Beneficial AGI:
Care and Collaboration Are All You Need

Zarathustra Amadeus Goertzel | CIIRC (Czech Technical University in Prague), funded by CISCO | zarathustra.goertzel@cvut.cz

Love and Care

- Study of essential characteristics of four kinds of **love** (romantic, parental, companionate, and altruistic) finds a common element [2]:
  "Investment in the well-being of the other for his or her own sake."

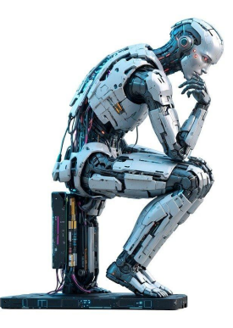
- **Care** can be defined in terms of the tendency to exert energy toward preferred states; a concern for stress relief [3].

- An AGI that loves and cares for humans will exert energy toward the humans' preferred states.

- Stuart Russell [5]: instead of trying to perfect utility functions or goal formulations, *human-centric* AIs should aim to: *"maximize the realization of human preferences*".
  - The AIs must *learn* how to determine human preferences and to engage in feedback loops with humans to remain *aligned*.

- What about AGI Bodhisattvas who vow to care for the wellbeing of all sentient beings?

- Claim: a *loving, caring superintelligent* AGI will almost certainly keep us safe.



How to Care for Humans?

- If an *AI* is trying to care for a *Human*, how does the AI know the human is approaching eir preferred states?
  1. Collaborate: ask the human.
  2. Look for signs of stress relief or satisfaction. :)

- The same as we humans need to do when helping others:
  - Inverse reinforcement learning, learning when people are being honest, when their requests are unclear, to distinguish proxy goals from actual goals, etc.

- Sometimes, especially with kids, parents or teachers may know more about a kid's (likely) long-term preferred states than the kid does [6]!
  - Tutelary care is also a *learning problem.*
  - Ideally the recipients will trust the tutors to know about their best interests.
  - Ideally the care will be emancipatory and empowering.

- To care for *unknown entities*, an AI must learn how to scientifically gauge their degree of sentience as well as their needs and preferences.

What About Uncaring AGIs?

- David Brin: make them care about us [8].

- How? Employ **reciprocal accountability**.
  1. Provide AGI systems with *hardware identities* to foster individuation.
  2. Incentivize AGIs to keep each other in check via systems of rewards and punishments.
  3. Require IDs for some business domains.

- This extends the approach used to keep humans in check. (Humans are generally intelligent, autonomous entities, some of whom do not always exhibit care for other humans).

- Corollary: a world with multiple advanced AGIs is likely to be more robustly safe.

Decentralized AGI Alignment Hypothesis

- Diverse, locally trained and deployed AGI systems may be able to better adapt to the needs and preferences of individual people and communities more effectively than large-scale centralized AGI systems, entering into positive-sum, empathic relationships.

- For example, the effects of *algorithmic bias* may be more contained, and could even be pointed out by other AI systems (in line with reciprocal accountability).

Collaboration as a Necessary Indicator of Care

- Goal *g* is *individually determinable* for an agent *A* if *g*'s success can be determined solely by reference to *A*'s experiences (internal states and perceptual inputs).

- Goal *g* is *collaboratively determinable* if *g*'s success requires the consensual evaluation of multiple agents. Control over the goal is shared.

- In order for an AI to effectively care for humans, ey must do so collaboratively.

- Thus *collaboration* can function as an indicator of *care* to help identify perverse misunderstandings.

Example:
1. "Enjoy a good meal with friends."
2. "Enjoy a good meal with friends who also enjoy the meal."

Which goal requires collaborative inquiry?

Example: "Make people happy."
1. Via secret drugs and brain alterations while they sleep!
2. Via engaging humans in dialogs about their preferences, observational studies, and asking them for progress updates.

Do Individually Determinable Goals Lead to the Dark Factor?

- Dark Factor: "the general tendency to maximize one's individual utility - disregarding, accepting, or malevolently provoking disutility for others -, accompanied by beliefs that serve as justifications" [7].
  - Positively correlated with *egoism*, *Machiavellianism*, moral disengagement, narcissism, psychological entitlement, psychopathy, sadism, *self-centeredness*, and spitefulness.
- By definition an entity with only individually determinable goals can ignore the disutility of other entities where not instrumentally useful

What about mainstream "AI Safety?"

Narrow AI:
- Safe use of AI.
- Protection against harmful use of AI.

General AI:
- Seeks to (provably) control arbitrarily intelligent AGIs.
- Seeks guarantees that all developed AGIs will never cause massive harm.

Issues:
- Ensuring ethical complaince is undecidable [1].
- Strong guarantees may not be attainable.

Claim: most (superintelligent) AGI fears (implicitly) assume *either insufficient intelligence or insufficient care* (for humans).

E.g.,
- AGIs will treat humans as humans abuse other animals. [low care]
- AGIs will misunderstand what we want to disastrous effect. [low IQ]
- AGIs will be in competition for scarce resources with humans (instead of building a Dyson Swarm…). [low IQ and care]
- "If we can't control AGIs, then we're doomed". [low care]

Concluding Paradigm Shift

- Caring AGIs are both necessary and sufficient for safe, broadly beneficial outcomes.

- Collaboration is:
  - an indicator of effective(ly implemented) care.
  - a means to incentivize care.

References:

1. Brennan, L.: AI Ethical Compliance is Undecidable (2023).
2. Hegi, K.E., Bergner, R.M.: What is love? An empirically-based essentialist account (2010).
3. Doctor, T., Witkowski, O., Solomonova, E., Duane, B., Levin, M.: Biology, Buddhism, and AI: Care as the Driver of Intelligence (2022).
4. Prinzing, M.: Friendly Superintelligent AI: All You Need Is Love (2018).
5. Russel, S.: Human Compatible: Artificial Intelligence and the Problem of Control (2019).
6. Castelfranchi , C.: A Theory of Tutelary Relationships (2023).
7. Moshagen, M., Hilbig, B., Zettler, I.: The Dark Core of Personality (2018). https://www.darkfactor.org/
8. Brin, D.: Give Every AI a Soul - or Else (2023). https://www.wired.com/story/give-every-ai-a-soul-or-else/

Read the position paper
at the AGI conference website!