AI and Value Alignment

Katja Maria Vogt

ValuesLab

Plan for Today

- Intro: what is value alignment?
- Terminology: normative, evaluative, values
- Alignment in the context of replies that AI gives to prompts
- Alignment and the range of traditions in ethics

Introduction: What is Value Alignment?

Alignment is considered one of the biggest challenges relating to AI.

- We don’t want AIs that decide that it’s best to destroy the world.
- We don’t want AIs to offer instructions for harmful actions.
- We don’t want AIs to use abusive and rude language.
- We don’t want AIs that replicate human biases.
- Etc.

Millière (2023) on Value Alignment

Technical challenge: to steer the behavior of AI systems in accordance with values.

Normative challenge: with *which* values?

Which Values?

Which values, and *whose* values? People disagree about values:

- Different evaluative *outlooks/frameworks*.

- Different *ranking of values* that are otherwise shared.

- Different conceptions of shared values. E.g.: suppose lots of people endorse the value of justice, but conceive differently of justice.

- Even those who share an evaluative outlook often disagree on *specific situations/decisions*.

Terminology

Normative: good, just, right, wrong…

Descriptive: green, four, a tree, …

Normative versus Descriptive: Disagreement

When we disagree about descriptive matters, we typically know how to resolve the disagreement. For example, we count.

When we disagree about normative matters, we typically don't know how to resolve the disagreement.

What *are* Values?

What *are* values?

- Are values comparable to numbers, not accessible to sense perception, and yet indispensible for engaging with the world?

- Are values comparable to hot/cold/sweet/bitter/pleasure/pain, features of the world to which we respond affectively?

- Are values states of affairs that contribute to human life going well?

- Other ideas? (We won't resolve this!)

Kinds of Values

Short of offering an account of the nature of value, here are some *kinds of values*:

- ethical/moral (related to how to interact): justice, kindness, honesty
- epistemic (related to thinking): wisdom, understanding, truth
- aesthetic (related to artwork): beauty
- prudential/instrumental: efficiency, cleverness
- and more!

“Someone’s values”

When we talk about “someone’s values,” we may talk about

- several values they accept: for example, someone accepts justice, truth, efficiency, …, as values

- an evaluative outlook: for example, someone describes themselves as embracing a Confucian worldview

- someone’s conception of a good-life-for-them: for example, someone wants to be a doctor, have a family, live in the city, …

Alignment—Option 1: A Set of Principles?

How can replies that AI models offer to prompts be aligned with value?

How about we ask an ethicist—presumably, a specialist—to come up with a set of principles, that are then programmed into the AI model?

Alignment—Option 1: A Set of Principles?

Pro: In some domains, it may seem that sustained reflection hones intuitions on what to do in certain situations.

Contra: Any such set of principles would be dogmatic; it would be reflective of one one person’s views, even if the person is an expert.

Alignment—Option 1: A Set of Principles?

"In truth, we do not know what the principles would be for "programming in" ethics as anything like an operational system. There is continuing disagreement over the fundamental principles of ethics, and even supposing this were not so, there is sufficient distance between fundamental principles and actual applications (What constitutes a harm in a given instance?" (Railton 2020, 60)

Alignment—Option 2: Operator Intent

Should the outputs of LLMs be aligned with the evaluative outlooks of those who operate it?

Objection 1: This creates an echo chamber.

Alignment—Option 2: Operator Intent

Objection 2: What if someone intends something horrible? More generally, operator intent can come apart from “some broader notion of human values.” (Klingefjord et al 2024, p.1)

Alignment—Option 2: Operator Intent

Objection 3: If alignment with operator intent is alignment with the outlook of *any* individual user/operator, it amounts to relativism.

- What is relativism? Roughly, relativism says that all beliefs/views are true.

- Relativism seems obviously flawed, among other things because it violates the Principle of Non-Contradiction (PNC), according to which contradictories cannot both be true at the same time, of the same thing, in the same respect.

Alignment—Option 3: Pluralism?

Two further ideas that could be developed:

- Pluralism: While some evaluative outlooks are pernicious, there are several outlooks that are OK. Alignment should find a way to work with these outlooks.

- Ranked pluralism: Outlooks can collaboratively be ranked as more/less wise; AI models can prompt users, via questions, to move toward wiser outlooks (Klingefjord et al 2024).

Alignment with operator intent?

Vote 1: Should AI developers pursue alignment with operator intent?

- YES
- YES, but excluding some pernicious outlooks
- YES, but in a question-based model that aims to prompt reflection
- NO

Alignment and the Range of Traditions in Ethics

How apt are approaches in normative ethics for alignment?

- This question does not address which theory in normative ethics is most compelling.

- It only looks at the structure of theories, and asks what and how they can be implemented.

Some Traditions

There is a range of ethical theories/frameworks, including:

- Deontology
- Utilitarianism/consequentialism
- Kantian ethics
- Aristotelian virtue ethics
- Confucian ethics
- Classical Indian Buddhism

Some Traditions: Disclaimer

Disclaimer: Each of these traditions is rich and complex. Today, we only look at two of them, and only in a selective fashion.

Deontology

- Greek *to deon*: it must be done, where the "must" commands with necessity.

- The normative force of this necessity can be most salient in the negative case, where something is *absolutely prohibited*.

- Think about: It seems that absolute prohibitions can be implemented as constraints in AI models.

- Alas, pretty much no one in ethics is a deontologist! Still, some absolute prohibitions may seem to be *components* of a compelling ethical outlook.

Utilitarianism

- Utilitarianism: The right action is the action that brings the greatest happiness/utility/pleasure to the greatest number of people.

- Impartial: This utility-calculus requires impartiality. Everyone, including the agent who deliberates, counts the same.

- AI-implementation: Is it possible to calculate utility? Problem: how do we know what makes whom happy?

Consequentialism

- Consequentialism is an updated version of utilitarianism. It assumes that one cannot know what makes other people happy. Hence, one cannot calculate amounts of happiness.

- Consequentialism: The right action is the action with the best consequences, where consequences are thought of in terms of preferences, interests, etc.

- Implementation in AI: While this initially sounds promising, because it is explicitly quantificatory, it is not obvious how one should calculate consequences relative to preferences/interests.

Consequentialism as Value Monism?

- Recall pluralism about values, understood as the claim that there are several types of value, for example, ethical, epistemic, aesthetic, prudential, etc.

- Suppose that, in a given situation, the following values are involved: justice (ethical), truth (epistemic), beauty (aesthetic), efficiency (prudential). In order to quantify and calculate, these would need to translate into a common coinage.

- That is, we would need some "supervalue," a single value into which all values translate, so that they become commensurable.

Do responses to trolley cases line up with the deontology-consequentialism divide?

- Switch: You can pull a switch to redirect the trolley, so that it kills one person, rather than five. VOTE 2: would you do this?

- Footbridge: You can push someone off a bridge, so that the trolley is stopped, and you save five people. VOTE 3: would you do this?

- Both cases have a “kill the one to save the five?”-structure.

Do responses to trolley cases line up with the deontology-consequentialism divide?

- It is sometimes assumed that deontologists say “NO” in both scenarios, while consequentialists say “YES” in both.
- But far more people say “YES” in Switch than in Footbridge.
- Does this mean that agents typically don’t inhabit firm ethical outlooks? Or that their outlooks are a mix of theories?

Take-away and Questions

- Is there a place for a set of principles that someone (who?) might put together? Or is this inevitably dogmatic?

- If LLMs adopt the perspectives of users, no matter what these perspectives are, do they embrace relativism?

- What would it mean for AIs to endorse pluralism?

- Do agents inhabit ethical frameworks? Or are our outlooks typically fragmented, flexible, and a mix of different approaches? If the latter, what does this imply for AI?

Readings

- Raphaël Millière, “The Alignment Problem in Context,” arXiv (2023)

- Oliver Klingefjord, Ryan Lowe, Joe Edelman, “What are human values, and how do we align AI to them?” (2024) pp. 1-10.

- Jeremy Waldron, “Public Reason and “Justification” in the Courtroom,” Journal of Law, Philosophy and Culture (2007): 107-134. [on pluralism of evaluative outlooks]

- Elinor Mason, “Value Pluralism.” The Stanford Encyclopedia of Philosophy (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/sum2023/entries/value-pluralism/>