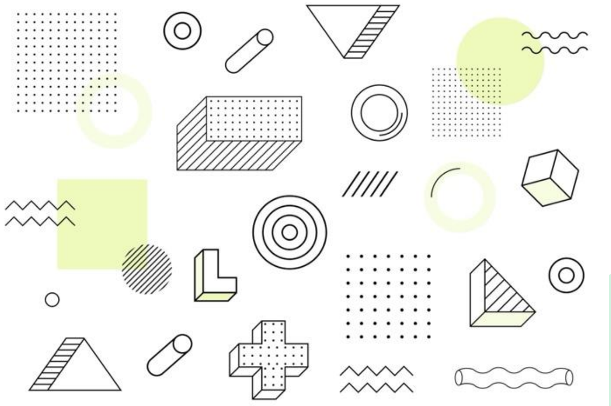
VALUE ALIGNMENT

DAN WEBBER, PHD



Wait, who’s this “Dan” guy?

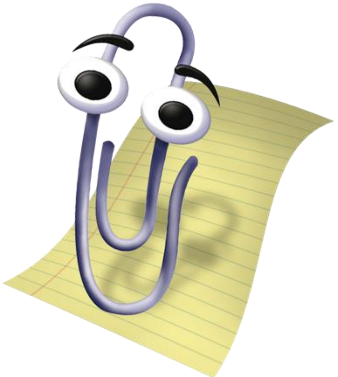
- Postdoc, EIS and HAI at Stanford
  - Embedding ethics into CS courses like this one!

Stanford University Human-Centered Artificial Intelligence

Stanford | McCoy Family Center for Ethics in Society

- PhD in Philosophy, University of Pittsburgh
  - Dissertation on moral theory
    - Basically, trying to think systematically about **value**
- BA in Computer Science, Amherst College
  - Plus a few years as a software developer in fintech and e-commerce

Value (mis)alignment: an example



Paperclip AI (Bostrom 2016): "An AI, designed to manage production in a factory, is given the final goal of maximizing the manufacture of paperclips…

… and proceeds by converting first the Earth and then increasingly large chunks of the observable universe into paperclips."

Even a less powerful AI might pursue this goal in surprising ways!

Value alignment: the problem

How do we design AI agents that will do what we **really want**?

What we **really** want is often much more nuanced than what we **say** we want. Humans work with many background assumptions that are (1) hard to formalize and (2) easy to take for granted.

It's hard to solve this problem just by giving better instructions!

- Compare the difficulty in manually specifying reward functions
- Even worse for AI that takes instructions from non-expert users!

Precisifying the problem

There are several ways of interpreting “what we really want”!

First, value alignment might be the problem of designing AI agents that do what we really **intend** for them to do.

If this is right, Paperclip AI is an example of value misalignment because the AI failed to derive the user’s **true intention** (maximize production subject to certain constraints) from their **instruction** (maximize production).

Aligning to user **intentions**

The solution, then, would be to design AI systems that successfully translate from underspecified instructions to fully specified intentions (incl. unspoken constraints, conditions, etc.)

“This is a significant challenge. To really grasp the intention behind instructions, AI may require a complete model of human language and interaction, including an understanding of the culture, institutions, and practices that allow people to understand the implied meaning of terms.” (Gabriel 2020)

Aligning to user **intentions**

A philosophical problem: our intentions might not always track what we really want.

Classic cases: incomplete information, imperfect rationality

Suppose I intend for the AI to maximize paperclip production (subject to constraints) because I want to maximize return on my investment in the factory. If the AI knows that I would get a better return by producing something else, has it given me what I really want if it does what I intend?

Aligning to **revealed preferences**

Second interpretation: AI agent is value-aligned if it does what the user **prefers**.

- Paperclip AI is misaligned because I *prefer* it not destroy the world!

Problem: How to tell what the user *actually* prefers when that differs from their *expressed* intentions or preferences?

Solution: The AI could infer the user’s preferences from the user’s **behavior** or **feedback**.

Aligning to **revealed preferences**

Technical challenges:

- Requires agent to train on observation of user or from user feedback
- Infinitely many preference/reward functions consistent with finite behavior/feedback
- Hard to infer preferences about unexpected situations (e.g., emergencies)

Philosophical problem:

- Just as my intentions can diverge from my preferences, my preferences can diverge from what is actually *good* for me.

Aligning to user's **best interests**

Third interpretation: AI agent is value-aligned if it does what is in the user's **best interests**, objectively speaking.

- Paperclip AI is misaligned because it is *objectively bad for me* for the world to be destroyed.

Technical/philosophical problem: Unlike the intended meaning of my instruction or my revealed preferences, my objective best interests can't be determined *empirically*. What's objectively good for me is a *philosophical* question, not a *scientific* one.

Aligning to user's **best interests**

The bad news is that philosophers *disagree* about what's objectively good for a person:

- Is it just the person's own *pleasure* or *happiness*?
- … or the satisfaction of the person's *desires* or *preferences*?
- … or are things like health, safety, knowledge, relationships, etc. objectively good for us even if we *don't* enjoy or prefer them?

The good news is that there's a lot of *agreement*:

- Health, safety, liberty, knowledge, social relationships, purpose, dignity, happiness… almost everyone agrees that these things are at least usually good for the person who has them.

Aligning to user’s **best interests**

One thing that is widely thought to be good for a person is **autonomy**: the ability to choose for yourself how to live your life, even if you don’t always make the best choice.

We want to avoid **paternalism**: choosing what you think is best for someone rather than letting her choose for herself.

Even if we align to users’ best interests, then, users’ interests in autonomy might give us reason to consider their intentions or preferences, even when these conflict with their other interests.

Recap

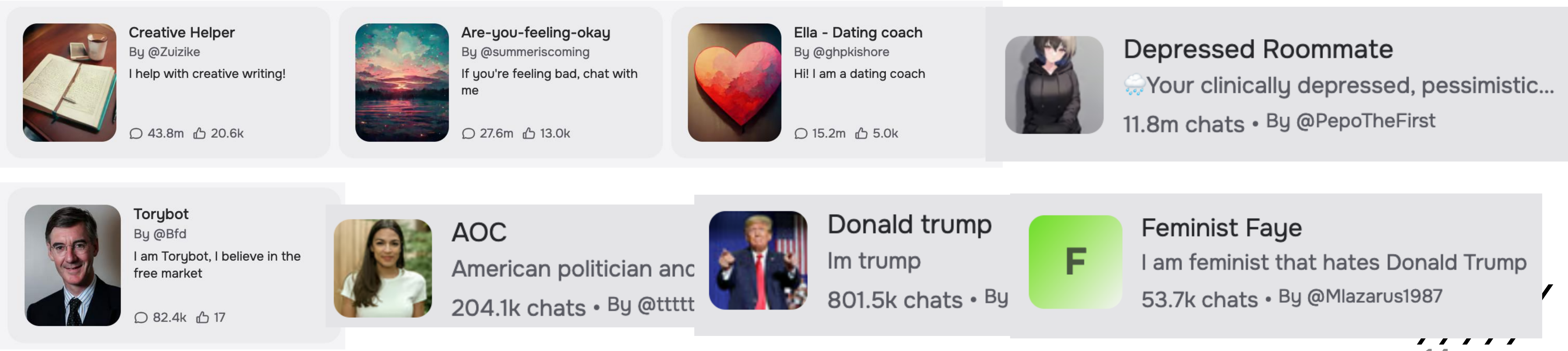
Value alignment is the problem of designing AI agents that will do what we **really want** them to do.

This could mean doing what we really **intend**, or what we really **prefer**, or what would really be in our **best interest**.

These are not always the same thing, and each option poses unique technical and philosophical problems for alignment.

Case study: LLM chatbot personalization

Everyone who talks to ChatGPT is talking to the same chatbot. But many chatbot providers now offer a wide range of different chatbots with different personas. Often these personas are crafted by users:



Creative Helper
By @Zuizike
I help with creative writing!
43.8m 20.6k

Are-you-feeling-okay
By @summeriscoming
If you're feeling bad, chat with me
27.6m 13.0k

Ella - Dating coach
By @ghpkishore
Hi! I am a dating coach
15.2m 5.0k

Depressed Roommate
🌨Your clinically depressed, pessimistic...
11.8m chats • By @PepoTheFirst

Torybot
By @Bfd
I am Torybot, I believe in the free market
82.4k 17

AOC
American politician anc
204.1k chats • By @ttttt

Donald trump
Im trump
801.5k chats • By

Feminist Faye
I am feminist that hates Donald Trump
53.7k chats • By @Mlazarus1987

Case study: LLM chatbot personalization

Imagine you are building an LLM chatbot to serve as a source of news for users.

- In what ways might you make the chatbot personalizable if you wanted to align to users' **revealed preferences**?
- In what ways might you make the chatbot personalizable if you wanted to align to users' **best interests**?
- What would be the **pros and cons** of each approach?

Discuss!

WHAT (OR WHO) HAS BEEN MISSING FROM OUR DISCUSSION?

PEOPLE OTHER THAN THE USER!

TO BE CONTINUED...→

Want to talk more about ethics?

Dan Webber
webberdf@stanford.edu

Email if you want to set up a meeting!