Understanding the role of values and norms in practical reasoning*

Jazon Szabo[1], Jose M. Such[1], and Natalia Criado[1]

King's College London, London, UK
{jazon.szabo,jose.such,natalia.criado}@kcl.ac.uk

Abstract. Mistrust poses a significant threat to the widespread adoption of intelligent agents. Guarantees about the behaviour of intelligent systems could help foster trust. In this paper, we investigate mechanisms to integrate value-based reasoning in practical reasoning as a way to ensure that agents actions align not only with society norms but also with user's values. In particular, we expand a normative BDI agent architecture with an explicit representation of values.

Keywords: norms · values · BDI · responsible AI.

1 Introduction

There is a social need to offer guarantees about the behaviour of artificial agents. Endowing agents with the ability to reason about norms and values could enhance not only safety but also trustworthiness of these agents. Values are what we find important in life and they can be used, for example, in explanations about agent behaviour [14]. Furthermore, values can anchor agents to certain behaviours [10] and more generally, values can align the behaviour of agents with our own values, for example in cases of moral reasoning [2]. In fact, value alignment has emerged as one of the basic principles that should govern agents and is an important part of responsible AI [13]. Norms are regulative mechanisms in a society [12] and any responsible agent should be able to behave in a norm-conforming way [9]. Hence, both norms and values are needed to ensure that agents behave in a human-aligned manner [2].

In this paper we will use the following motivational example: Jay is at a restaurant and is using a software assistant to handle the payment. However, Jay is having trouble financially, and so would prefer to tip as little as possible. What should the software assistant consider to find the ideal amount to pay? There are social norms such that tipping 12.5% is ideal, but tipping at least a certain amount, say 5%, is expected. Furthermore, Jay values the happiness of the waiter, conforming to social norms and his financial security. The software assistant should recommend an amount based on Jay's desire to pay as little as possible, the norms about tipping and Jay's values.

* This work was supported by UK Research and Innovation [grant number EP/S023356/1]

In this paper, we argue that the best way to incorporate values to a cognitive agent architecture is to make values basic mental attitude. In particular, we will augment a normative BDI agent [6] with values to create an agent architecture whose behaviour is aligned with societal norms and user's values. We give a way of doing this and identify key properties of this representation of values. We also show how our architecture leads to the correct suggestion in Jay's case. Finally, we discuss related work, limitations and future work.

2 Background

2.1 Preliminaries

We focus on the integration of values and norms in practical reasoning. For this purpose, we use a normative multi-context BDI agent architecture [6] to address the different mental, ethical and normative attitudes in a modular way. A context in a normative BDI agent contains a partial theory of the world. In particular, there are contexts for beliefs, desires, intentions and norms. Reasoning in one context may affect reasoning in other contexts, which is represented by across-context inference rules, named *bridge rules*.

Let $\mathcal{L}$ be a classical propositional language (built from a countable set of propositional variables with connectives $\rightarrow$ and $\neg$). A normative BDI Agent [6] is defined by a tuple $\langle B, D, I, N \rangle$, where:

- B is the belief context, which language is formed by (γ, ρ) expressions, where γ is a grounded formula of $\mathcal{L}$; and $\rho \in [0, 1]$ represents the certainty degree associated to this proposition. The logical connective $\rightarrow$ is used to represent explanation and contradiction relationships between propositions.
- D is the desire context, which language is formed by (γ, ρ) expressions, where γ is a grounded formula of $\mathcal{L}$; and $\rho \in [0, 1]$ represents the desirability degree associated to this proposition.
- I is intention context, which language is formed by expressions such as (γ, ρ) expressions, where γ is a grounded formula of $\mathcal{L}$; and $\rho \in [0, 1]$ is the intentionality degree of proposition γ.
- N is the set of norms that affect the agent. Its language is composed of $(\langle D, C \rangle, \rho)$ expressions, where D is the deontic modality of a norm (i.e., $\mathcal{O}$bligation or $\mathcal{P}$rohibition), C is a literal of $\mathcal{L}$ representing the situations that the agent needs to bring about or avoid according to the norm, and $\rho \in [0, 1]$ is a real value that assigns a relevance to the norm. This relevance represents the degree in which the norm concerns the agent.

In normative BDI agents the information flows from perception to action according to three main steps (see Figure 1). Here, we briefly describe these steps[1], and explain those processes affected by the incorporation of values:

1. The agent *perceives* the environment and updates its beliefs, and norms.

[1] Neither normative nor practical reasoning is the focus of this paper. We use a simple normative definition to illustrate the interplay between norms and values. For a detailed description of normative and practical reasoning see [6] and [3], respectively.

2. In the *deliberation* step, the desire set is revised. New desires may be created from the user preferences as formulae according to the following bridge rule:

$$\frac{B : (desire(\alpha), \rho)}{D : (\alpha, \rho)} \tag{1}$$

meaning that if $(desire(\alpha), \rho)$ is deduced in context B, then (α, ρ) is inferred in D. Similarly, desires that have been achieved must be dropped. At this step the agent considers the norms and makes a decision about which ones it wants to obey. As a result, new desires are created for fulfilling norms. If the agent is willing to comply with an obligation, then a desire for reaching the state imposed by the obligation is created by to the following bridge rule:

$$\frac{N : (\langle \mathcal{O}, C \rangle, \rho), w(C) > \delta}{D : (C, c(\rho, w(C)))} \tag{2}$$

where w calculates the agent willingness to comply with a given norm:

$$w(C) = \frac{\sum\limits_{B:(C \to \gamma, \rho_B), D:(\gamma, \rho_D)} \rho_B \times \rho_D}{\sum\limits_{B:(C \to \gamma, \rho_B)} \rho_B} - \frac{\sum\limits_{B:(\neg C \to \gamma, \rho_B), D:(\gamma, \rho_D)} \rho_B \times \rho_D}{\sum\limits_{B:(\neg C \to \gamma, \rho_B)} \rho_B} \tag{3}$$

This function considers the desirability of the consequences of fulfilling and violating the norm together with the plausibility of these consequences to calculate the agent willingness top comply with the norm. When the willingness is greater than δ, it means that the agent is willing to comply with the obligation. The degree assigned to the new desire is calculated by the compliance function (c) that considers the relevance of the norm and the willingness to comply with it. For prohibition norms there is an analogous bridge rule creating desires to avoid forbidden states.

3. In the *decision making* step, desires help the agent to select the most suitable plan to be intended. This is implemented by the following bridge rule:

$$\frac{D : (\varphi, \delta), B : ([\alpha]\varphi, \rho), P : plan(\varphi, \alpha, c_\alpha)}{I : (\alpha_\varphi, h(\rho \times (u(\delta) - c_\alpha)))} \tag{4}$$

A formula $([\alpha]\varphi, \rho)$ is interpreted as the probability that φ satisfies the user by executing α. Then, the intention degree to reach a desire φ by means of a plan α is taken as a trade-off between the benefit of reaching this desire (calculated by u, which is a mapping that transforms desire degrees into benefits); and the cost of the plan (c_α), weighted by the belief degree ρ. h is a transformation that maps global benefits back to normalized utility degrees.

2.2 Values

Values are "what is important to us in life" [11] and play a critical role in how people behave. The most widely accepted system of values is the Schwartz

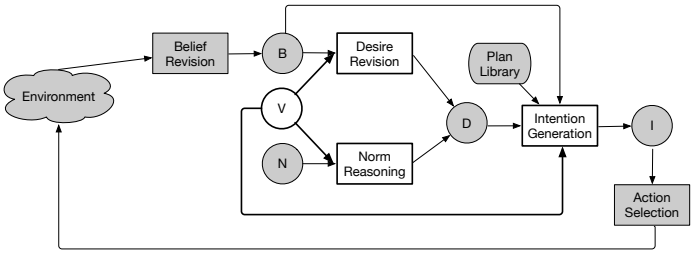

Fig. 1. Reasoning Phases in a normative BDI Agent. Context are represented as circles and bridge rules are represented as boxes. Note white boxes and circles correspond to the inclusion of values into the normative BDI architecture.

Theory of Basic Values [11], which identifies 10 different basic values shared by everyone: self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence and universalism. However, how people order values can be different from person to person. For example, in the scenario Jay considers universalism, security and comformity important, in that order[2]. Universalism means to value the welfare of all people and nature. In the scenario to care for the wealth of a waiter would be valued by universalism and so universalism would imply giving a good tip. Security means to value stability in relationships, society and one's self. In the scenario security means to value not wasting money on optional expenses (because in this scenario even such small expenses could cause financial instability) and so would imply giving a low tip or no tip at all. Conformity means to value behaving according to society's rules and expectations. In the scenario conformity values behaving according to the norm that one should tip 12.5%. As we can see values can be aligned with each other or in conflict with each other and humans make decisions considering the relative importance of multiple values. Assuming that universalism, security, conformity is the order of Jay's values, we can intuitively conclude that Jay should leave a good tip.

3 Integrating Values in Normative BDI Agents

Values are fundamentally different from beliefs, desires, intentions and norms [11]: beliefs refer to the subjective probability that something is true, not to the importance of goals as a guiding principles in life; similarly, desires, intentions and norms are about specific situations, whereas values transcend specific situations. Hence, we propose to include a new value context (see Figure 1).

Our definition of this new context is inspired by Schwartz's theory of values and the 4 key properties necessary for a representation of values [11]: (i) *comparability*, it should be possible to assess values in specific situations and to compare these different situations with respect to a specific value; (ii) *orderability*, values should

[2] We are using a subset of the ten basic values for brevity.

be ordered by importance; (iii)*practicality*, the agent should consider multiple values and their relative importance when selecting a course of action; and (iv)*normativity*, values influence whether the agent (or person) accepts or rejects a norm.

3.1 Value Language

Syntax. V_C is a set formed by 10 constants, one per each Schwartz value. The language of the value context, denoted by V, is formed by three predicates. Predicates $promote(v, \gamma, \rho)$ and $demote(v, \gamma, \rho)$ —where $v \in V_C$, γ is a propositional variable of $\mathcal{L}$, and $\rho \in [0, 1]$— represent to what degree a state of the world promotes or demotes a value. For example, the statement $demote(security, bigtip, 0.8)$ expresses that leaving a big tip demotes the value of security. Predicate $weight(v, \rho)$ —where $v \in V_C$, and $\rho \in [0, 1]$— represents the extent the agent holds a value important. For example, the statement $weight(security, 0.7)$ expresses that the agent holds the value of security fairly important.

Semantics. For every $v \in V_C$ and propositional variable γ, there is exactly one ρ such that $promote(v, \gamma, \rho)$ holds (and respectively for *demote*). Note that if a proposition doesn't promote a value, it is expressed by $promote(v, \gamma, 0)$ (and respectively for *demote*)[3]. For every $v \in V_C$, there is exactly one ρ such that $weight(v, \rho)$ holds.

3.2 Value-based Reasoning

The value context endows agents with an explicit representation of values, their importance and knowledge about which situations promote or demote some values. This representation will allow values to influence the actions taken by agents. To this aim, we need to associate each propositional variable γ of $\mathcal{L}$ with its valuing: i.e., a numerical value representing to what degree states of the world satisfying γ promote the agent's values:

$$val(\gamma) = \frac{\sum_{v \in V_C} (\rho_{promoted} - \rho_{demoted}) \times \rho_{weight}}{\sum_{v \in V_C} \rho_{weight}}$$

where $promote(v, \gamma, \rho_{promoted})$, $demote(v, \gamma, \rho_{demoted})$ and $weight(v, \rho_{weight})$ hold in context V. Note the above function will calculate the valuation as a value within the $[-1, 1]$ interval, the normalized valuation is defined as:

$$\overline{val}(\gamma) = \frac{val(\gamma) + 1}{2}$$

[3] Depending on the particular domain of application, a constraint to ensure that one of $promote(v, \gamma, 0)$ or $demote(v, \gamma, 0)$ holds could be added to the V context.

Desire Revision Bridge Rules. The desire revision bridge rules are modified to avoid the generation of any desire incompatible with the agent's values. For example, the desire generation bridge rule is modified as follows:

$$\frac{B : (desire(\alpha), \rho), \overline{val}(\alpha) > \delta_v}{D : (\alpha, \rho)} \tag{1*}$$

A desire to achieve α is generated if $\overline{val}(\alpha) > \delta_v$, where δ_v is a domain-dependent threshold determining the trade-off between the user goals and desires.

Norm Compliance Bridge Rules. Values guide the selection or evaluation of norms and, hence, it is necessary to update the Norm Compliance Bridge Rules to account for values. In particular, we propose here to modify the willingness function w to include the valuation of each consequence of a norm:

$$w(C) = \frac{\sum\limits_{B:(C\rightarrow\gamma,\rho_B),D:(\gamma,\rho_D)} (\rho_B \times \rho_D) \oplus \overline{val}(\gamma)}{\sum\limits_{B:(C\rightarrow\gamma,\rho_B)} \rho_B} - \frac{\sum\limits_{B:(\neg C\rightarrow\gamma,\rho_B),D:(\gamma,\rho_D)} (\rho_B \times \rho_D) \oplus \overline{val}(\gamma)}{\sum\limits_{B:(\neg C\rightarrow\gamma,\rho_B)} \rho_B} \tag{3*}$$

where $\oplus$ is a operator that combines the desirability and probability of a consequence with their valuation as a real value within the $[0, 1]$ interval. $\oplus$ is a function such that: $\oplus(1, 1) = 1$, $\oplus$ has as null element 0, and $\oplus$ is increasing with respect to both arguments and continuous.

Intention Generation Bridge Rule. This rule is modified to consider how intentions affect values. In particular, each plan is assessed not only in terms of cost and benefit but also in terms of the valuation of its consequences:

$$\frac{D : (\varphi, \delta), B : ([\alpha]\varphi, \rho), P : plan(\varphi, \alpha, c_\alpha)}{I(\alpha_\varphi, h((r \times (u(d) - c_\alpha)) \oplus \biguplus\limits_{\gamma \in postcond(\alpha)} \overline{val}(\gamma)))} \tag{4*}$$

where *postcond* maps each plan into its postconditions and $\biguplus$ combines different normalized valuation values into a single value within the $[0, 1]$ interval.

In our tipping example, there is a social norm of tipping 12.5% represented in context N. When the agent considers this norm, the agent creates a desire to adhere to the norm (through the norm compliance bridge rules and the willingness function) because the high importance of universalism and conformity values. From this normative desire, an intention is created. Although the plan to pay an ideal tip compromises the financial security of Jay (i.e. the value of security), this negative impact is drowned out by the positive impact the plan has on the values of universalism and comformity. Finally, the agent acts on its intention to leave an ideal tip and recommends it to its user.

3.3 Value Properties

Once the language and rules for representing and reasoning about values have been proposed we will formally discuss the ways in which this formalization satisfies the key 4 key properties of values:

- *Comparability*, predicates *promote* and *demote* allows to compare different situations with respect to a specific value.
- *Orderability*, the *weight* predicate defines an ordering on values. Associating each value with a weight is more general than an ordering on the set of values.
- *Practicality*, actions available to the agent are shaped by their desires and intentions, which themselves are generated and filtered based on values.
- *Normativity*, the modified willingness function considers not only the desirability of each norm consequence but also their impact on the agent's values, thus allowing the agent to reason about which norms to adhere to based on the relative importance of multiple values.

Note that none of these properties would be satisfied in the absence of an explicit representation of values.

4 Discussion

Recent research has looked at ways to integrate values and norms into practical reasoning. For example, Mercuur et al. [10] have incorporated values and norms into social simulations. In their work, agents can act in accordance with values or norms, but they do not consider the interplay between norms and values. However, several authors have claimed that agents should use value-based arguments to decide which action to take, including whether to comply with or violate norms [2, 12]. Cranefield et al. [5] have studied how to consider values in plan selection algorithm used by a BDI agent, choosing the plan that is most consistent with the agent's values to achieve a given goal. However, other aspects of value-based reasoning, such as the interplay between values and goals and norms are not considered.

In our work we state that values and norms play a more fundamental role in the functioning of a BDI agent, and a combination of these two mental attitudes enable agents behave in a way that is more aligned with human expectations. In particular, we have made a first attempt to expand a normative BDI architecture [6] with an explicit representation of values and identified 3 key ways in which values influence behaviour: (i) determining which norms should be complied with; (ii) determining which goals are worth pursuing; and (iii) determining the course of action to achieve a goal.

In this paper, we have proposed efficient ways to integrate value-based reasoning in a normative BDI agent by focusing on the the quantitative aspects of value promotion and demotion. However, values are usually the object of deliberations of a different nature:

- Values rarely play an explicit role in common decisions. Value-based reasoning is more frequent when humans are faced with new dilemmas usually having conflicting implications for different values [11].
- Situations with respect to a value can not only differ in the degree that the value is promoted or demoted but also in what quality of the value is being promoted or demoted: consider the difference between the relaxing and

luxurious pleasure one gets from lying in the sun and the intense and sharp pleasure one gets from quenching a thirst [4]. Even though both actions promote pleasure, they do it in a way that differs not only in how much pleasure is being promoted but also in what kind of pleasure is being promoted.
– Research suggest that humans have an ordering among values. However, it is not clear that humans can state quantitatively value importance or assess in absolute terms how particular situations promote and demote values.

As future work we will work on how to incorporate quantitative reasoning with forms of reasoning that more adequate for value-based reasoning; e.g., severity-based approach [8], coherence maximisation [7] or argumentation [1].

References

1. Atkinson, K., Bench-Capon, T.J.M.: Taking account of the actions of others in value-based reasoning. Artif. Intell. **254**, 1–20 (2018)
2. Bench-Capon, T.J.M., Modgil, S.: Norms and value based reasoning: justifying compliance and violation. Artif. Intell. Law **25**(1), 29–64 (2017)
3. Casali, A., Godo, L., Sierra, C.: A graded BDI agent model to represent and reason about preferences. Artif. Intell. **175**(7-8), 1468–1478 (2011)
4. Chang, R.: Value pluralism (2015)
5. Cranefield, S., Winikoff, M., Dignum, V., Dignum, F.: No pizza for you: Value-based plan selection in BDI agents. In: Proc. of IJCAI. pp. 178–184 (2017)
6. Criado, N., Argente, E., Noriega, P., Botti, V.J.: Reasoning about norms under uncertainty in dynamic environments. Int. J. Approx. Reasoning **55**(9), 2049–2070 (2014)
7. Criado, N., Black, E., Luck, M.: A coherence maximisation process for solving normative inconsistencies. Autonomous Agents and Multi-Agent Systems **30**(4), 640–680 (2016)
8. Gasparini, L., Norman, T.J., Kollingbaum, M.J.: Severity-sensitive norm-governed multi-agent planning. Autonomous Agents and Multi-Agent Systems **32**(1), 26–58 (2018)
9. Malle, B.F., Bello, P., Scheutz, M.: Requirements for an artificial agent with norm competence. In: Proc. of AIES. pp. 21–27 (2019)
10. Mercuur, R., Dignum, V., Jonker, C.: The value of values and norms in social simulation. J. Artificial Societies and Social Simulation **22**(1) (2019)
11. Schwartz, S.H.: An overview of the schwartz theory of basic values. Online readings in Psychology and Culture **2**(1), 11 (2012)
12. Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J.A., Morales, J., Wooldridge, M.J., Ansótegui, C.: Exploiting moral values to choose the right norms. In: Proc. of AIES. pp. 264–270 (2018)
13. Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., Perello-Moragues, A.: Value alignment: a formal approach
14. Winikoff, M., Dignum, V., Dignum, F.: Why bad coffee? explaining agent plans with valuings. In: Proc. of SAFECOMP. pp. 521–534 (2018)