm promotion function π_{lin} in Fig. 4, we can further inspect π_{lin} in the 3D space. We do so in Fig. 5 , which depicts Obligations, Permissions, and Prohibitions for $\epsilon = 0.5$.

In each of these cases, π_{lin} is a two-variable function depending on possible different values of $\alpha_v^+(a)$ and $\alpha_v^-(a)$. Because of the definition of the moral value's

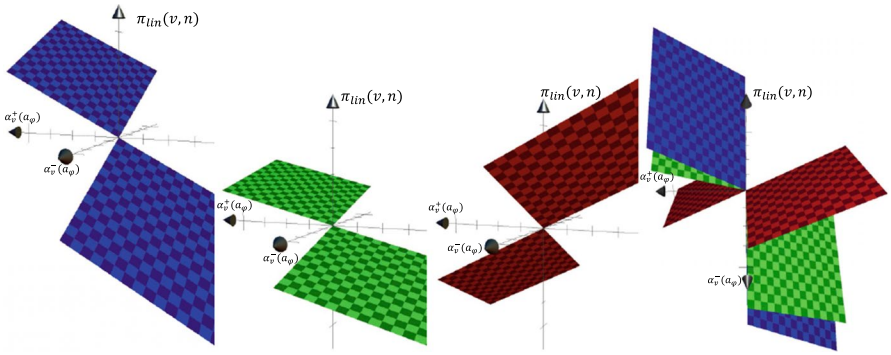

Fig. 5 Plots of each of the cases of the linear promotion function using $\epsilon = 0.5$. All axis represent values in $[-1, 1]$. The arrow on all axis marks point 1.

judgement functions (see Eq. 1), the promotion function is only defined when value judgements are of opposite sign or zero, thus $\alpha_v^+(a) \cdot \alpha_v^-(a) \leq 0$ must hold.

The surface in Fig. 5a represents the promotion function for obligations. It is a plane that is positive for actions praiseworthy to perform and blameworthy to skip, $\alpha_v^+(a) > 0$ and $\alpha_v^-(a) < 0$. On the other hand, the plane is negative for actions blameworthy to perform and praiseworthy to skip, $\alpha_v^+(a) < 0$ and $\alpha_v^-(a) > 0$. Note that in particular, given a value v and a norm $n = Obl(a)$:

- if $\alpha_v^+(a) = 1$ and $\alpha_v^-(a) = -1$, then it has a maximum promotion, $\pi_{lin}(v, n) = 1$;
- if $\alpha_v^+(a) = -1$ and $\alpha_v^-(a) = 1$, then it has a maximum negative promotion (or demotion), $\pi_{lin}(v, n) = -1$; and
- if $\alpha_v^+(a) = 0$ and $\alpha_v^-(a) = 0$, then it is neutral to the value, $\pi_{lin}(v, n) = 0$.

On the other hand, the surface in Fig. 5c represents the promotion degrees for prohibitions. Note that the promotion takes the opposite value than the promotion function does for obligations. Thus, given $n = Prh(a)$, $n' = Obl(a)$, then $\pi_{lin}(v, n) = -\pi_{lin}(v, n')$.

Finally, the surface in Fig. 5b represents the promotion function for permissions. Notice that the formula of this surface is the same as the one for obligations but scaled by $\epsilon \in [0, 1]$ (in this case $\epsilon = 0.5$). Thus, the promotion degree for a permission will be lower than for an obligation when $\alpha_v^+(a) > 0$ and $\alpha_v^-(a) < 0$, whereas it will be larger when $\alpha_v^+(a) < 0$ and $\alpha_v^-(a) > 0$, as shown in the combined plot of all cases in Fig. 5d. Note that ϵ marks the upper bound of the promotion function when evaluating permissions (and $-\epsilon$ the lower bound), that is, $\forall n = \langle Per(a) \rangle \in N$, $\pi_{lin}(v, n) \in [-\epsilon, \epsilon]$. Therefore, a smaller ϵ must be used in cases where the decision-maker prefers enforcing norms (obligations and prohibitions). In comparison, a larger ϵ must be used if the decision-maker wants to set larger promotion degrees for permissions.

6 Computing the Value-Alignment of a Norm System

At this point, we are ready to pose our central problem formally. Given a set of candidate norms and a value system, recall that our goal, as outlined in Fig. 1, is to compute the *most value-aligned* sound norm system.

6.1 Computing Value Alignment

To reason about norm systems based on moral value preferences, we must be able to compare them in terms of the moral values that they promote. The fundamental principle that we adopt for this is: *the more preferred the moral values promoted by a norm system and the higher the promotion degree, the more value-aligned the norm system*. Thus, a decision maker will opt for sound norm systems that promote the most preferred moral values, and hence are more aligned with the value system on hand.

Let $\langle \mathbb{A}, N, \mathfrak{R} \rangle$ be a normative domain. In order to quantitatively compute the value alignment of a norm system (out of the candidate norms N) with a value system $VS = \langle V, \succeq \rangle$, we will proceed as follows. First, we obtain the relevance of each moral value in VS from the value ranking $\succeq$. The relevance of a value is a numerical utility to encompass how preferred the value is (see the following paragraph). Second, we compute the value alignment of any norm system using the norm promotion of its norms to the values and the relevance of the promoted values.

To compute quantitative preferences over the moral values in VS, we define a relevance function $r : V \to \mathbb{R}$ that translates the qualitative preferences expressed by $\succeq$ to value relevance. Specifically, we require that, for $v, v' \in V$, if v is more preferred than v', then its relevance $r(v)$ must be greater than $r(v')$. Following the same reasoning, if v and v' are indifferently preferred, they have equal relevance $r(v) = r(v')$. Ultimately, by setting a relevance for each moral value, we can compare all the moral values in V.

There is no universal relevance function. Nonetheless, this function should at least satisfy two conditions. Consider the value equivalence classes in $V/\sim$ and their quotient order $\succ$. First, a relevance function must assign the same relevance to all values in each equivalence class $\eta \in V/\sim$. Second, the more preferred the equivalence class, the more relevance their values have. With these two conditions in mind, we define an example relevance function. Say that v is a value in equivalence class η. Then, we compute the relevance of v as:

$$r(v) = r(\eta) = \sum_{\eta \succ \eta'} r(\eta') + 1 \tag{3}$$

Example 7 The values in the value system of Example 6 would have the following relevance (applying Equation 3): $r(Tim) = 1$ and $r(Civ) = r(Tim) + 1 = 2$.

By using value relevance, we can calculate the value alignment of a norm system by aggregating the relevance of the moral values each norm it promotes/demotes, being the relevance of each moral value weighted by the degree of promotion/demotion from the norm to the moral value. Formally:

Definition 9 (*Value alignment score*) Given a norm system $\Omega \subseteq N$, a value system $VS = \langle V, \succeq \rangle$, and a promotion function π, we define the value alignment score of Ω with respect to VS as:

$$va(\Omega) = \sum_{n \in \Omega} \sum_{v \in V} \pi(v, n) \cdot r(v) \tag{4}$$

The following example illustrates how to compute the value alignment score of some norm systems in our running example.

Example 8 Following our running example, we now calculate how the promotion/demotion of the candidate norms towards the values using the promotion function π_{lin} from Definition 8 (in this example, we consider $\epsilon = 1$ because we do not favour obligations over permissions):

$$\begin{array}{ll} \pi(Civ, Obl(cg)) = 0.4 & \pi(Tim, Obl(cg)) = -0.5 \\ \pi(Civ, Obl(ca)) = 0.5 & \pi(Tim, Obl(ca)) = -1 \\ \pi(Civ, Per(kg)) = -1 & \pi(Tim, Per(kg)) = 1 \end{array}$$

We now assess the value alignment score of the sound norm systems in Example 4 and the values' relevance of Example 7:

$$va(\{Obl(cg)\}) = 0.3 \qquad va(\{Obl(ca)\}) = 0 \qquad va(\{Per(kg)\}) = -1$$

6.2 Problem Formalisation

In Sect. 6.1 we learned how to compute the value alignment score of a norm system in terms of the values it promotes. Now we are ready to define the so-called *value-aligned norm system selection problem* as an optimisation problem as follows:

Problem 1 (Value-aligned norm system selection problem (VANS)) Given a normative domain $\langle \mathbb{A}, N, \mathfrak{R} \rangle$, a value system $\langle V, \succeq \rangle$, and a promotion function π, the value-aligned norm system selection problem is that of finding a sound norm system $\Omega \subseteq N$ maximising value alignment. This amounts to solving:

$$\max_{\Omega \subseteq N} \Big(va(\Omega) \Big) \text{s.t. } \Omega \text{ is sound.} \tag{5}$$

7 Computing Value-Aligned Norm Systems with Optimisation Techniques

Notice that solving the VANS problem amounts to solving the optimisation problem in equation 5. Next, we show how to solve such optimisation problem as a binary integer program. A binary integer program (BIP) (Lieberman & Hillier, 2005) encodes an optimisation problem in which the decision variables take values in $\{0, 1\}$. A VANS problem can be encoded as a BIP where each decision variable represents a norm. Thus, we would have the binary decision variables $\{x_1, \ldots, x_{|N|}\}$ [9], where each x_i encodes the decision on whether a norm $n_i \in N$ is selected (taking value 1) or not (taking value 0). Thus, the VANS problem can be solved by the following binary integer program:

$$\max_{x_i \in \{0,1\}} \sum_{i=1}^{|N|} x_i \cdot va(\{n_i\}) \tag{6}$$

Subject to the following constraints:

- *Incompatibility constraints* prevent that two incompatible norms are jointly selected as part of a norm system. Thus, the following constraints must hold:

$$x_i + x_j \leq 1 \quad \text{for each } (n_i, n_j) \in \mathfrak{R}_i. \tag{7}$$

- *Generalisation constraints* ensure that two redundant norms (one generalising the other) cannot be simultaneously selected, namely:

$$x_i + x_j \leq 1 \quad \text{for each } (n_i, n_j) \in \mathfrak{R}_g. \tag{8}$$

The BIP encoding the VANS problem requires $|N|$ binary decision variables; and $|\mathfrak{R}_i| + |\mathfrak{R}_g|$ pairwise constraints (Equations 7 and 8);

Notice that the specification above corresponds to a maximization problem whose constraints are all inequalities. Hence, it is in standard form, and it can be solved with state-of-the-art solvers such as CPLEX (IBM, 1988) or Gurobi (GurobiOptimization, 2010).

Example 9 Considering the normative domain of Example 3, the value system in Example 6, and the value alignment scores of Example 8. The optimisation function to solve the running example would be:

$$0.3 \cdot x_{Obl(cg)} + 0 \cdot x_{Obl(ca)} - x_{Per(kg)}$$

where variable $x_n \in \{1, 0\}$ represents norm n. The constraints to consider in this case are: $x_{Per(kg)} + x_{Obl(ca)} \leq 1$, $x_{Per(kg)} + x_{Obl(cg)} \leq 1$, and $x_{Obl(cg)} + x_{Obl(ca)} \leq 1$. In this case, the solution is $\{Obl(cg)\}$.

[9] We assume N to be finite.

In Serramia (2021), we provide the implementation of an algorithm for encoding a VANS problem into a BIP and solve it subsequently.

8 Empirical Analysis

The purpose of the analysis in this section is to introduce a potential application employing actual-world data to illustrate how decisions on regulations vary depending on the value system at hand.

8.1 The European Value Study

The European Value Study (EVS) (EVS, 2021) is a large-scale survey research programme on the values of European citizens. Starting in 1981, EVS collects data every nine years from up to 47 European countries/regions, and since 2017, it has collaborated with the World Value Survey (WVS, 2021). EVS provides free-of-charge accessible data to foster the study of the variety of positions and trends that citizens from different countries have regarding fundamental values such as well-being, solidarity, or democracy. These values are grouped within six main topics (life, family, work, religion, politics, and society). Historically, EVS researchers aimed at " exploring the moral and social values underlying European social and political institutions and governing conduct."

Here, we aim to illustrate how the findings on values gathered from these surveys can be used to guide the legislation of specific aspects within a country. Thus, we advocate for establishing regulations that reflect the value preferences of the citizens. As a proof of concept, we focus on two values and norms in the context of three different European countries. In particular, we exploit the data from Study (2017) to analyse the relative preferences between the moral values of *permissiveness* (Knill et al., 2015) (v_{per}) and *religion* (or *religiosity* (Molteni et al., 2021), v_{rel}) to envisage a situation where policy-makers decide on the regulation of the actions of *adoption* (*adp*) by homosexual couples and *divorcing* (*div*) when justified[10].

Thus, a policy-maker may consider as candidate norms those norms permitting (*Per*) and prohibiting (*Prh*) such actions[11]:

[10] Notice that along the lines of Hanel et al. (2018), these specific values of *religion* (Molteni et al., 2021) and *permissiveness* (Knill et al., 2015) can be related to those of *tradition* and *tolerance* from Schwartz's revised model of values (Schwartz et al., 2012). They are more specific since tradition also includes cultural and family traditions, and tolerance is described as "acceptance and understanding of those who are different from oneself." Thus, we choose them to fit EVS's data better. In fact, one may even think that secularism seems a better alternative to permissiveness when comparing it to religion. However, we argue that permissiveness (Knill et al., 2015) is better suited, as it is specifically related to sexual freedom (Wikipedia, 2021), and the data from EVS we use relates to homosexual couples and divorce.

[11] Considering the norm definition in Definition 2, we exclude the norms obliging these actions (i.e., $\theta = Obl$) as they do not seem prima-facie reasonable candidate norms. Moreover, this example is by no means complete, so we assume there will be other norms in place such as those allowing heterosexual couples to adopt.

Table 1 Comparison of the value preferences of the Swedish, Albanian, and Czech populations

Country	rel. ($r(v_{rel})$)	not rel. ($r(v_{per})$)	value ranking
Sweden	371 (0.31)	818 (0.69)	$v_{per} \succeq v_{rel}$
Albania	1029 (0.72)	400 (0.28)	$v_{rel} \succeq v_{per}$
Czech Republic	374 (0.22)	1361 (0.78)	$v_{per} \succeq v_{rel}$

The "rel." column shows the number of religious citizens (i.e., those that consider religion important or very important), whereas the "not rel." column counts non-religious ones (i.e., those that consider it to be not important or not important at all)

- $n_{Per(adp)} = \langle \{homosexual\_couple\}, Per(adoption) \rangle$,
- $n_{Prh(adp)} = \langle \{homosexual\_couple\}, Prh(adoption) \rangle$,
- $n_{Per(div)} = \langle \{justified\}, Per(divorce) \rangle$,
- $n_{Prh(div)} = \langle \{justified\}, Prh(divorce) \rangle$

Moreover, norms permitting and prohibiting the same action are incompatible: $\mathfrak{R}_i = \{(n_{Per(adp)}, n_{Prh(adp)}), (n_{Per(div)}, n_{Prh(div)})\}$.

8.2 Building Value Systems from Data and Deciding on Regulations

Citizens from different countries may have different notions of the moral values under consideration (i.e., *permissiveness* and *religion*). We aim to compare the data gathered by the 2017 EVS surveys (Study, 2017) in three distinct European countries (Sweden, Albania, and the Czech Republic) to infer alternative value systems and illustrate how these would affect the selection of different regulations. We follow a three-step process for each country. First, to define each value system (see Definition 6), we specify the values' relative importance and corresponding relevance. Second, we characterise moral values by defining their value judgement functions. Third, we choose a value-aligned sound norm system for each country. In what follows, we further detail these steps and analyse the differences in the resulting regulations.

First, we resort to the EVS question "Q1F: How important is religion in your life?" to gather information about each country. Since respondents can choose between four[12] answers on a scale ranging from "Very important" to "Not at all important", we partition answers into two sets: those for which religion is important and those for which it is not. Columns "rel." and "not rel". in Table 1 detail this partition in terms of the total number of answers (and, in parentheses, their corresponding percentage) obtained for each country. Then, by assuming that those respondents in the partition that awarded importance to religion are, in fact, religious, we can infer that they would favour the ranking $v_{rel} \succeq v_{per}$, where religion is preferred over permissiveness[13]. Conversely, the non-religious partition

[12] We omit "don't know" and "no answer" options for the sake of simplicity.

[13] A thorough discussion over how these specific values and assumptions are chosen is out of the scope of this paper, where we simply use them to illustrate the overall application of our proposal (so, the general method would still apply if some of these assumptions required any modification).

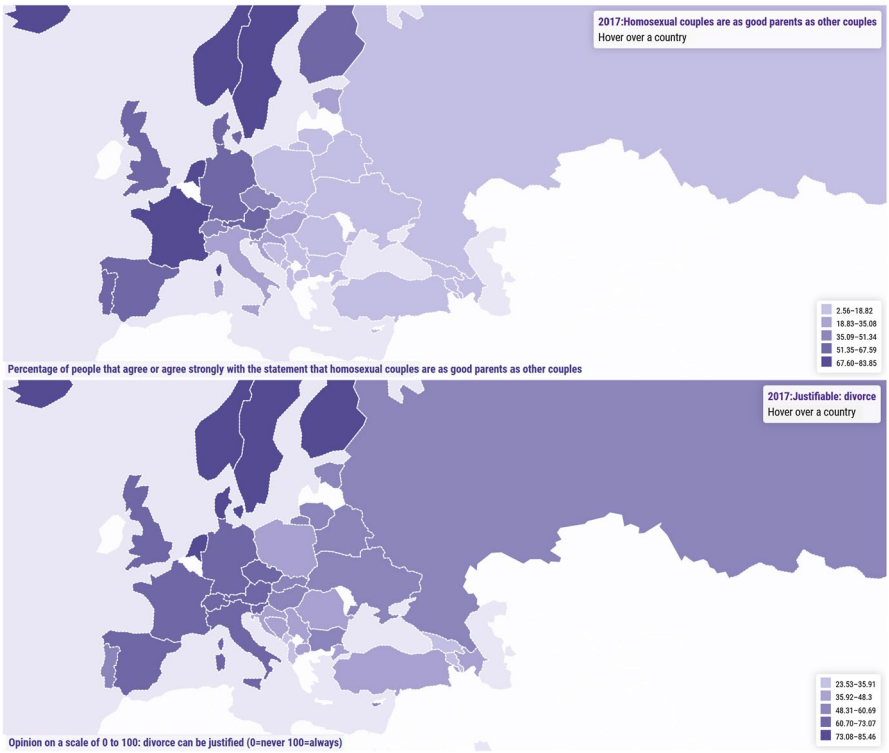

Fig. 6 Distribution of answers to questions Q27A (top) and Q44G (bottom) in the 2017 EVS survey (available at http://www.atlasofeuropeanvalues.eu/maptool.html in a normalised scale range of [0,100])

Table 2 Comparison of Swedish, Albanian, and Czech value judgement functions

Country	$\alpha^{+}_{rel}(adp)$	$\alpha^{+}_{rel}(div)$	$\alpha^{+}_{per}(adp)$	$\alpha^{+}_{per}(div)$
Sweden	0.44	0.56	0.61	0.76
Albania	− 0.64	− 0.33	− 0.58	− 0.18
Czech Republic	− 0.09	0.01	0.12	0.32

would favour $v_{per} \succeq v_{rel}$. Thus, the larger partition sets the value preferences of a country, as shown in the fourth column of Table 1, where most Swedish and Czech citizens prefer permissiveness over religion. In contrast, most Albanian have the opposite value preference.

As for value relevance, although Sect. 6.1 provides a general method for computing it, we can be more fine-grained here since we can also use the relative weight of each partition to infer the relevance of the values at hand. In this manner, we consider the relevance of the religion value ($r(v_{rel})$) to be the percentage of religious participants (again, second column in Table 1) and analogously, the relevance of v_{per} to be the percentage of non-religious participants (third column in Table 1).

Table 3 Comparison of Swedish, Albanian, and Czech norm-value alignments

Country	$va(\{n_{Per(adp)}\})$	$va(\{n_{Prh(adp)}\})$	$va(\{n_{Per(div)}\})$	$va(\{n_{Prh(div)}\})$
Sweden	0.56	− 0.56	0.7	− 0.7
Albania	− 0.62	0.62	− 0.29	0.29
Czech R.	0.07	− 0.07	0.25	− 0.25

Second, value judgement functions describe how actions are judged (see Definition 5). Here, we can resort to the specific questions in the survey that cover the actions of adoption (*apt*) and divorce (*div*) mentioned above. These questions are "Q27A: How much do you agree or disagree with the statement: Homosexual couples are as good parents as other couples?" (see Fig. 6 top) and "Q44G : Can divorce be always justified, never justified, or something in between?" (Fig. 6 bottom) respectively. Interestingly, the data allows us to correlate responses to partition responses by religious and non-religious citizens. Moreover, responses can be mapped to the $[-1, 1]$ interval to infer a judgement (with respect to the value) over the actions[14]. Thus, Table 2 details, for each country, how religious citizens evaluated the adoption by homosexual couples (see second column) and divorce (third column) together with the same evaluations by non-religious people (fourth and fifth columns resp.).

Although there are no specific questions regarding the non-performance of actions, here we may assume that $\alpha_v^-(a) = -\alpha_v^+(a)$ for these precise actions ($a \in \{adp, div\}$) and values ($v \in \{rel, per\}$), since if, for instance, $\alpha_v^+(a) < 0$ means performing action a is detrimental for value v, not performing action a may be interpreted as promoting v[15].

From these value functions, we are now ready to compute promotion function $\pi(v, n)$ (see Sect. 5.2) and asses the corresponding value alignment score for each candidate norm previously defined in Sect. 8.1. Table 3 shows the specific values for each country and norm, so that, for each norm n, we compute $va(\{n\})$ (based on the equation from Definition 9) as:

$$va(\{n\}) = \pi(rel, n) \cdot r(v_{rel}) + \pi(per, n) \cdot r(v_{per}).$$

where π corresponds to the linear promotion function in Definition 8 with $\epsilon = 1$ because we disregard obligations for this specific case. For instance, for Sweden, the value alignment score for norm $n_{Per(adp)}$ is computed as $va(\{n_{Per(adp)}\}) = \pi(rel, n_{Per(adp)}) \cdot r(v_{rel}) + \pi(per, n_{Per(adp)}) \cdot r(v_{per})$ $=0.31 \cdot 0.44 + 0.69 \cdot 0.61= 0.56$, as shown in second column and row in Table 3.

[14] Q27A responses range from 1 (meaning strong agreement) to 5 (strong disagreement), so we apply the formula $f(x) = -\frac{x-3}{2}$ to map each answer x into the $[-1, 1]$ interval. As for Q44G, since its responses range from 1 (never justifiable) to 10 (always justifiable), the mapping formula we use is $g(x) = \frac{x-5,5}{4,5}$.

[15] We take this assumption because of the preceptive nature of majoritarian European religions. Alternatively, we may consider $\alpha_v^-(a) = 0$ to signal that not performing the action is neutral to the value, which would not imply a significant change in this specific example.

The third step in deciding on regulations corresponds to computing each country's value-aligned sound norm system. This requires solving the binary integer program encoded per country as described in Sect. 7. As a result, the norm system that our method would recommend to both the Swedish and Czech policy-makers is $\{n_{Per(adp)}, n_{Per(div)}\}$, whereas it would recommend $\{n_{Prh(adp)}, n_{Prh(div)}\}$ for Albania. The next subsection provides some insights on these recommended decisions.

8.3 Discussion

The study above on EVS data evidences the need for value judgements and the sensitivity of decisions to the value system at hand.

First, our case study illustrates the need for an expressive specification of value system, counting not only on value preferences but also on value judgement functions. As we have observed, it is crucial to know the society's preferences over values and the society's interpretation of these values (which we articulate through its judgement function). Thus, while Sweden and the Czech Republic agree on their preferences on values ($v_{per} \succeq v_{rel}$), their interpretation of what v_{rel} means is significantly different (see second and fourth rows in Table 2). This is even more apparent when comparing the interpretations of values for Sweden and Albania (see second and third rows in Table 2) since the two countries exhibit opposite interpretations of each value. Furthermore, it is interesting that the Swedish interpretation of v_{rel} is more positive towards the actions under analysis than the Albanian interpretation of v_{per}.

Second, our case study shows that, given some candidate norms, the decision on the value-aligned norms to enact is sensitive to different value systems. In the case of Sweden, the two values considered judge positively all the actions, resulting in a norm system permitting them all. Contrarily, in the case of Albania, both values judge negatively all actions, and hence the resulting norm system prohibits all actions. Finally, in the case of the Czech Republic, all actions are judged positively by v_{per}, but *adp* is judged negatively by v_{rel}. Nonetheless, since v_{per} is a much more preferred value, the most value-aligned norms are those permitting all actions. Overall, even though the value systems are quite different in the three cases, particularly the value judgement functions, we have obtained the norms that would better align with the participants' values in each country.

9 Related Work

In this work, we take inspiration from the philosophy literature to use moral values (and their relative preferences) to guide the evaluation and selection of norms. Specifically, we consider moral values as ethical principles that guide the evaluation of actions (Timmons, 2012). Alternatively, Hartman (1967) formalises goodness not related to actions but to the descriptive properties of entities. Thus, a pen that cannot write is considered flawed. Defining moral values such as benevolence or security regarding their properties is a rather complex task despite its contribution.

As previously mentioned, values have also been studied from a psychological or a sociological perspective (Cheng & Fleischmann, 2010; Schwartz, 2012; Hanel et al., 2018). Thus, for example, Schwartz (2012) provides an overview of ten basic moral values recognised in cultures worldwide. Cultural differences strive in the prioritisation assigned to these values. In much the same way, we advocate for consdidering the relative priority of values when facing complex decision-making processes that involve several values. This perspective is also aligned with others from the moral agents literature , which consider value contexts (Liscio et al., 2022) or take into account value preferences when considering moral situations in specific domains, such as elderly care (Cranefield et al., 2017), as well as in general settings, such as intelligent systems design (Dignum, 2017). In fact, the work by Cranefield et al. (2017) is the closest to ours since they also propose an optimisation mechanism. However, they apply it to plan selection instead of norm selection, and they optimise (minimise) the sum of the current importance of all values.

The AI research community has been increasingly active in the study of moral agency. Thus, besides the work by Wallach and Allen on moral machines (Wallach & Allen, 2008), that of Moniz-Pereira and Saptawijaya on machine ethics (Pereira-Moniz & Saptawijaya, 2016), and that of Russell et al. on beneficial AI (Russell et al., 2015), several research papers focus on moral values. Just to mention a few others in addition to those previously discussed, Murukannaiah et al. (2020) provide an excellent roadmap to guide research on ethics and multi-agent systems. In Murukannaiah et al. (2020) terms, in this paper, we focus on developing a model of ethics based on values and norms that supports system-level ethical judgements. However, we do not focus on supporting the decision making of individual agents taking into account individual preferences, but on supporting the system designer while considering a given value system. Similarly, while we shape the environment for the agents to act accordingly to the society's values, Ajmeri (2018) tackles the question of engineering agents that can reason about values, and act ethically. Floridi and Sanders (2004) use values as thresholds: an agent is morally good if all its actions respect that threshold; and it is morally evil if some action violates it. Kohler et al. (2014) include artificial moral agents (Wallach & Allen, 2008) in multi-agent institutions to accomplish fair resource allocation. Abel and MacGlashan (2016) formalize the ethical learning and decision-making problem as solving a partially observable Markov decision process. Cointe et al. (2016) propose a judgement ability for agents to evaluate the rightness and/or goodness of both their own behaviour and those of others. Moreover, moral values are often related to norms. For instance, Sun et al. (2019) propose a regulation enforcement mechanism based on ethical considerations. Kasenberg et al. (2018) infer values (expressed as norms) by behaviour observation. Tielman et al. (2018) present a method to derive norms from actions, values and domain. Mercuur et al. (2019) compare human behaviour with agents endowed with moral values and norms which are expressed with the "non-standard" deontic operator *should*. Finally, Aydoğan et al. (2021) introduce a negotiation framework to establish norms based on individual values. Interestingly, they define value promotion (resp. demotion) as a relation between values and propositions describing context and actions. While this is similar to our value judgement functions, we consider promotion as a relation between values and norms (where the deontic operator has to be taken into account) and quantify the relation's strength. Nevertheless, there is still

room for advancing the state of the art in the formalisation of value alignment. We believe it is necessary to formalise the notion of moral value judgement of actions to formalise value alignment.

Considering the broader perspective of decision-making support, works such as Pitt et al. (2014) or Petruzzi et al. (2015) operationalise ethical considerations in resource allocation settings by proposing fairness and social capital metrics, respectively.

Argumentation constitutes another research area that has studied values. Some representative examples include the work of Bench-Capon and Atkinson (2009); Atkinson et al. (2006) and Modgil (2006), which use variations of the "Value-based Argumentation Framework" to determine the truth of a statement or to assess the goodness of specific actions. Regarding agents (moral agency), Bench-Capon extends agent reasoning with values (Bench-Capon, 2016). Specifically, value promotion or demotion is associated with changes in system states when agents perform actions. In Luo et al. (2017), this idea is further explored as authors introduce agents with opportunistic behaviour that take advantage of less informed agents to reach those state transitions that further promote their values. Although both approaches consider the impact of values and their preferences, these works consider decision-making as an individual process while we take a system-wide perspective.

To our knowledge, no other research approach tackles the same problem of value-aligned norm decision-making as stated in this paper. However, notice that w e first framed the problem of selecting the set of norms to enact in a society (Lopez-Sanchez et al., 2017). Nevertheless, selection in this work only considered norm relationships and deployment costs. Subsequently, we advanced towards considering moral values by reformulating the problem of "choosing the *right* norms to establish" in Serramia et al. (2018, 2018). Specifically, in Serramia et al. (2018), we proposed moral values as additional (explicit) preference criteria and discussed how norms can be established in new-born or highly dynamic social groups. Then, in Serramia et al. (2018), we cast this initial approach as an optimisation problem and studied its empirical hardness. In this paper, we build upon this background and extend it. Firstly, we discuss the philosophical foundations of value-aligned norm selection. Secondly, we use this philosophical basis to formalise better our theoretical approach and computational methods to selecting value-aligned norm systems. Finally, we illustrate our method with actual data from the European Value Study. We provide a detailed empirical study by considering a more comprehensive range of decision scenarios and a more fine-grained analysis. Overall, this paper advances state of the art by providing: philosophical and theoretical foundations to computing value-aligned norm systems; and computational tools to yield value-aligned norm systems.

10 Conclusions and Future Work

This paper provides the theoretical foundations for selecting norm systems that promote the most preferred moral values in a society. Additionally, it offers practical mechanisms to support the automatic selection of those norm systems. We do

so by posing the so-called *value-aligned norm system selection problem* (VANS) grounded in two structures: the normative domain, defining norms and their relationships, and the value system, containing prioritised sets of moral values. We connect these structures via the norm promotion function, which is grounded on the praiseworthiness (or value judgement) of actions and allows us to quantify the value alignment of norm systems. Then solving the problem amounts to finding the sound norm system (i.e., without conflicting nor redundant norms) that maximises value alignment. To find the solution, we encode the VANS problem as a binary integer program and solve it with a state-of-the-art solver. We illustrate our proposal with real data from the European Value Study.

10.1 Limitations

We want to inform the reader of some limitations of the presented approach. First, our approach is inherently utilitarian. This brings several risks, as commonly discussed in the literature (Rachels & Rachels, 2012). These criticisms range from the possible inability to quantify "how good" norms are towards values to the narrow view that the action's consequences concerning how they promote/demote value are all that matter when deciding the norms to enact. Another limitation of our approach is that we assume we know the value system we want the norms to align with. While eliciting value systems is still an open problem, there have been some advancements recently. Liscio et al. (2023) described a pipeline to elicit individual value systems and aggregate them into a shared value system. Importantly, researchers have already addressed some of the parts of this pipeline, such as context-specific value detection (Liscio et al., 2022) and value system aggregation (Lera-Leri et al., 2022). Finally, we assume a known set of candidate norms and compute the value-aligned set of norms out of these candidates. While our approach is useful in cases where system designers have full control, it might not be as useful in less controlled environments, for example, when norms (or values) emerge. In these cases, our contribution is limited to providing a measure of value alignment between the emerged norms and values.

10.2 Future Work

Although we treat the value-aligned norm selection problem from a theoretical point of view and illustrate it with EVS data, the framework presented in this paper could have other potential practical applications worth exploring in the future. For example, budget allocation in participatory systems (Serramia et al., 2019) (where given a budget, proposals are accepted or rejected based on their alignment with common moral values), moderation of online communities through norms (Morales et al., 2015), or value-driven modelling of public policies (Perello-Moragues & Noriega, 2020). Furthermore, an exciting addition could be that of explainability. In this regard, we could build explanations for why an individual norm or a set of norms did not get selected. These explanations could be based on the utilities of norms or norm systems and suggest changes in the settings, such as norm-value promotions or

value preferences, leading to the norm/norm system selection. Finding these alternative settings could be addressed using optimisation techniques.

Funding This work has been funded by projects VAE (TED2021-131295B-C31), VALAWAI (HE-101070930), LOGISTAR (H2020-769142), AI4EU (H2020-825619), Crowd4SDG (H2020-872944), ACISUD (PID2022-136787NB-I00), COREDEM (H2020-785907), Fairtrans (PID2021-124361OB-C33), AUTODEMO (SR21-00329), TAILOR (H2020-952215), assist-Decidim (23S02974-001), 2021 SGR 00313, and 2021 SGR 00754. F. Bistaffa was supported by the H2020-MSCAIF- 2016 HPA4CF project. The authors have no competing interests to declare that are relevant to the content of this article. Maite Lopez-Sanchez belongs to the WAI research group (University of Barcelona) an associated unit to CSIC through the IIIA.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on machine learning. ICML '04* (p. 1.) New York: ACM. https://doi.org/10.1145/1015330.1015430.

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society* (vol. 16, p. 02). London: AAAI

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*.

Ajmeri, N. (2018). Engineering multiagent systems for ethics and privacy-aware social computing. PhD thesis, North Carolina State University

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511978036

Atkinson, K., Bench-Capon, T. J. M., & McBurney, P. (2006). Parmenides : Facilitating deliberation in democracies. *Artificial Intelligence and Law, 14*, 261–275.

Audi, R. (1999). *The Cambridge dictionary of philosophy*. Cambridge University Press.

Aydoğan, R., Kafali, O., Arslan, F., Jonker, C. M., & Singh, M. P. (2021). Nova: Value-based negotiation of norms. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3465054

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics*. Oxford University Press.

Bench-Capon, T. J. M. & Atkinson, K. (2009). Abstract argumentation and values. In *Argumentation in Artificial Intelligence* (pp. 45–64). Boston: Springer. https://doi.org/10.1007/978-0-387-98197-0_3

Bench-Capon, T. (2016). Value-based reasoning and norms. In *Workshop on artificial intelligence for justice (AI4J)* (pp. 9–17). The Hague: ECAI

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bostrom, N., & Yudkowsky, E. (2011). Ethics of artificial intelligence. In *Cambridge Handbook of Artificial Intelligence*.

Campos, J., López-Sánchez, M., Salamó, M., Avila, P., & Rodríguez-Aguilar, J. A. (2013). Robust regulation adaptation in multi-agent systems. *ACM Transactions on Autonomous and Adaptive Systems, 8*, 1–27.

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Sombetzki, J., Winfield, A. F. T., & Yampolskiy, R. (2017). Towards moral autonomous systems. arxiv:1703.04741

Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., & Yeung, K. (2021). Trustworthy AI. In *Reflections on artificial intelligence for humanity* (pp. 13–39). Cham: Springer

Cheng, A.-S., & Fleischmann, K. R. (2010). Developing a meta-inventory of human values. *Proceedings of the American Society for Information Science and Technology, 47*(1), 1–10.

Chisholm, R. M. (1963). Supererogation and offence: A conceptual scheme for ethics. *Ratio (Misc.), 5*(1), 1.

Cointe, N., Bonnet, G., & Boissier, O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems* (pp. 1106–1114). Singapore: International Foundation for Autonomous Agents and Multiagent Systems

Cooper, D. (1993). *Value pluralism and ethical choice*. St. Martin Press Inc.

Cranefield, S., Winikoff, M., Dignum, V., & Dignum, F. (2017). No pizza for you: Value-based plan selection in BDI agents. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17* (pp. 178–184). Melbourne: International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2017/26.

Dignum, V. (2017). Responsible autonomy. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17* (pp. 4698–4704). Melbourne: International Joint Conference on Artificial Intelligence https://doi.org/10.24963/ijcai.2017/655.

European Comission: Artificial Intelligence Act. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed: 2023-07-13 (2023)

European Comission: Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed: 2023-07-13 (2019).

EVS: European Values Study. https://europeanvaluesstudy.eu, last visited on Sept. 2021. (2021)

Fieser, J., & Dowden, B. (2023). Ethics. https://iep.utm.edu/category/value/ethics/ (The Internet Encyclopedia of Philosophy)

Fitoussi, D., & Tennenholtz, M. (2000). Choosing social laws for multi-agent systems: Minimality and simplicity. *Artificial Intelligence, 119*(1–2), 61–101.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Frankena, W. K. (1973). *Ethics* (2nd ed.). Prentice-Hall.

Gert, B., & Gert, J. (2020). The Definition of Morality. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Fall 2020 edn. Metaphysics Research Lab, Stanford University

Grimes, M., & Marquardson, J. (2019). Quality matters: Evoking subjective norms and coping appraisals by system design to increase security intentions. *Decision Support Systems, 119*, 23–34. https://doi.org/10.1016/j.dss.2019.02.010

Grossi, D., & Dignum, F. (2005). From abstract to concrete norms in agent institutions. In *Proceedings of the third international conference on formal approaches to agent-based systems. FAABS'04* (pp. 12–29). Berlin, Heidelberg: Springer

GurobiOptimization: Gurobi. http://www.gurobi.com/ (2010)

Hadfield-Menell, D., Dragan, A., Abbeel, P. & Russell, S. (2016). Cooperative inverse reinforcement learning. In *Proceedings of the 30th international conference on neural information processing systems. NIPS'16* (pp. 3916–3924). Red Hook: Curran Associates Inc.

Hanel, P. H., Litzellachner, L. F., & Maio, G. R. (2018). An empirical comparison of human value models. *Frontiers in Psychology, 9*, 1643.

Hansson, S. O., & Hendricks, V. (2018). *Introduction to formal philosophy*. Springer.

Hansson, S. O. (2001). *The Structure of Values and Norms. Cambridge Studies in Probability, Induction and Decision Theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511498466

Hartman, R. S. (1967). *The Structure of Value: Foundations of Scientific Axiology*. Southern Illinois University Press.

IBM: CPLEX. https://www.ibm.com/analytics/data-science/prescriptive-analytics/cplex-optimizer (1988)

IEEE Standards Association: The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html. Accessed: 2023-07-13 (2016)

Kasenberg, D., Arnold, T., & Scheutz, M. (2018). Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the 2018 AAAI/ACM Conference*

on AI, Ethics, and Society. AIES '18 (pp. 184–190). New York: Association for Computing Machinery. https://doi.org/10.1145/3278721.3278774.

Keller, T., & Savarimuthu, B. T. R. (2017). Facilitating enhanced decision support using a social norms approach. *Journal of Electronic Commerce in Organizations (JECO), 15*(2), 1–15.

Knill, C., Adam, C., & Hurka, S. (2015). *On the Road to Permissiveness?* Change and Convergence of Moral Regulation in Europe. Oxford University Press.

Kohler, T., Steghoefer, J.-P., Busquets, D., & Pitt, J. (2014). The value of fairness: Trade-offs in repeated dynamic resource allocation. In *2014 IEEE eighth international conference on self-adaptive and self-organizing systems (SASO)* (pp. 1–10). London: IEEE.

Kollingbaum, M. J., Norman, T. J., Preece, A., & Sleeman, D. (2006). Norm conflicts and inconsistencies in virtual organisations. In *2006 International workshop on coordination, organizations, institutions, and norms in agent systems* (pp. 245–258). Berlin: Springer.

Lera-Leri, R., Bistaffa, F., Serramia, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. (2022). Towards pluralistic value alignment: Aggregating value systems through ℓ_p-regression. In *Proceedings of the 21st international conference on autonomous agents and multiagent systems. AAMAS '22* (pp. 780–788). Richland: International Foundation for Autonomous Agents and Multiagent Systems

Lieberman, G. J., & Hillier, F. S. (2005). *Introduction to Operations Research*. McGraw-Hill.

Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R. I. J., Jonker, C. M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., & Murukannaiah, P. K. (2023). Value inference in sociotechnical systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems. AAMAS '23* (pp. 1774–1780). Richland: International Foundation for Autonomous Agents and Multiagent Systems

Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., & Murukannaiah, P. K. (2022). What values should an agent align with? an empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems, 36*(1), 23.

Lopez-Sanchez, M., Serramia, M., Rodriguez-Aguilar, J.A., Morales, J. & Wooldridge, M. (2017). Automating decision making to help establish norm-based regulations. In *Proceedings of the 16th conference on autonomous agents and MultiAgent systems (AAMAS'17)* (pp. 1613–1615). São Paulo: International Foundation for Autonomous Agents and Multiagent Systems.

Luo, J., Meyer, J. C., & Knobbout, M. (2017). Reasoning about opportunistic propensity in multi-agent systems. In: Sukthankar, G., Rodríguez-Aguilar, J. A. (eds.) *Autonomous agents and multiagent systems—AAMAS 2017 Workshops, Best Papers*, São Paulo, Brazil, May 8-12, 2017, Revised Selected Papers. Lecture Notes in Computer Science (vol. 10642, pp. 203–221). Brazil: Springer, São Paulo. https://doi.org/10.1007/978-3-319-71682-4_13.

Maio, G. R. (2016). *The psychology of human values*. Routledge.

McNamara, P. (2011). Praise, blame, obligation, and DWE: Toward a framework for classical supererogation and kin. *Journal of Applied Logic, 9*(2), 153–170. https://doi.org/10.1016/j.jal.2009.09.007

Meinard, Y., & Cailloux, O. (2020). On justifying the norms underlying decision support. *European Journal of Operational Research, 285*(3), 1002–1010. https://doi.org/10.1016/j.ejor.2020.02.022

Mercuur, R., Dignum, V., Jonker, C., et al. (2019). The value of values and norms in social simulation. *Journal Artificial Societies and Social Simulation, 22*(1), 1–9.

Meyer, J.-J.C., & Wieringa, R. J. (Eds.). (1993). *Deontic logic in computer science: Normative system specification*. John Wiley and Sons Ltd.

Modgil, S. (2006). Value based argumentation in hierarchical argumentation frameworks. In *Proceedings of the 2006 Conference on Computational Models of Argument: Proceedings of COMMA 2006* (pp. 297–308). Amsterdam: IOS Press

Molteni, F., Ladini, R., Biolcati, F., Chiesi, A. M., Sani, G. M. D., Guglielmi, S., Maraffi, M., Pedrazzani, A., Segatti, P., & Vezzoni, C. (2021). Searching for comfort in religion: insecurity and religious behaviour during the COVID-19 pandemic in Italy. *European Societies, 23*(sup1), 704–720. https://doi.org/10.1080/14616696.2020.1836383

Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Vasconcelos, W., & Wooldridge, M. (2015). Online automated synthesis of compact normative systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS), 10*(1), 2.

Morales, J., Mendizábal, I., Sánchez-Pinsach, D., López-Sánchez, M., & Rodriguez-Aguilar, J. A. (2015). Using iron to build frictionless on-line communities. *AI Communications, 28*(1), 55–71.

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. https://doi.org/10.1007/s10458-018-9390-3

Murukannaiah, P. K., Ajmeri, N., Jonker, C. M. & Singh, M. P. (2020) New foundations of ethical multiagent systems. In *Proceedings of the 19th international conference on autonomous agents and MultiAgent systems. AAMAS '20* (pp. 1706–1710). Richland: International Foundation for Autonomous Agents and Multiagent Systems

Nadelmann, E. A. (1990). Global prohibition regimes: The evolution of norms in international society. *International organization, 44*(4), 479–526.

Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Kush, R., Campbell, M., Singh, M., & Rossi, F. (2019). Teaching ai agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development*. https://doi.org/10.1147/JRD.2019.2940428

Pereira-Moniz, L., & Saptawijaya, A. (2016). *Programming machine ethics* (Vol. 26). Springer.

Perello-Moragues, A., & Noriega, P. (2020). Using agent-based simulation to understand the role of values in policy-making. In *Advances in social simulation: looking in the mirror* (pp. 355–369). Cham: Springer. https://doi.org/10.1007/978-3-030-34127-5_35

Petruzzi, P. E., Busquets, D., & Pitt, J. (2015). A generic social capital framework for optimising self-organised collective action. In *Proceedings of the 2015 IEEE 9th international conference on self-adaptive and self-organizing systems. SASO '15* (pp. 21–30). USA: IEEE Computer Society. https://doi.org/10.1109/SASO.2015.10.

Pitt, J., Busquets, D., & Macbeth, S. (2014). Distributive justice for self-organised common-pool resource management. *ACM Transactions on Autonomous and Adaptive Systems (TAAS), 9*(3), 14.

Rachels, J., & Rachels, S. (2012). The debate over utilitarianism. In: The elements of moral philosophy. McGraw Hill,

Rodriguez-Soto, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2020). A structural solution to sequential moral dilemmas. In *Proceedings of the 19th international conference on autonomous agents and MultiAgent systems. AAMAS '20* (pp. 1152–1160). Richland: International Foundation for Autonomous Agents and Multiagent Systems.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine, 36*(4), 105–114.

Savarimuthu, B. T. R., Cranefield, S., Purvis, M. A., & Purvis, M. K. (2013). Identifying prohibition norms in agent societies. *Artificial Intelligence and Law, 21*(1), 1–46. https://doi.org/10.1007/s10506-012-9126-7

Schwartz, S. H. (2012). An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture, 2*(1), 2307–0919.

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., et al. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology, 103*(4), 663.

Serramia, M. (2021). Algorithm to generate the BIP encoding of a VANS problem. https://gitlab.iiia.csic.es/marcserr/vans-quant

Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J. A., & Escobar, P. (2019). Optimising participatory budget allocation: The decidim use case. Frontiers in Artificial Intelligence and Applications. In *Artificial Intelligence Research and Development* (Vol. 319, pp. 193–202). Amsterdam: IOS Press.

Serramia, M., Lopez-Sanchez, M., Rodriguez-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. AIES '18* (pp. 264–270). New York: Association for Computing Machinery. https://doi.org/10.1145/3278721.3278735.

Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Rodriguez, M., Wooldridge, M., Morales, J., & Ansotegui, C. (2018). Moral values in norm decision making. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems. AAMAS '18* (pp. 1294–1302). Richland: International Foundation for Autonomous Agents and Multiagent Systems.

Sethi, R., & Somanathan, E. (1996). The evolution of social norms in common property resource use. *The American Economic Review, 86*(4), 766–788.

Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J. & Perello-Moragues, A. (2019). Value alignment: A formal approach. In *Responsible artificial intelligence agents workshop (RAIA) in AAMAS*. Montreal: IFAAMAS

Singh, M. P. (2014). Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology, 10*(1145/2542182), 2542203.

Study, E. V. (2017). European Values Study 2017: Integrated Dataset (EVS 2017), GESIS Data Archive, Cologne, ZA7500 Data file Version 4.0.0, https://doi.org/10.4232/1.13560.

Sun, F.-Y., Chang, Y.-Y., Wu, Y.-H. & Lin, S.-D. (2019). A regulation enforcement solution for multi-agent reinforcement learning. In *Proceedings of the 18th international conference on autonomous agents and MultiAgent systems. AAMAS '19* (pp. 2201–2203). Richland: International Foundation for Autonomous Agents and Multiagent Systems

Svegliato, J., Nashed, S. B., & Zilberstein, S. (2021). Ethically compliant sequential decision making. In *Proceedings of the 35th AAAI international conference on artificial intelligence*

Tielman, M., Jonker, C., & van Riemsdijk, B. (2018). What should i do? deriving norms from actions,values and context. In: *Tenth international workshop modelling and reasoning in context* (vol. 2134, pp. 35–40). Stockholm: CEUR Workshop Proceedings.

Timmons, M. (2012). *Moral theory: An introduction*. Rowman & Littlefield Pub.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2021). Implementations in machine ethics: A survey. *ACM Computing Surveys*. https://doi.org/10.1145/3419633

Urmson, J. O. (1958). Saints and heroes. In A. I. Melden (Ed.), *Essays in moral philosophy.* University of Washington Press.

van de Poel, I., & Royakkers, L. (2011). *Ethics, technology, and engineering: An introduction*. Wiley-Blackwell.

Vasconcelos, W. W., Kollingbaum, M. J., & Norman, T. J. (2009). Normative conflict resolution in multi-agent systems. *Autonomous Agents and Multi-Agent Systems, 19*(2), 124–152.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wikipedia: Definition of Permissive Society. https://en.wikipedia.org/wiki/ Permissive_society (2021)

WVS: World Values Survey. https://www.worldvaluessurvey.org/wvs.jsp, last visited on Sept. 2021. (2021)

Yazdanmehr, A., & Wang, J. (2016). Employees' information security policy compliance: A norm activation perspective. *Decision Support Systems,92*, 36–46. https://doi.org/10.1016/j.dss.2016.09.009. A Comprehensive Perspective on Information Systems Security - Technical Advances and Behavioral Issues