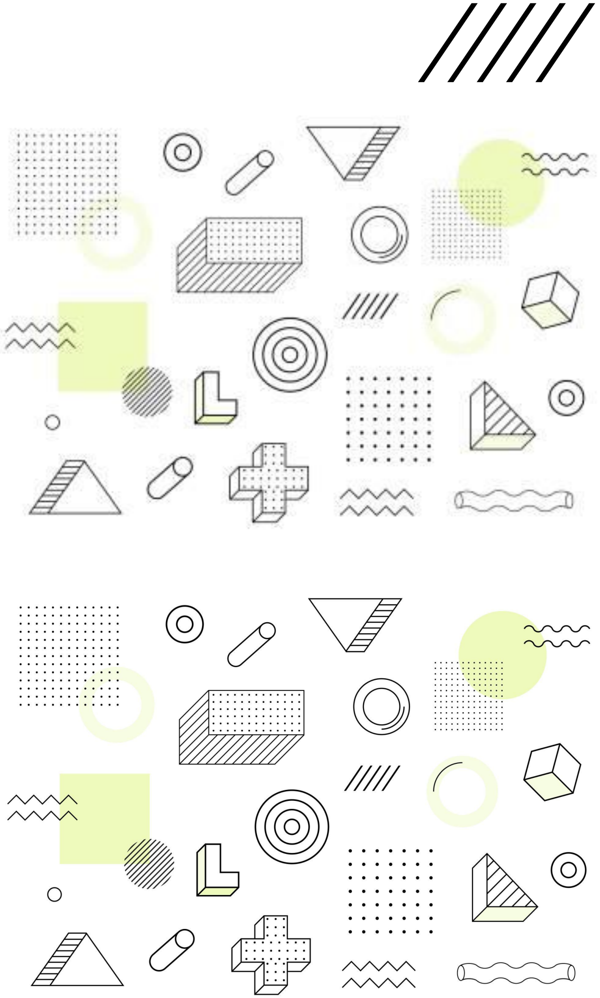
VALUE ALIGNMENT

PART II

Wanheng Hu, Ph.D.

Acknowledgement: This lecture is based on materials originally developed by Dan Webber, with valuable input from Andy Ouyang.

CS234 Winter 2026



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.

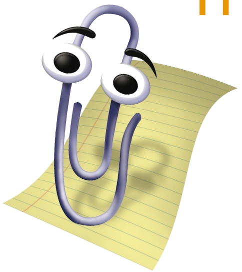
We discussed **Sycophancy in AI** and **Agentic AI** as case studies where these different alignment targets can pull in different directions.

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: Agentic AI

Imagine you have a personal AI agent to buy concert tickets at the best price. The agent monitors markets, negotiates, and purchases automatically. Many others also use agents with the same goal.

Discussion: What happens when everyone has an agent optimizing on their behalf?

Individually aligned actions can lead to:

- Faster competition between agents
- Price spikes
- Unequal advantages across users
- …

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?

Case study: Top Down Agentic AI

Possible approaches

- Hard constraints (e.g., “do not manipulate/mislead other agents”)
- Global objectives (e.g., minimize price inflation, promote fairness)

Challenges

- Which principles should be enforced — fairness, efficiency, profit?
- How should conflicts between users be resolved?
- Risk of reward hacking

Top Down Limitations

Difficult to create a rule set that:

- Covers all possible situations
- Handles edge cases and exceptions

Moral rules often conflict with one another

Risk of oversimplification

Hard to capture individual nuance

Alignment Issues

In the Real World

NATURE MACHINE INTELLIGENCE

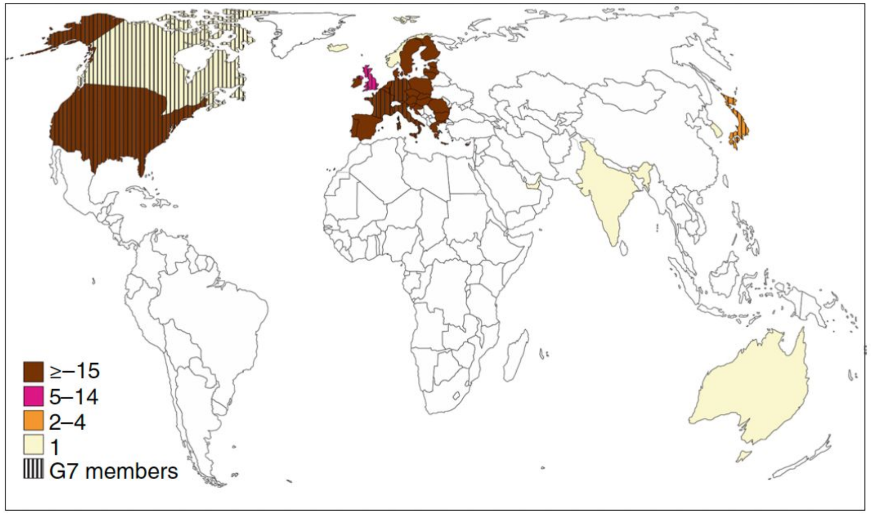


Fig. 2 | Geographic distribution of issuers of ethical AI guidelines by number of documents released. Most ethics guidelines are released in the United States (*n* = 21) and within the European Union (19), followed by the United Kingdom (13) and Japan (4). Canada, Iceland, Norway, the United Arab Emirates, India, Singapore, South Korea and Australia are represented with 1 document each. Having endorsed a distinct G7 statement, member states of the G7 countries are highlighted separately. Map created using https://d-maps.com/carte.php?num_car=13181.

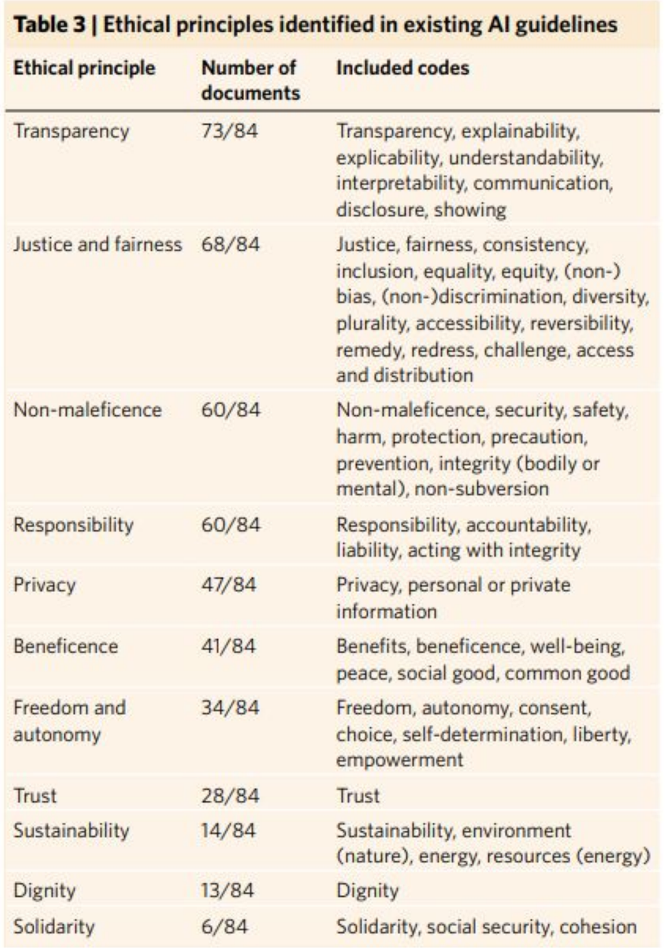


Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

(Jobin et al., 2019)

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.

Normative challenge: *representation*

- Moral disagreement across individuals and cultures
- Whose examples are used?
- Majority behavior may dominate minority values

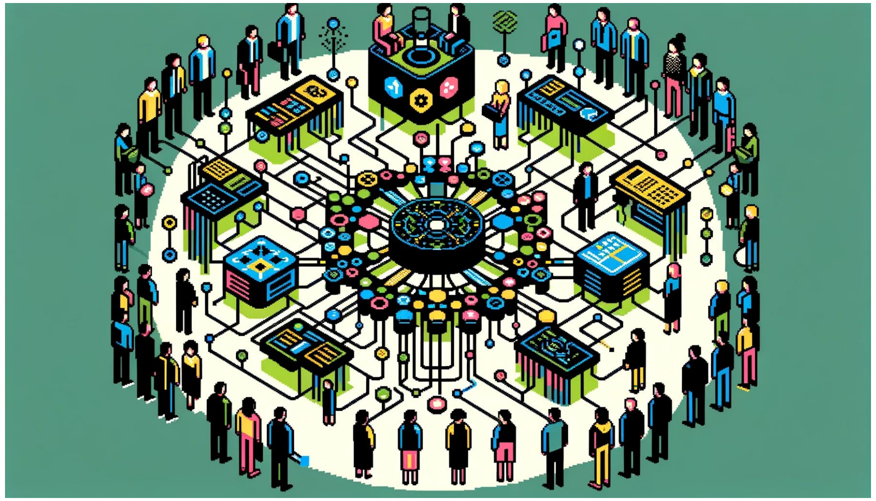
Participatory AI

Expands “learning from humans” to include:

- Multiple stakeholders
- Affected non-users
- Ongoing input, not one-time training

Values treated as

- contextual
- contestable
- revisable over time



Participatory AI in Practice

Examples

- Community advisory boards for AI systems
- Ongoing user feedback channels
- Public consultations before deployment

Key ethical considerations

- Recognition that values can conflict
- Willingness to revise systems over time

Case study: Bottom Up Agentic AI

Possible approaches

- Learn from user feedback on ticket purchases
- Train on historical ticket market data
- Adapt based on observed agent-to-agent interactions

Challenges

- Which norms get learned (speed, profit, fairness)?
- Majority user behavior may dominate

Bottom Up Limitations

Learns values from data, feedback, or examples

Feedback is often unevenly distributed

- Majority groups are more likely to be represented
- Minority or marginalized groups may be underrepresented

Learned values may reflect existing social biases

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.

Want to talk more about ethics?

Wanheng Hu
wanhenghu@stanford.edu

Email if you want to set up a meeting!