On Conforming and Conflicting Values

Kinzang Chhogyal[1] , **Abhaya Nayak**[1] , **Aditya Ghose**[2] , **Mehmet Orgun** [1] , **Hoa Dam** [2]

[1]Macquarie University, Sydney, Australia
[2]Univ. of Wollongong, Wollongong, Australia
{kin.chhogyal, abhaya.nayak,mehmet.orgun}@mq.edu.au, {aditya, hoa}@uow.edu.au

Abstract

Values are things that are important to us. Actions activate values - they either go against our values or they promote our values. Values themselves can either be conforming or conflicting depending on the action that is taken. In this short paper, we argue that values may be classified as one of two types - *conflicting* and *inherently conflicting* values. They are distinguished by the fact that the latter in some sense can be thought of as being independent of actions. This allows us to do two things: i) check whether a set of values is consistent and ii) check whether it is in conflict with other sets of values.

1 Introduction

The pervasiveness of AI in society has benefitted us but has also spurred on much public debate about important issues including ethics and trust in AI systemts. This has resulted in a growing interest in the research on values such as in *value sensitive design* [Friedman *et al.*, 2013] where systems are designed by identifying values that are important and then translating them into design requirements [Van de Poel, 2013]. Others have focussed on the use of values in argumentation [Bench-Capon and Atkinson, 2009], on their relation with norms [Serramia *et al.*, 2018; Ghose and Savarimuthu, 2012], in characterising opportunistic propensity [Luo *et al.*, 2017], and in plan selection for BDI agents [Cranefield *et al.*, 2017]. However, in most of these works except [Cranefield *et al.*, 2017], the treatment of values is limited - they are cast as abstract entities, usually with some preference ordering - and the research is more on the use of values as a means to an end rather than on the values themselves. It seems then that we must try to get a deeper understanding of values and in this paper, we look at one particular aspect of values.

In Psychology, Schwartz's *Theory of Basic Human values* [Schwartz, 2012] paints a richer picture of values. It is assumed that all values share certain features. Of them, three are particularly relevant: i) values can be activated and cause emotions to arise, ii) they can influence our goals and therefore the choice of our actions and, iii) they may have a preference ordering. He also identifies ten broad (abstract) values under which it is assumed most concrete values fall. What is of most interest to us is the dynamics of values; it is stated that *actions in pursuit of any value have consequences that conflict with some values but are congruent with others* [Schwartz, 2012]. This paper is motivated by that particular statement.

We begin by presenting a simple formalisation that captures what it means for values to conform or conflict with each other when actions are executed. We show that adopting this formalisation, leads us to special pairs of values that are always in conflict which we call *inherently conflicting* values and we end by briefly discussing some implications of this work.

2 Values

We assume there is a set of all values, $\mathcal{V} = \{a, b, \ldots\}$, from which agents draw their values. These values represent concrete values which can be thought to fall under the ten broad values [Schwartz, 2012]. We also assume there is a set of possible actions $\mathcal{A} = \{a, a', \ldots\}$ that agents can execute. Let $\mathcal{S}$ be the set of states that the world can be in and by $S(a)$ we denote a subset of $\mathcal{S}$ where a is executable. The symbol a can represent both a value or an action but it is usually clear from the context what a is referring to.

2.1 Conformance and Conflict

We begin with the definition of a *value state* of a value which is inspired by the one in [Cranefield *et al.*, 2017].

Definition 1 *Give a value v, the value state of v is denoted as $VS(v)$ where $VS(v) \in \mathbb{N}^+$.*

The actual representation of the value state is not important and neither are the bounds. What is important is that value states can increase and decrease. We said in the Introduction that one of the properties of values is that they can be activated and stir up emotions [Schwartz, 2012]. Actions have the potential to activate values which results in an increase or decrease in the corresponding value states. We say potential because the state under which the action is executed may determine which values are activated. In some states under certain actions, all values might get activated and in some few or none of the values might get activated. An action that causes the value state of a value to increase is interpreted as one that *promotes* the value where as if it decrease the value state, it *acts against* the value. We use the following notation to show how the value state of a value v changes given an action a and a state $s \in S(a)$:

$$(s, a, v) \rightarrow v^{*} \tag{1}$$

where $*$ is $\uparrow$, $\downarrow$ or $\leftrightarrow$ and indicates whether the value state of v increases, decreases or remains unchanged respectively.

Definition 2 (Indifference) *A value v is indifferent to an action a in state $s \in S(a)$ if whenever a is executed in s, $VS(v)$ remains unchanged.*

$$(s, a, v) \rightarrow v^{\leftrightarrow} \tag{2}$$

We take it that it is not possible to have a value v that is indifferent to every action in every state. This would make v meaningless and it shouldn't have been included as a value in the first place. We state it as the following condition:

Condition-1: For any value v, there must be at least one action a and one state $s \in S(a)$ which activates v.

Definition 3 (Conflicting Values) *Given an action a, a state $s \in S(a)$, and two values $v, v' \in \mathcal{V}$, if whenever a is executed in s, $VS(v)$ increases (decreases) and $VS(v')$ decreases (increases) then we say v and v' are conflicting values with respect to a and s.*

$$\begin{aligned} &(s, a, v) \rightarrow v^{\uparrow} \textit{ and } (s, a, v') \rightarrow v'^{\downarrow} \quad \textit{or} \\ &(s, a, v) \rightarrow v^{\downarrow} \textit{ and } (s, a, v') \rightarrow v'^{\uparrow} \end{aligned} \tag{3}$$

Ex 1 *Consider that you value both* frugality *and* quality. *If you decide to buy a pine dining table that costs \$100 over an oak table that costs \$200, it increases the value state of frugality but decreases the value state of quality. Thus, they are conflicting values in this situation.*

Definition 4 (Conforming Values) *Given an action a, a state $s \in S(a)$, and two values v and v', if whenever a is executed in s, $VS(v)$ increases (decreases) and $VS(v')$ also increases (decreases) then we say v and v' are conforming values with respect to a and s.*

$$\begin{aligned} &(s, a, v) \rightarrow v^{\uparrow} \textit{ and } (s, a, v') \rightarrow v'^{\uparrow} \quad \textit{or} \\ &(s, a, v) \rightarrow v^{\downarrow} \textit{ and } (s, a, v') \rightarrow v'^{\downarrow} \end{aligned} \tag{4}$$

Ex 2 *Again take the values of frugality and quality but this time you notice that the oak table is being offered at a discount of 50%. If you buy the oak table, it will increase the value states of both frugality and quality. The two values are conforming in this situation.*

Definition 5 (Inherently Conflicting Values) *Given two values $v, v' \in \mathcal{V}$, if for all actions $a \in \mathcal{A}$ and for all states $s \in S(a)$, v and v' are conflicting values or if both v and v' are indifferent, we say v and v' are inherently conflicting values. We also say v inherently conflicts with v' and vice versa.*

Ex 3 *Consider that you value a* free market economy *over a* regulated economy. *If you are a legislator and you decide to support legislation that increases the tariff on imported goods, it decrease the value state of free market economy and increases the value state of regulated economy. In fact, any action that promotes one will go against the other and thus they are always in conflict.*

Note that because of Condition-1 previously stated, it is not possible to have inherently conflicting values that are only indifferent and not conflicting. By conflicting, we will mean values that are conflicting but not inherently conflicting unless otherwise stated.

Definition 6 (Inherently Conforming Values) *Given two values v and v', if for all actions $a \in \mathcal{A}$ and for all states $s \in S(a)$, v and v' are conforming values or if both v and v' are indifferent, we say v and v' are inherently conforming values. We say v inherently conforms with v' and vice versa.*

Observation 1 *Given two values v' and v'' that are are inherently conforming and another value v:*

a) if $v'(v'')$ conflicts (conforms) with v, then $v''(v')$ also conflicts(conforms) with v, and

b) if $v'(v'')$ inherently conflicts with v, then $v''(v')$ also inherently conflicts with v.

Observation 2 *Given two values v' and v'' that are are inherently conflicting and another value v:*

a) if $v'(v'')$ conflicts (conforms) with v, then $v''(v')$ conforms(conflicts) with v, and

b) if $v'(v'')$ inherently conforms with v, then $v''(v')$ also inherently conflicts with v.

Proposition 1 *If two values v' and v'' are inherently conflicting with v, then v' and v'' are inherently conforming.*

Proof 1 *Assume the antecedent. For contradiction, assume v' and v'' are not inherently conforming. So there must be an action a and a state $s \in S(a)$ such that $(s, a, v') \rightarrow v'^{\uparrow}$ and $(s, a, v'') \rightarrow v''^{\downarrow}$ or $(s, a, v') \rightarrow v'^{\downarrow}$ and $(s, a, v'') \rightarrow v''^{\uparrow}$. However, since v is inherently conflicting with v' and v'', it must be that either $(s, a, v) \rightarrow v^{\uparrow}$ or $(s, a, v) \rightarrow v^{\downarrow}$ and both v' and v'' must simultaneously either increase or decrease their value state, which results in a contradiction.* □

This could be contentious but it seems that representing two inherently conforming values separately doesn't offer us much; from Observation 1 we can see that whatever we can say of one value - with respect to the other values they conflict or confirm with - is true of the other. This suggests that perhaps two inherently conforming values should be collapsed into one as they virtually represent the same value. On the other hand, if we have multiple values that are inherently in conflict with another value, then from Proposition 1, we know they are inherently conforming, and again from Observation 1 we get the same argument for collapsing them into one. On the basis of this discussion, we introduce some further conditions that we assume to hold henceforth.

Condition-2: For any value v, if it inherently conforms with a value v', then v and v' are the same value.

Condition-3: For any value v, if it inherently conflicts with v', then v' is the only value it inherently conflicts with.

Ex 4 *Let $\mathcal{S} = \{s1, s2\}$, $\mathcal{A} = \{a1, a2\}$ and let $V = \{a, b, c, d\}$. Let $a1$ and $a2$ be executable in every state, i.e. $S(a1) = S(a2) = \mathcal{S}$. The dynamics of the values are shown in Table 1. Values that are indifferent to actions are not shown in the state-action cells. The following observations can be made:*

1. *a and b are inherently conflicting because in every state where $a^{\uparrow}$, $b^{\downarrow}$ or $a^{\downarrow}$, $b^{\uparrow}$ or $a^{\leftrightarrow}, b^{\leftrightarrow}$.*
2. *a and c are conflicting in $(s1, a')$ as $a^{\uparrow}, c^{\downarrow}$; they are conforming in $(s2, a')$ as $a^{\downarrow}, c^{\downarrow}$.*
3. *a and d are not inherently conflicting because in $(s1, a'')$ and $(s2, a'')$, we see $d^{\uparrow}$ and $d^{\downarrow}$ respectively, whereas $a^{\leftrightarrow}$ in both cases.*

	state $s1$	**state** $s2$
action a'	$a\uparrow, b\downarrow, c\downarrow, d\downarrow$	$a\downarrow, b\uparrow, c\downarrow, d\uparrow$
action a''	$c\downarrow, d\uparrow$	$d\downarrow$

Table 1: Conforming, Conflicting and Inherently Conflicting Values. Details in Ex. 4.

Given a set $\mathcal{A}$ of possible actions and the set $\mathcal{S}$ of all states, we can divide the set of all values $\mathcal{V}$, into two sets: one set containing all pairs of inherently conflicting values and the second set consisting of the remaining values. The symbol $\mathcal{V}^{\perp} \subseteq \mathcal{V}$ represents the set of all inherently conflicting values. From Condition-3, we know that for each value in $v \in \mathcal{V}^{\perp}$, v has exactly one inherently confliciting $v' \in \mathcal{V}$ which we will denote as $\overline{v}$. Pairs of inherently conflicting values are special because even though we know they are activated by actions, since they are always in conflict or indifferent to any action, this allows us to talk about them without mentioning actions and in this sense they can be seen as being independent of actions. This perspective allow us to do two things:

a) given a set of values, it allows us to define what it means for that set to be consistent and,

b) given two sets of values, it allows us to define when one value set is in conflict with another.

This would not be possible if we tried doing the same thing with conflicting values without also talking about the particular actions involved.

Definition 7 *A set of values V is inconsistent iff there exists values $v, v' \in V$ such that $v,\ v' \in \mathcal{V}^{\perp}$ and $v' = \overline{v}$. Otherwise, it is consistent.*

Definition 8 *Two sets of values V and V' are conflicting iff there exists values $v \in V$ and $v' \in V'$ such that $v, v' \in \mathcal{V}^{\perp}$ and $v' = \overline{v}$. Otherwise, V and V' are non-conflicting.*

Ex 5 *Let $\mathcal{V} = \{a, \overline{a}, b, c, \overline{c}, d, \overline{d}, e, f\}$. We have $\mathcal{V}^{\perp} = \{a, \overline{a}, c, \overline{c}, d, \overline{d}\}$. Also, let $V = \{a, b, c\}$, $V' = \{\overline{a}, d, e\}$ and $V'' = \{d, \overline{d}, f\}$. We can say the following:*

1. *V and V' are consistent as neither contain inherently conflicting values.*
2. *V'' is inconsistent as it contains both d and $\overline{d}$.*
3. *V and V' are conflicting because of $a \in V$ and $\overline{a} \in V'$.*
4. *V and V'' are non-conflicting.*

3 Discussion and Conclusion

We conclude this paper by briefly addressing two issues that result from our presentation:

1. When talking about the dynamics of values, aside from the examples that were provided, the definition of conforming, conflicting and inherently conflicting values was entirely based on actions and states with little mention of agents. This alludes to the idea that perhaps values are Platonic Ideals that exist as separate entities outside the agent and yet we know without agents, values would be meaningless. We don't really see a problem to this separation and it has been done previously in other areas of AI. For instance, consider the various *action languages* [Gelfond and Lifschitz, 1998] used for talking about the effects of actions. Even though actions are executed by agents, actions are generally talked about in terms of their pre-conditions and post-conditions without referring to any agent at all.
2. We said that there are inherently conflicting values and this bring up question of whether all values have a value that they inherently conflict with? If we start with a set of values $\mathcal{V}$, and there is a value v that doesn't have an inherently conflicting value, there is nothing to stop us from introducing a new value $\overline{v}$ in $\mathcal{V}$. An argument against it could be that pairs of inherently conflicting values should correspond to values that occur in society and are "naturally" in conflict. For example, take *pro-choice* and *pro-life*, values related to abortion - any action that promotes one is clearly going to go against the other. On the other hand, it is hard to think of a value like *healthy lifestyle* as having an inherently conflicting value. If we try to construct one, we end up with what feels like an "unnatural" and "artificial" value like *unhealthy lifestyle* which no person would hold.

In this short paper, we presented a simple formalisation of the dynamics of values which lead to values being classified as conforming, conflicting, inherently conforming and inherently confirming. We also argued that there are distinct pairs of values that are inherently in conflict and no other ones. We hope this brief paper will stimulate further discussion on the meaning of conflicts between values and on the topic of values in general.

References

[Bench-Capon and Atkinson, 2009] Trevor Bench-Capon and Katie Atkinson. Abstract argumentation and values. In *Argumentation in artificial intelligence*, pages 45–64. Springer, 2009.

[Cranefield *et al.*, 2017] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. 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*, pages 178–184, 2017.

[Friedman *et al.*, 2013] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*, pages 55–95. Springer, 2013.

[Gelfond and Lifschitz, 1998] Michael Gelfond and Vladimir Lifschitz. Action languages. *Electronic Transactions on Artificial Intelligence*, 3:195–210, 1998.

[Ghose and Savarimuthu, 2012] Aditya Ghose and Tony Bastin Roy Savarimuthu. Norms as objectives: Revisiting compliance management in multi-agent systems. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 105–122. Springer, 2012.

[Luo *et al.*, 2017] Jieting Luo, John-Jules Meyer, and Max Knobbout. Reasoning about opportunistic propensity in multi-agent systems. In *Autonomous Agents and Multiagent Systems*, pages 203–221, Cham, 2017. Springer International Publishing.

[Schwartz, 2012] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

[Serramia *et al.*, 2018] Marc Serramia, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, Manel Rodriguez, Michael Wooldridge, Javier Morales, and Carlos Ansotegui. Moral values in norm decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 1294–1302, 2018.

[Van de Poel, 2013] Ibo Van de Poel. Translating values into design requirements. In *Philosophy and engineering: Reflections on practice, principles and process*, pages 253–266. Springer, 2013.