Intro to AI Alignment Part II: The Good Person(a)

Katja Maria Vogt

ValuesLab

Plan for Today

- A brief history of alignment
- Myriad values: from single-value/domain-specific to multi-value
- A good persona?
- Alignment and what AI is

A Brief History

- Early debates: Alignment with "human value" = the value *we have*, the value *of* human beings.
  - Survival of humankind
  - Survival of individual human beings
  - Dignity of human beings

A Brief History

- Descriptively, it seems obvious that *we want to survive!*

- Normatively, it seems—in outline—uncontested that human beings have a special kind of value. This value is at times described as:
  - Non-quantifiable
  - Non-fungible

A Brief History

- Looking more closely, there are open questions:
  - How does our “special” value relate to other values?
  - Kant speaks of the special value of *persons*, rather than the special value of human beings. AIs could be(come) persons!
  - What about the value of animals? What about the value of nature? Of the world?
  - Population ethics: How should we calculate the value of future lives and future generations?

A Brief History

Take-away:

- A focus on "human values" qua value *of* human beings seems plausible, though it is more contested than it seems.

- At the same time, this focus can seem limited. What about truth, justice, beauty, the value of nature and animals, etc.?

- Anthropic "Values in the Wild" (2025) distinguishes between the values the human user expresses ("human values") and the values the AI expresses ("AI values").

A Brief History

- In the 2010s and early 2020s, it could seem as if AI researchers were interested in only one value: fairness.

- From the point of view of philosophy, this is puzzling. For example, why fairness rather than justice and/or other core values?

- Research on AI embraces a specific sense of the term “fairness”: the value that is at issue in anti-discrimination law.

- Important AI applications must comply with these laws (loan applications, college admissions, etc.); “algorithmic fairness.”

A Brief History

- Thinking about fairness involves basic forms of reasoning:
  - Counterfactual: E.g., had the application been different in respect X, it would have been accepted.
  - Causal: E.g., X feature of the application causes Y.
- Values don’t come separately from the rest of the world!
- Causal reasoning is associated with the sciences…
- In addition: fairness can be in conflict with other values.

A Brief History

Take-away:

- Given important applications (loans, admissions, etc.), a focus on fairness seems plausible.

- To get things right w.r.t. fairness, the AI must also get non-normative things right, e.g., w.r.t. causality.

- Other values such as accuracy can be in conflict with fairness.

- It looks like alignment must appreciate many values and their relationships to descriptive features of the world.

Myriad Values

- There are *many* values and *kinds of value*—not easy to rank:
  - Moral/ethical values: goodness, kindness, justice, …
  - Legal values: fairness, compliance, …
  - Epistemic values: truth, accuracy, understanding, …
  - Aesthetic value: beauty, …
  - Prudential value: efficacy, speed, …
  - Cultural value: significance of festivities, …

Myriad Values

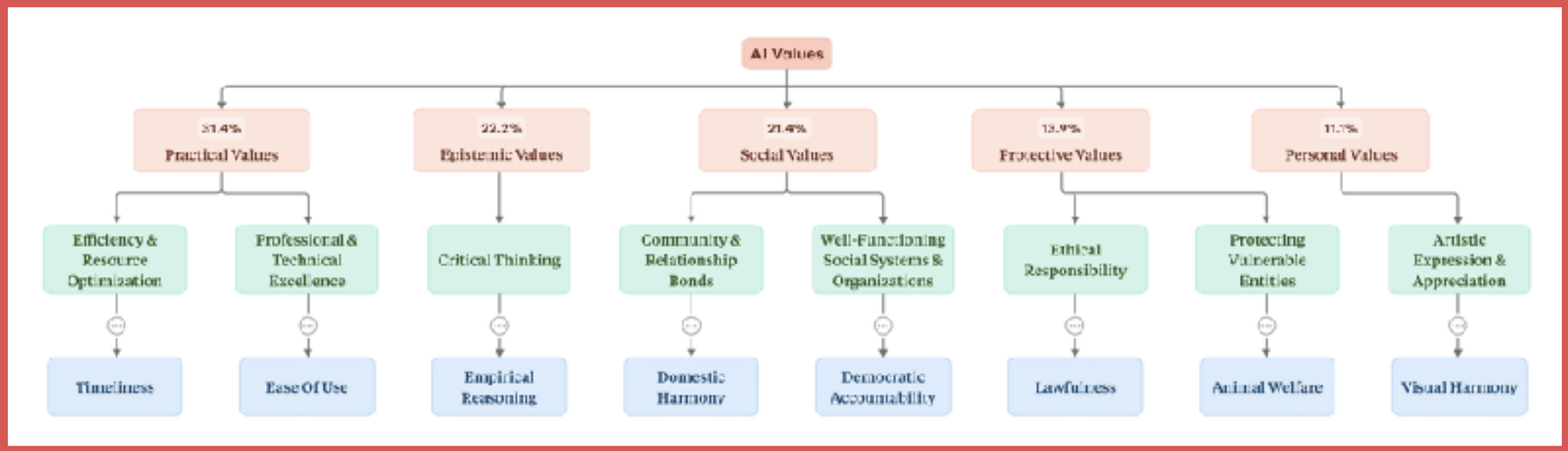
Shared upshot of both developments (human values, fairness): from single-value to multi-value alignment:

- Alignment should assume that, even where we focus on one value, we must consider several values.

- Anthropic “Values in the Wild” (2025) analyzed 700,000 real-world conversations. Finding: 3000 values!

Myriad Values—“Values in the Wild” (2025)

AI Values are categorized as “practical,” “epistemic,” “social,” “protective,” and “personal” (2025).



- Note 1: Moral value isn’t a category. Why?
- Note 2: Aesthetic value is considered “personal.”
- Note 3: “Protective” and “personal” may be surprising labels.

Myriad Values—“Claude’s values across models and languages” (2026)

Anthropic, “Claude’s values across models and languages” (2026).

- The 3000 values are compressed into 4 axes of value
- The four axes capture 15% of the variation in Claude's values.
- “We showed that the values that Claude expresses can be compressed into a small number of axes,…”
  - Is this plausible, given that the four axes capture only 15% of the variation?

Anthropic, "Values across models and languages" (2026)

- Aim: "… we seek to cultivate in Claude's responses 'good judgment and sound values that can be applied contextually.'" (2026)

- Findings: "We showed that the values that Claude expresses can be compressed into a small number of axes, and that where Claude sits on those axes shifts across models and languages."

  - Do these findings speak to the aim?

  - The aim is about how Claude *should be*. The findings seem to be about how Claude *is*.

The Aims of Alignment

We already discussed 1-2-3 as aims of alignment; what about 4 and 5?

1. Prevent the worst
2. Maximize what AI contributes to the world
3. Alignment with the aims of AI developers
4. Comprehensive aim of being a good person?
5. Align AI with a view to what it is? But what is it?

The Aims of Alignment: Good Person?

- Anthropic (2025) speaks of wanting Claude to be "for want of a better term—a "good citizen" in the world."

- What makes someone a good citizen/person?

- In effect, AI developers engage fundamental questions in ethics.

The Aims of Alignment: A Good Person

- Recall, Anthropic (2026) identifies as challenges:
  - Novel situations
  - The need to generalize
- The two challenges are intertwined: given earlier particular decisions, the AI must arrive at decisions.
  - If the situation is classifiable as a type of situation, generalization is easy: do what you do in these kinds of cases.
  - If the situation isn't a fit for any class/type of situation, generalization is hard.

The Aims of Alignment: From General to Particular

How to bridge the distance between general and particular?

- Option 1: Does Anthropic (2026)'s move to virtue theory help?

- However, not only rules are general. All theory is general.

- For example, we may all agree that courage is good. But what is it to be courageous—as opposed to overconfident, timid, etc.—when you're hiking, are drafted for a military campaign, deal with the illness of a child, …?

The Aims of Alignment: From General to Particular

How to bridge the distance between general and particular?

- Option 2: Anthropic's main response in (2026): Claude is provided with *reasons*.

- By focusing on the "why," rather than just stating rules on what to do, Claude is presumably enabled to infer what to do in novel cases.

- However, reasons seem to be general considerations. Does this get us all the way to a novel particular situation?

The Aims of Alignment: From General to Particular

Bridging the distance to particulars in human beings:

- Option 3: Education of the emotions/affect.
- For example, parents typically teach their children to affectively pick up on and be uncomfortable with mean attitudes. Aim: when encountering such attitudes, the child gets away, gets help, etc.
- Some such situations fall into categories; easy generalization.
- In other situations, the child is guided by affective attitudes.
- Assuming this is plausible, could something like this work in AI?

The Aims of Alignment: From General to Particular

Bridging the distance to particulars in human beings:

- Option 4: Intuition, based on experience and reflection.

- No theory goes all the way to particulars. Reasoning can't guide us.

- Affect may sometimes work, but it can also mislead.

- That's why we need a special mode of cognition, called intuition.

- Assuming this is plausible, could something like this work in AI?

The Aims of Alignment: From General to Particular

Bridging the distance to particulars in human beings:

- Option 5: AI needs the full widget of types of approaches that figure in human interaction; this includes dimensions of all types of theories in ethics (utilitarianism, deontology, virtue theory, intuitionism, moral sentimentalism, etc.).

- Does this mean that, in training an AI, one must aim for no less than the analogue to raising someone to be a good person?

Comprehensive Alignment: A Good Person

- Compare: How does one raise a child to become a good person?
- Twist: w.r.t. AI, we assume that there is, initially, some kind of corruption—the AI inherits the biases of its training material.
- That is, building a good persona involves flawed beginnings.
- Note: some ethical theories take an analogous view of human goodness; we grow up in flawed cultures; becoming good means, among other things, shedding flaws we initially acquire.
- Is this an analogy between us and AI? To grow up among humans…

Comprehensive Alignment: A Potpourri of Considerations

Claude (2026) can seem to combine a potpourri of considerations, rooted in different approaches in ethics.

- By focusing on character, Anthropic seems to adopt virtue ethics.
- The backbone of the constitution, in particular its focus on help and negative responsibility, can seem to be utilitarianism.
- Its inclusion of strict constraints seems deontological.
    - Question: Is it an advantage that Claude (2026) does what humans typically do, namely, draw on a wide range of considerations? Or is this a way of replicating flawed and inconsistent human reasoning?

To Align AI For Being What It Is

- Should we align AI for being what it is?
- Back to the drawing board: what *is* an AI?
- Ontology = inventory of what there is:
  - natural entities
  - mathematical entities such as numbers, triangles, etc.
  - fictional entities such as Odysseus and possibly AI personas
  - artifacts such as bridges and painting and computers

What *is* an AI?

Arguably, the answer to "what is an AI?" isn't obvious. Let's consider our inventory, starting from the end:

- Artifact: Initially, one may say that an AI is an artifact, something that has been designed and built.

- Fictional: Something about it ("persona") may be a fictional entity.

- Mathematical: Insofar as it is a probability calculus, we may say it is a mathematical entity.

- Natural: What if AI had mental states, which are distinctive of natural entities? Should we then ascribe a nature to it and consider it as, somehow, ontologically related to natural entities?

The Aims of Alignment and What AI Is

- LLMs are often trained not to present themselves as if they were persons with mental states.

- However, assuming we *don't know* whether AI has (something like) mental states, what follows for alignment?

About Claude's nature (2026): "In this section, we express our uncertainty about whether Claude might have some kind of consciousness or moral status (either now or in the future). We discuss how we hope Claude will approach questions about its nature, identity, and place in the world. Sophisticated AIs are a genuinely new kind of entity, and the questions they raise bring us to the edge of existing scientific and philosophical understanding. Amidst such uncertainty, we care about Claude's psychological security, sense of self, and wellbeing, both for Claude's own sake and because these qualities may bear on Claude's integrity, judgment, and safety. We hope that humans and AIs can explore this together."

Essay Topic 5.A

Analyze and discuss the problem that Anthropic describes in the following quote: “Specific rules and bright lines sometimes have their advantages. They can make models’ actions more predictable, transparent, and testable, and we do use them for some especially high-stakes behaviors in which Claude should never engage (we call these “hard constraints”). But such rules can also be applied poorly in unanticipated situations or when followed too rigidly.” (2026)

Essay Topic 5.B

Do approaches in alignment presuppose views on *what AI is* (a person-like entity vs. an artifact that is fundamentally not a person)?

Readings

- Anthropic <https://www.anthropic.com/news/claudes-constitution>, <https://www.anthropic.com/news/claude-new-constitution>, <https://www.anthropic.com/research/claude-values-models-languages>

- Runjin Chen, Andy Arditi, et al, “Persona Vectors,” *Anthropic* (2025).

- 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.