ICRES 2018: International Conference on Robot
Ethics and Standards, New York, USA, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.021

Demystifying "Value Alignment": Formally Linking Axiology to Ethical Principles in a Deontic Cognitive Calculus

Selmer Bringsjord, Naveen Sundar G and Atriya Sen

1. Introduction

A lot of people in and around AI who are concerned about immoral and/or destructive and/or unsafe AIs are running around calling for "value alignment" in such machines.* Our sense is that most of the time such calls are vapid, since those issuing the calls don't really know what they're calling for, and since the phrase is generally nowhere to be found in ethics itself. Herein we carry out formal work that puts some rigorous flesh on the calls in question. This work, quite limited in scope for now, forges a formal and computational link between axiology (the theory of value) and ethical principles (propositions that express obligations, prohibitions, etc.).

$\mathscr{CC}_{\mathcal{D}}$ is the space of deontic cognitive calculi. Hitherto, in particular calculi in this space that have been specified, and from there in some cases implemented, there is an absence of principled axiology. A case in point is $\mathcal{DCEC}^*$; this calculus, presented and used in (Govindarajulu & Bringsjord 2017*a*), specifies and implements what may so far be the most expressive, nuanced ethical principle to be AI-ified† — yet there is no principled axiology in this calculus, and hence none reported in the paper in question. Instead, an intuitive notion of positive and negative value, based on elementary arithmetic, and consistent with at least naïve forms of consequentialist ethical theories, is given and employed. In the present short abstract, we encapsulate our formally forging a link between Chisholm's

*E.g. see https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/.
†The Doctrine of Double-Effect.

(1975) intrinsic-value axiology,[‡] and ethical principles, in order to introduce the road to doing this in a rich way, subsequently.

2. Chisholm's Intrinsic-Value Axiology, Absorbed

Chisholm (1975) gives a theory of *intrinsic* value expressed in a quantified propositional logic that takes as primitive a binary relation *Intrinsically Preferable*; this allows him to e.g. write $P(p, q)$, where here P is the primitive relation, and p and q are of course propositional variables. To absorb Chisholm's system into a deontic cognitive calculus, we begin by immediately recasting Chisholm's theory in quantified multi-modal logic, in which all his first-order relations are modal operators, and his propositional variables are variables ranging over unrestricted formulae (which may themselves have modal operators and quantifiers within them). E.g., P becomes the binary modal operator $\mathbf{Pref}$.[§] We cast Chisholm's six definitions as axioms, and translate his five axioms as described immediately above. This yields 11 axioms, specified as follows:

A1 $\forall \phi, \psi\ [\mathbf{Same}(\phi, \psi) \leftrightarrow \neg\mathbf{Pref}(\phi, \psi) \wedge \neg\mathbf{Pref}(\psi, \phi)]$
A2 $\forall \phi\ [\mathbf{Indiff}(\phi) \leftrightarrow \neg\mathbf{Pref}(\phi, \neg\phi) \wedge \neg\mathbf{Pref}(\neg\phi, \phi)]$
A3 $\forall \phi\ [\mathbf{Neutral}(\phi) \leftrightarrow \exists \psi\ (\mathbf{Indiff}(\psi) \wedge \mathbf{Same}(\phi, \psi)]$
A4 $\forall \phi\ [\mathbf{Good}(\phi) \leftrightarrow \exists \psi\ (\mathbf{Indiff}(\psi) \wedge \mathbf{Pref}(\phi, \psi)]$
A5 $\forall \phi\ [\mathbf{Bad}(\phi) \leftrightarrow \exists \psi\ (\mathbf{Indiff}(\psi) \wedge \mathbf{Pref}(\psi, \phi)]$
A6 $\forall \phi, \psi\ [\mathbf{ALAG}(\phi, \psi) \leftrightarrow \neg\mathbf{Pref}(\psi, \phi)]$
A7 $\forall \phi, \psi\ [\mathbf{Pref}(\phi, \psi) \rightarrow \neg\mathbf{Pref}(\psi, \phi)]$
A8 $\forall \phi, \psi, \gamma\ [(\mathbf{ALAG}(\psi, \phi) \wedge \mathbf{ALAG}(\gamma, \psi)) \rightarrow \mathbf{ALAG}(\gamma, \phi)]$
A9 $\forall \phi, \psi\ [(\mathbf{Indiff}(\phi) \wedge \mathbf{Indiff}(\psi)) \rightarrow \mathbf{Same}(\phi, \psi)]$
A10 $\forall \phi\ [(\mathbf{Good}(\phi) \wedge \mathbf{Bad}(\neg\phi)) \rightarrow \mathbf{Pref}(\phi, \neg\phi)]$
A11 $\forall \phi, \psi\ [\neg(\mathbf{Pref}(\phi, \phi \vee \psi) \wedge \mathbf{Pref}(\psi, \phi \vee \psi)) \wedge \neg(\mathbf{Pref}(\phi \vee \psi, \phi) \wedge \mathbf{Pref}(\phi \vee \psi \vee \psi))]$

3. Some Axiological Theorems, Machine-Discovered/Verified

In order to obtain some object-level theorems from A1–A11, we of course must have a proof theory to anchor matters; and in order for these theorems to be automatically obtained we shall of course need this theory to

[‡]The earlier version of which is (Chisholm & Sosa 1966).
[§]Later, it will become necessary to move beyond Chisholm by allowing this operator to take into account agents α, and thus it will become ternary: $\mathbf{Pref}(\phi, \psi, \alpha)$. The other operators in the modalized axiology of Chisholm would of course need to include a placeholder for agents.

be implemented. We use a simple proof theory, $\mathcal{R}_{\mathcal{A}}$, for now. Our first ingredients are inference schemata needed for quantificational reasoning over the 11 axioms. We thus allow the standard natural-deduction schemata allowing introduction and elimination of $\exists$ and $\forall$ in A1–A11; and of course we allow the remaining natural-deduction schemata for first-order logic: *modus ponens*, indirect proof, and so on, where the wffs allowed in these schemata may contain operators. In this mere abstract, we don't discuss any of the theorems that are now reachable, nor do we show that these theorems can be automatically proved by ShadowProver (Govindarajulu & Bringsjord 2017*b*, Govindarajulu, Bringsjord, Ghosh & Peveler 2017).

We in addition invoke the inference schemata (resp., as S_g and S_b)

$$\frac{\vdash_{\mathcal{R}_{\mathcal{A}}} \phi}{\mathbf{Good}(\phi)}$$

and

$$\frac{\vdash_{\mathcal{R}_{\mathcal{A}}} \neg\phi}{\mathbf{Bad}(\phi)}$$

4. Bridging to Ethical Principles

So far there is no connection between value and traditional ethical categories, such as the *obligatory* and *forbidden*. In the full paper, we introduce "bridging" principles that take us from value to not only these two categories, but to the complete spectrum of ethical categories in (Bringsjord 2015). Here, put informally, are two of the principles we formalize and explore:

P1/P2 Where ϕ is any good (bad) state-of-affairs, it ought to be (is forbidden) that ϕ.

5. Some "Value-Alginment" Theorems, Machine-Discovered/Verified

In the full paper, we explore the automated proving of the theorems in §4 by way of ShadowProver.

6. On Deriving an 'Ought' From an 'Is'

Hume famously maintained in his *Treatise of Human Nature* that an ought cannot be dervied from an is. Yet it appears that, one, those calling for

"value alignment" are calling for what Hume declared impossible, and that, two, we have nonetheless accomplished the very thing, for we have:

> **Theorem**: It ought to be that $\phi \rightarrow \phi$.
>
> **Proof**: Trivial: $\phi \rightarrow \phi$ is a theorem, and hence by S is **Good**, and thus by S_g is obligatory. **QED**

7. Conclusion and Next Steps

We have sought to explicitly and formally forge a connection between axiology and deontic concepts and propositions, in order to rationalize calls for "value alignment" in AI. Have we succeeded? At this point, confessedly, the most that could be said in our favor is that we have put on the table an encapsulation of a candidate for such forging. What additional steps are necessary?

Some are rather obvious. Goodness of states-of-affairs, and badness of them as well, would seem to fall into continua; yet what we have adapted from Chisholm and Sosa allows for no gradations in value. That goodness and badness comes in degrees appears to be the case even if attention is restricted to states-of-affairs that are intrinsically good (bad). For instance, knowledge (at least of "weighty" things, say the stunning truths about the physical world in relativity theory) on the part of a person would seem to have intrinsic value, but the selfless love of one person for another would seem to be something of even greater intrinsic value. Yet at this point, again, our axiology admits of no gradations in goodness and badness. We know that we need a much more fine-grained axiology.

Our suspicion is that goodness and badness should be divided between what has intrinsic value/disvalue, and what has value instrumentally. Instrumental value would be parasitic on intrinsic value. More specifically, we are inclined to think that both the instrumental and the intrinsic is graded.

One final remark: Even at this early phase in the forging of a formal connection between value and ethical principles based on deontic operators, it seems patently clear that because actual and concrete values in humanity differ greatly between group, nation, culture, religion, and so on, any notion that a given AI or class of AIs can be aligned with *the* set of values in humanity isn't only false, but preposterous. For many Christians, for example, the greatest intrinsic good achievable by human persons is direct and everlasting commuion with the divine person: God. For many others (e.g. (Thagard 2012)), this God doesn't exist, and the greatest goods are achieved by living, playing, and working in the present worldm in which our lives are short and end forever upon earthly death. In the context formed

by the brute fact that values among human persons on our planet vary greatly, and are, together, deductively inconsistent, our focus on formality is, we submit, prudent. Once the formal work has advanced sufficiently, presumably the alignment of AIs with values can be undertaken relative to a concretely instantiated axiology, from which ethical principles flow. For a given class of AIs, then, their behavior would be regulated by ethical principles that flow from a particular instantiation of the axiology.

References

Bringsjord, S. (2015), A 21st-Century Ethical Hierarchy for Humans and Robots: $\mathcal{EH}$, *in* I. Ferreira, J. Sequeira, M. Tokhi, E. Kadar & G. Virk, eds, 'A World With Robots: International Conference on Robot Ethics (ICRE 2015)', Springer, Berlin, Germany, pp. 47–61. This paper was published in the compilation of ICRE 2015 papers, distributed at the location of ICRE 2015, where the paper was presented: Lisbon, Portugal. The URL given here goes to the preprint of the paper, which is shorter than the full Springer version. **URL:** *http://kryten.mm.rpi.edu/SBringsjord_ethical_hierarchy_0909152200NY.pdf*

Chisholm, R. (1975), 'The Intrinsic Value in Disjunctive States of Affairs', *Noûs* **9**, 295–308.

Chisholm, R. & Sosa, E. (1966), 'On the Logic of "Intrinsically Better"', *American Philosophical Quarterly* **3**, 244–249.

Govindarajulu, N. & Bringsjord, S. (2017*a*), On Automating the Doctrine of Double Effect, *in* C. Sierra, ed., 'Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)', International Joint Conferences on Artificial Intelligence, pp. 4722–4730. **URL:** *https://doi.org/10.24963/ijcai.2017/658*

Govindarajulu, N. S. & Bringsjord, S. (2017*b*), On Automating the Doctrine of Double Effect, *in* C. Sierra, ed., 'Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17', Melbourne, Australia, pp. 4722–4730. Preprint available at this url: `https://arxiv.org/abs/1703.08922`. **URL:** *https://doi.org/10.24963/ijcai.2017/658*

Govindarajulu, N. S., Bringsjord, S., Ghosh, R. & Peveler, M. (2017), Beyond The Doctrine Of Double Effect: A Formal Model of True Self-Sacrifice. International Conference on Robot Ethics and Safety Standards.

Thagard, P. (2012), *The Brain and the Meaning of Life*, Princeton University Press, Princeton, NJ.