Moral Disagreement and Artificial Intelligence

Pamela Robinson, Australian National University

*Moral disagreement is **especially challenging** because it's unclear whether it calls for a political or an epistemic solution.*

preference disagreements, in which people have conflicting preferences → usually call for political solutions, which aim at a fair compromise.

descriptive disagreements, in which people disagree over descriptive facts → usually call for epistemic solutions, which aim at the truth.

moral disagreements, in which people disagree over moral facts → might call for one or the other kind of solution.

- political solutions are most popular among current AI ethics researchers.
- but examples of both solutions can be found.

How do we choose between them?

imagine that we've built, or are about to build, an 'AI Decider'

When an AI's decisions will affect people who disagree about relevant moral facts:

should we design AI to aim at mutual acceptance?

or should we design AI to aim at the moral truth?

***Choosing between** political and epistemic solutions to moral disagreement…*

Some potential grounds to make the choice:

pragmatic grounds

- *mutual acceptance*
- *predictability*
- *safety*

moral grounds

- *procedural justice*
- *proximity to moral truth*
- *metaethical disagreement*

but both solutions can be defended on these grounds.

I argue that the choice between political and epistemic solutions is ultimately a choice between **morally risky design choices**.

- building an AI Decider is *never free from moral risk*.
- adopting one solution over the other takes a stand on which is less morally risky.

This work aims to explain the **problem** posed by moral disagreement for designing moral and value-aligned AI.

Next step: which kind of solution is least morally risky?

Australian National University

HMI