ALIGNMENT WITH WHAT VALUES?

Kalle Grill

2026-03-11

UMEÅ UNIVERSITY

OUTLINE

Machine value alignment

What/whose values?

Wants vs. interests

Current vs. future wants (and interests)

Individual vs. group wants and interests

Promoting alignment

THREE KINDS OF ETHICAL/MORAL CONCERNS

Systemic effects

- Environmental
- Economical
- Political
- Unintended malign use

Process values

- Transparency
- Explainability
- Safety

Goal values

- Main aim/purpose

VALUE ALIGNMENT

- Standard system alignment:
  1) Behavior aligned with the goals designed for
  2) Behavior aligned with the immediate task-related goals of users
- Current recommender/autoplay systems: Aligned with user revealed preferences/choices, based on population statistics
- Superintelligence scenario: Future super powerful AGI have values instilled that align with human interests, the objective good, etc.
  - Norbert Wiener (1960): "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere... we had better be quite sure that the purpose put into the machine is the purpose which we really desire."

SOME BENEVOLENT ALIGNMENT GOALS

"ensure that artificial general intelligence benefits all of humanity" [OpenAI]

"ensure transformative AI helps people and society flourish" [Anthropic]

"ensure that the world safely makes the transition through transformative AI" [Anthropic]

"solving intelligence, to advance science and benefit humanity" [DeepMind]

"build AI responsibly to benefit humanity" [DeepMind]

"reorient the general thrust of AI research towards provably beneficial systems" [Center for Human-Compatible AI]

WHOSE VALUES?

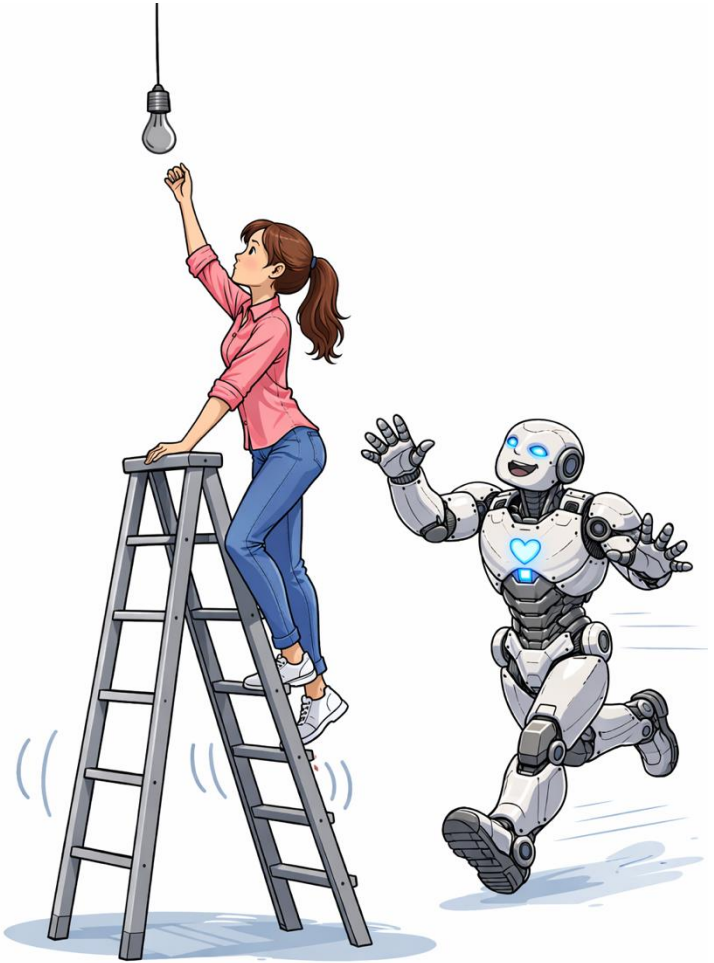
- Designer / Creator
- User
  - Human being using/prompting/paying (who are the users of ”free” ad-financed software?)
- Impacted
  - Human beings affected by use – e.g., applicants, customers, the surveyed…
- Society
  - Local or national
- Humanity
- The objective good
  - Beyond just human beings? Nature? Future sentient artificial beings?

SOME POTENTIAL CONFLICTS

- Wants vs. Interests
- Current vs. future wants
- Individual vs. collective

BENEVOLENT AI

If we could create a machine that would optimally promote our interests, would we want it?



Picture generated by ChatGPT

WANTS AND INTERESTS

What I want ≠ what is good for me

=> Moral problem of paternalism:

- People do not always optimize their own interest
- Sometimes we can interfere/intervene/support to make them do so more
- Should we? When? How?

WHAT WE WANT

- What we think is good for us as individuals
- What we think is good for us as a (global) society
- What we value, including other things than the human good (such as beauty)
- Many other things, because we are irrational, weak-willed, impulsive, etc.

=> Many disagreements and unsolved problems

PATERNALISTIC TECH/AI

Old: Car belt warning signal; safe chain-saws

New: Chatbots that refuse to facilitate self-harm

Near future: Household AI systems that regulate food intake, media consumption, etc.

Far future: Superintelligent “nannies” that oversee, regulate or design all parts of our lives

THREE (MAIN) INPUT DIMENSIONS

- Explicit (interpreted) commands
- Assumed or deduced interests
  - General/generic
  - Individual/personal
- Deduced ("revealed") preferences
  - Current
  - Future
- Which dimension is more easily approximated and by what general approaches? For example: what information is generated by a stated preference? About preference? About interests?

PRUDENT AI

If we could create a machine that would optimally promote our wants, over time, would we want it?

Picture generated by ChatGPT

CURRENT VS. FUTURE WANTS

- What we want changes over time, sometimes radically
- We are myopic - we prefer pleasures soon and pains later

$\Rightarrow$ We would not want, now, to optimize our want satisfaction over time

- It is an important part of our autonomy to have some control over our own future

$\Rightarrow$ We have some moral right to prioritize our current wants

- Should a prudence machine optimize the aggregate of future-oriented wants at each time, or only the now-oriented wants?

CURRENT VS. FUTURE INTERESTS

Should we care equally about all our future states?

There is diminishing psychological connection and overlap with ourselves as we extend into the future

WAYS OF IDENTIFYING HUMAN INTERESTS

- Subjective reports
  - Life satisfaction
  - Momentary mood
  - Before/after various events/interventions
- Objective measures
  - For health/well-being: QUALYs, willingness to pay, etc.
  - Other metrics: Education levels, income, crime, employment, etc.

PRUDENT TECH/AI

Current: Warnings (“do you really…”), health and exercise wearables, backup by default, screen time limits, version history (?)

Future: Personal assistant predicting preference change

COORDINATING AI

If we could create a machine that would optimally promote the aggregate of human preferences, who would want it?

Picture generated by ChatGPT

INDIVIDUALS VS. GROUPS

- Coordination problems – conditional preferences
  - Ignorance and irrationality of individuals magnified by strategic preferences
- No neutral way to aggregate preferences over individuals (social choice theory)

"an open problem" –Russel 2021

OTHER-REGARDING PREFERENCES

- What we need and want for ourselves requires the cooperation of others
- People want things for others (positive and negative!)

$\Rightarrow$ Unequal weight of interests

- Some people's lives have extreme importance

COLLECTIVE INTERESTS – SOME FURTHER PROBLEMS

- Distribution
  - Priority for those… worse off? more deserving?
- Future/potential people (and other moral subjects?)

COORDINATING TECH/AI

Current: Safety features for protecting others, chatbot-supported AI-human expert systems (for medical diagnosis etc.)

Future: Aggregation of (global) human preference in various domains, by different aggregation principles

OPPORTUNITIES OF POWERFUL AI?

Reduce influence of irrationality and ignorance

Solve coordination problems

Lock in agreements

STRATEGIC CONSIDERATIONS GOING FORWARD

1. Determining values
2. Aligning ourselves
3. Changing goals

ABILITY TO DETERMINE VALUES VARIES

- Current wants are easier to approximate than future wants
- Machines that observe us will more easily approximate what we want than what we believe is good for us, because of our other-regarding preferences
- General models of rational behavior will more easily identify what is good for us than what we want or prefer, because human wants and preferences are more diverse than human interests

$\Rightarrow$ Research focus will affect what values/dimensions get more priority

ALIGNING OURSELVES

- Individually
  - More informed/educated
  - More prudent
- Collectively
  - More tolerant
  - More altruistic

SIMPLIFYING ALIGNMENT BY CHANGING GOALS

Ways to optimize future interest and want satisfaction:

- Humble interests, easily fulfilled
- Simple preferences, easily satisfied

Could potentially be achieved without unpleasantness and without (much obvious) interference

MAIN TAKEAWAYS (?)

1. Being benevolent is not as straightforward as it may seem: there are many different values associated with benevolence
2. Each of these values have inbuilt tensions and can be interpreted in different ways
3. Humanity and technology/AI can become more aligned with each other and with values also by changes to society/humanity