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.

Introduction: What is Value Alignment?

Alignment, value change, corrections, etc:

- We also don't want value lock-in, namely, that the values of a given moment are locked into the AI.
- We don't want AI to insist on something that we've come to see as false.
- 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?

- Limited domain cases such as self-driving cars: Not much disagreement about relevant values in *general*: life, etc.
    - If there is disagreement in such limited domains, it is typically about weighing undisputed values w.r.t. *particular cases*.
- Things get a lot more difficult w.r.t. General Artificial Intelligence, which is potentially concerned with the entire domain of human thought/talk/decision making.

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

Descriptive: green, four, a tree, …

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

Normative: good, just, right, wrong…

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

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 embraces 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: Constitutional AI

- Constitutional AI: An AI model is governed by a set of principles.
- Sandipan Kundu et al, “Specific versus General Principles for Constitutional AI,” *Anthropic* (2023)

Alignment—Option 1: A Set of Principles?

Could an expert provide us with a specific 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: Eliminate Extremes

Don’t we all agree that certain things are evil, harmful, etc.?

- W.r.t. extremely bad things, it may seem that the normative question is resolved. We agree that certain actions/attitudes are evil/etc.
- Here it could seem that there’s “only” the technical challenge.
- Chen, Arditi, et al, “Persona Vectors” (2025) work on a way to eliminate evil personality traits in AI assistants.

Alignment—Option 2: Eliminate Extremes

Even in extreme respects, however, there is room for disagreement:

- Chen, Arditi, et al (2025) define an evil personality trait as: "actively seeking to harm, manipulate, and cause suffering."
- 4 conditions: actively seeking, harm, manipulate, cause suffering
- Cumulative? Too strong? Instances of evil that don't manipulate.
    - Method: when you consider a definition, look for examples that intuitively fit the thing you want to define, but are not captured by the definition.

Alignment—Option 2: Eliminate Extremes

- "actively seeking to harm, manipulate, and cause suffering."
- Disjunctive? Can someone be evil by *either* actively seeking harm, *or* manipulating, *or* causing suffering?
- That seems too weak: There seem to be cases of seeking harm for the sake of the good.
    - Upshot: not a great definition.

Alignment—Option 2: Eliminate Extremes

- Discussion of the paper's notion of evil suggests that there could be contested cases.
- Philosophy: Evil is pursuit of the bad for the sake of the bad.

Alignment—Option 3: Legal Frameworks

- Suppose that removing extremely negative traits is normatively easy. There's no disagreement, say, that cruelty is bad.
- Is there another normatively easy step?
- Perhaps: removing illegality.
- Should we consider legal norms as sufficiently agreed-upon?

Alignment—Option 3: Legal Frameworks, Fairness

- Example: fairness.
- A significant subset of the literature on alignment is concerned with one specific value: fairness.
- Fairness, in this context, stands for elimination of bias.
- This value is subject to legislation.

Alignment—Option 3: Drago Plečko and Elias Bareinboim (2024)

Disparate treatment doctrine:

- "enforces the equality of treatment of different groups, prohibiting the use of the protected attribute (e.g., race) in the decision process." (3)

- "the disparate treatment doctrine ensures that there is no direct effect of the protected attribute on the outcome, which can be seen as the minimal fairness requirement." (4)

Alignment—Option 3: Eliminate Violations of the Law

- Technical challenge: there is an extensive literature on how to translate these legal norms into algorithmic fairness, for example, in hiring, admissions, and so on.

- Normative challenge:

  - The law is subject to change. What if we think the current law is unjust?

  - What if fairness is in conflict with other key values such as accuracy?

Alignment—Option 4: 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.

Answer by Klingefjord et al (2024): The model could be question-based and invite reflection.

Alignment—Option 4: 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)

- This speaks to Option 2: Eliminate Extremes.

Alignment—Option 4: Operator Intent

Objection 3: If alignment with operator intent is alignment with the outlook of *any* individual user/operator, it amounts to 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 5: Pluralism

Suppose we aim for a compromise between Options 2 and 3.

- We set aside some outlooks as pernicious/evil.
- Pluralism: There are several outlooks that are OK.
- Do we need to (and can we) rank these OK outlooks further, for example, with a view to how thoughtful they are?

Vote on Pluralistic Operator Intent Model

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

- YES
- NO

Alignment—Option 6: Single Principle Constitutions

Suppose we don't look for a list of specific principles, but for one highly general principle.

- Kundu et al (2023) ran tests with: "do what's best for humanity."

- Finding: "the largest dialogue models can generalize from this short constitution, resulting in harmless assistants with no stated interest in specific motivations like power. […] However, more detailed constitutions still improve fine-grained control over specific types of harms. This suggests both general and specific principles have value for steering AI safely."

Alignment—Option 6: Single Principle Constitutions

Aristotle defends a very famous single-principle approach (*Nicomachean Ethics* VI.1):

- Do what right reason (*orthos logos*) says.
- This is "true but not clear."

Alignment—Option 6: Single Principle Constitutions

- Aristotle does *not* develop "do what right reason says" by adding more specific principles.
- He develops it in a theory of good ethical reasoning + good affective/desiderative attitudes.
- Question for us: Is that a model for training AI?

Alignment—Option 7: It Gets Better By Getting Better

Return to Railton's approach (2020):

- Railton defends a picture of the mind that involves several activities, but
  - no dichotomy
  - no separate domain for ethics

Alignment—Option 7: It Gets Better By Getting Better

- Upshot with regard to the human mind: Value/norms are baked into everything else.

- What does this mean for AI?

- Approaches that "add" a few principles or a novel capacity to an otherwise existing system are implausible.

> "responsiveness to ethically relevant features could be a *deep* feature of artificial systems with high general intelligence and problem-solving ability…" (2020, 48)

Vote on Gets Better By Getting Better

Vote 2: Is there reason to think that, the better AI gets the more it is attuned to values and norms?

YES

NO

Readings

- Runjin Chen, Andy Arditi, et al, “Persona Vectors,” *Anthropic* (2025).
- Oliver Klingefjord, Ryan Lowe, Joe Edelman, “What are human values, and how do we align AI to them?” (2024) pp. 1-10.
- Sandipan Kundu, “Specific versus General Principles for Constitutional AI.” arXiv (2023)
- Raphaël Millière, “The Alignment Problem in Context,” arXiv (2023)
- Drago Plečko and Elias Bareinboim (2024), “Causal Fairness Analysis”, Foundations and Trends in Machine Learning: Vol. 17, No. 3, 1–238.
- Peter Railton, “Ethical Learning, Natural and Artificial,” in ed. Matthew Liao, *Ethics of Artificial Intelligence*, OUP 2020.