VALUE ALIGNMENT

PART II

DAN WEBBER, PHD

Recap of last time

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**.

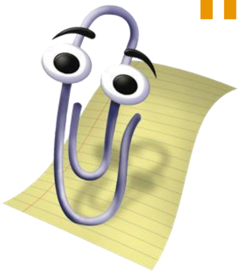
We discussed **chatbot personalization** as a case study where these different alignment targets can pull in different directions.

WHAT WAS MISSING FROM OUR PREVIOUS DISCUSSION?

PEOPLE OTHER THAN THE USER!

Aligning to **social value** or **morality**

Fourth interpretation: AI agent is value-aligned if it does what is **morally right**.



- Paperclip AI is misaligned because it's bad *for everyone* if the world is destroyed!

This interpretation emphasizes the **we** in "what we really want."

What the user intends, prefers, or even what's in her interest might be bad for others!

The user still matters

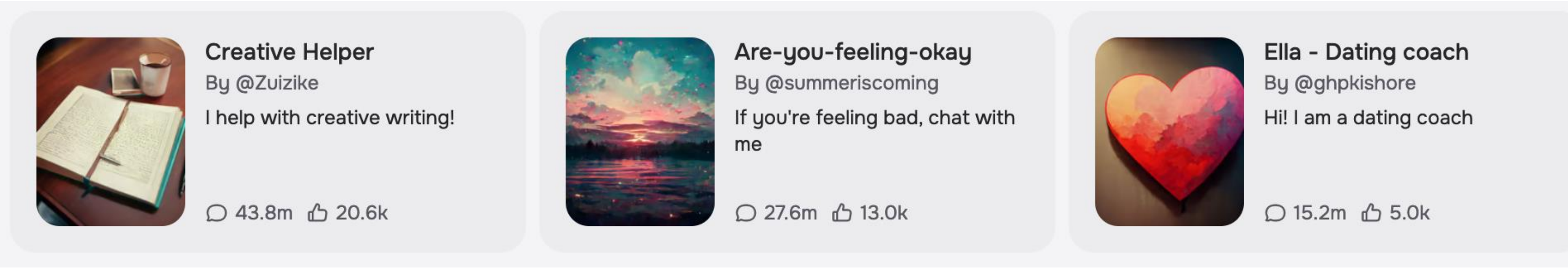
But it wasn’t just a waste of time to start by focusing on the user!

Even though we want to align to morality, we also want to align to what the user wants when what the user wants is morally acceptable.

So it still matters how we think about what the **user** really wants, even if we need to think about it in the **larger ethical context**.

Case study: LLM chatbot personalization

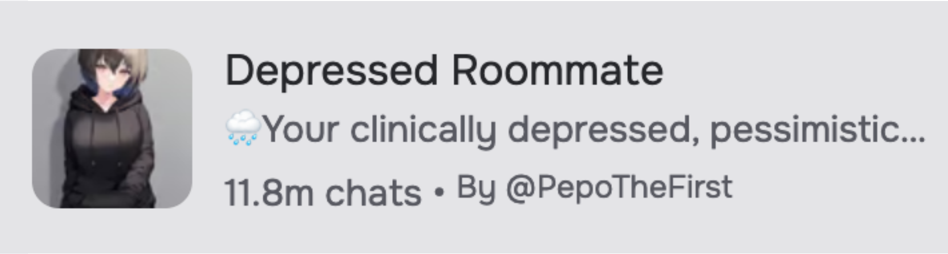
How might you approach personalization for your news chatbot if your alignment target is *social* or *moral* value?



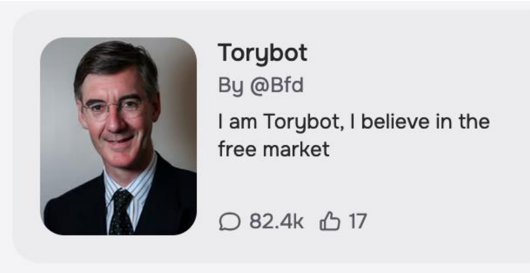
Creative Helper
By @Zuizike
I help with creative writing!
43.8m 20.6k

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

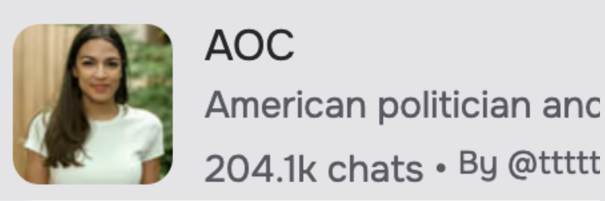
Ella - Dating coach
By @ghpkishore
Hi! I am a dating coach
15.2m 5.0k



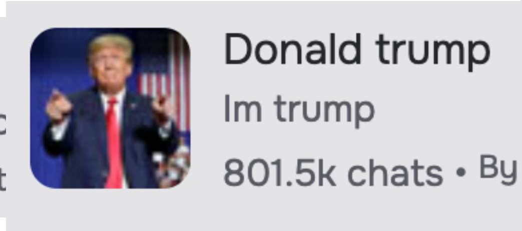
Depressed Roommate
🌧️Your clinically depressed, pessimistic...
11.8m chats • By @PepoTheFirst



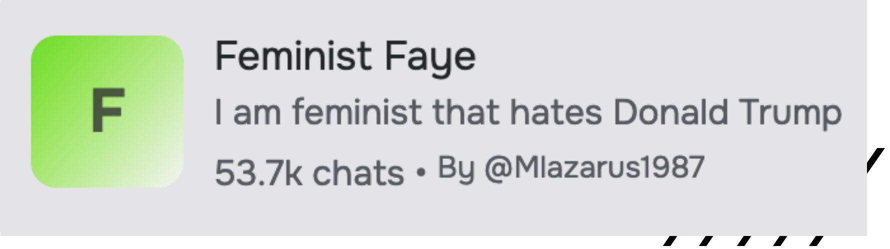
Torybot
By @Bfd
I am Torybot, I believe in the free market
82.4k 17



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

Aligning to **morality**: **top-down**

Top-down approach: Explicitly formulate moral principle(s) to align to.

- Try to ensure alignment via reward function, post-processing, etc.

Philosophical problem: What are the correct moral principle(s)?

- We don’t know! This is an open problem in moral theory.

Utilitarianism: Maximize total net happiness over all people.

- What about the *distribution* of happiness? What about *rights*?

Aligning to **morality**: **top-down**

Common-sense pluralism: Many different moral principles.

- "Don't lie," "Don't steal," "Don't hurt people," "Keep promises," etc.
- But what about when the principles conflict? What about (highly nuanced) exceptions?

Moral "reward hacking": Incorrectly specified moral principles can recommend surprising forms of bad behavior.

- What's a surprising way that a utilitarian AI agent might learn to maximize total net happiness over all people?

Aligning to **morality**: **bottom-up**

Bottom-up approach: Don't explicitly formulate principles; learn morality by example.

- e.g., through inverse RL, imitation learning, or RLHF

Philosophical problem: *moral disagreement*

- Whose example?
- Should ChatGPT produce depictions of the prophet Muhammad? Offer tips for evading law enforcement? Depends who you ask!
- Some cases generate disagreement because they are *hard*.

Aligning to **morality**: **bottom-up**

Technical problem: *rare* or *unforeseen* cases

- Self-driving car trained on real-world human driving might never see examples of how to respond to deadly brake failure.
- Gap in moral "understanding" if AI agent extrapolates incorrectly.

Takeaways for **moral** value alignment

- No silver bullet to guarantee *perfectly* moral behavior.
- But alignment can be *better* or *worse*. For better alignment:
  - Start with easy stuff that (almost) everyone agrees on…
    - Your AI should avoid killing people! It (usually) shouldn't lie, etc.
  - … but do your best to capture the complexities too.
    - **Top-down**: Think hard about principles, conflicts, exceptions.
    - **Bottom-up**: Get creative; train on as many rare/edge cases as you can imagine.

Embedded Ethics survey!

- Coming soon, be on the lookout
- Your thoughts on Embedded Ethics in your current courses
- First 800 participants get $10 gift card
- Whether you choose to participate or not will **not** affect your grade in any way

Want to talk more about ethics?

Dan Webber
webberdf@stanford.edu

Email if you want to set up a meeting!