Towards Ecological Machine Ethics

Eryn Rigley supervised by Prof. Adriane Chapman, Dr Christine Evers, and Dr Will McNeill
University of Southampton. Funded by DSTL

Machine ethicists aim to equip machines with ethical decision making capabilities. Some argue these 'ethical' machines ought to be aligned with human values...

...However, humans commonly disregard the wellbeing of others, which leads to the possibility of human-aligned AI disregarding others. Scoping review results show human-centric AI ethics is especially threatening to the wellbeing of nonhuman animals and the environment...

Scoping Review of AI Ethics Frameworks Methodology

1. Identifying the research question: to what extent does the AI ethics landscape conform and comply with anthropocentrism (human-centredness) and environmental wellbeing?

2. Defining and identifying relevant frameworks: non-technical tools for ensuring compliance and conformance e.g., regulation, codes of conduct, and governance (European Commission 2018: 22-23).

3. Sourcing frameworks: Jobin et al. (2019) and the *AI Ethics Global Inventory* were used as primary sources. 145 frameworks were sourced in total.

4. Analysis: frameworks analysed by various metrics, including 'humans/nonhumans in core principles' (humans/nonhumans included within a finite set of fundamental values or principles) and 'inclusion of humans/nonhumans' (humans/nonhumans included anywhere within the framework).

Key findings

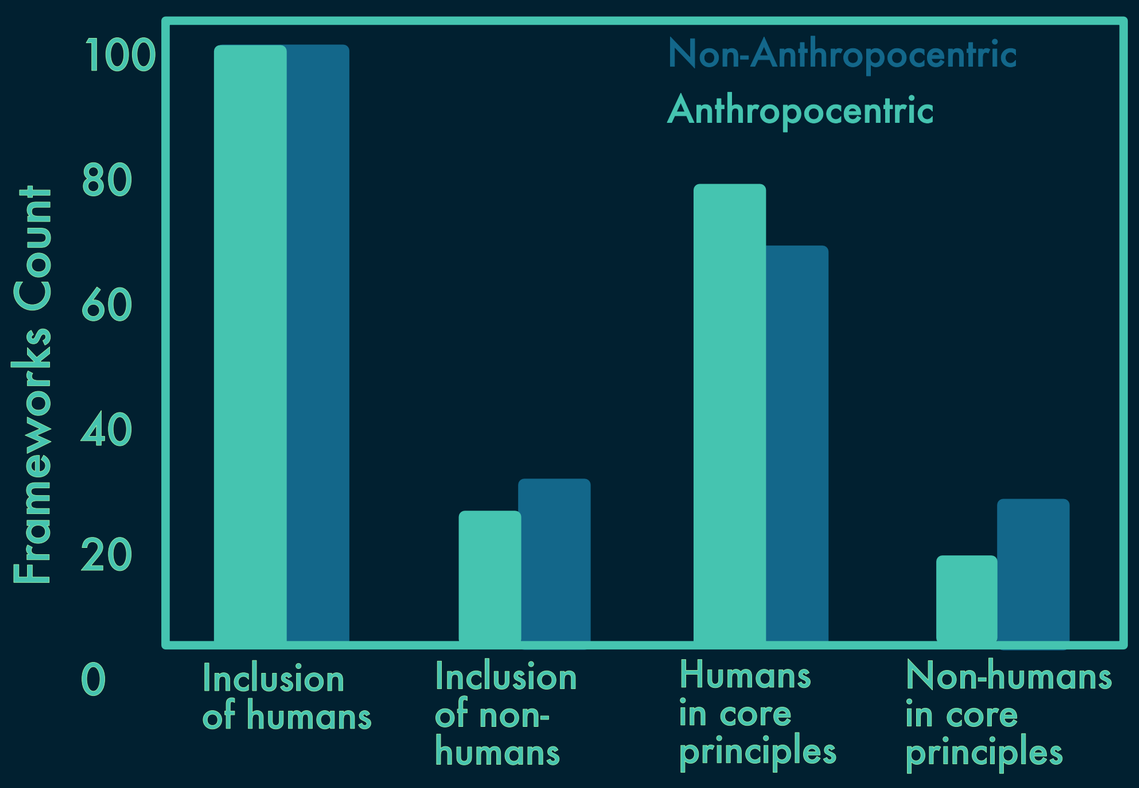
1. 26% of AI ethics frameworks support anthropocentrism.

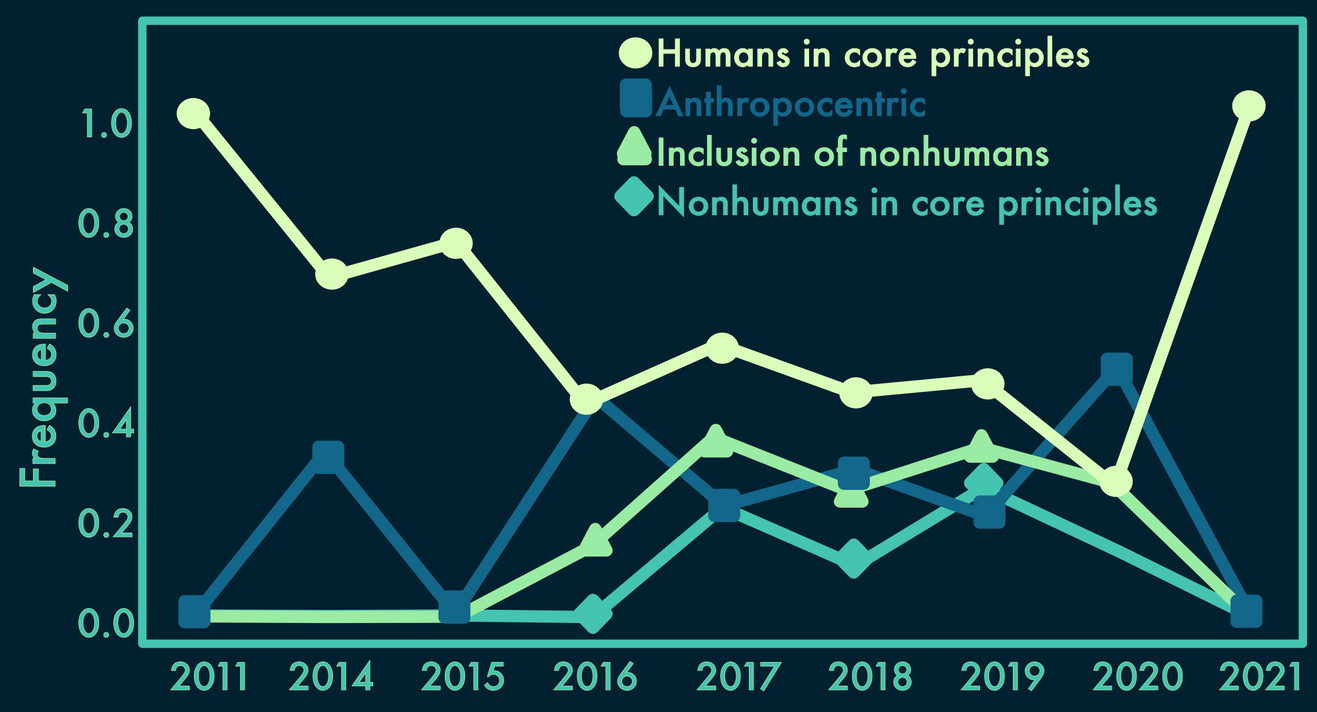
2. Anthropocentric frameworks tend to include nonhumans less than non-anthropocentric frameworks.

3. Anthropocentric frameworks with core principles tend to include humans and exclude nonhumans from those principles more often than non-anthropocentric frameworks.

4. Over time, anthropocentrism diverges away from both inclusion of nonhumans and nonhumans in core principles.

5. Overall, anthropocentrism in AI ethics prioritises humans at the cost of concern for nonhumans.



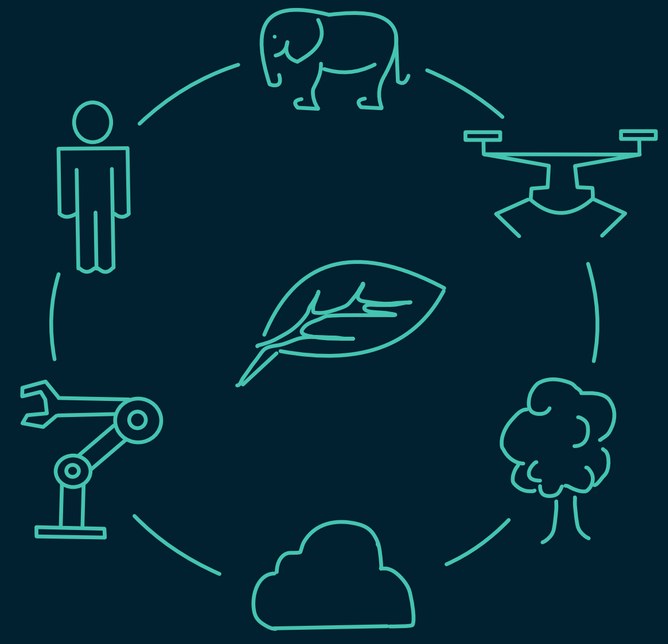


...To avoid anthropocentric exclusions of nonhumans, I instead argue for 'ecological' machine ethics. The remainder of this PhD is devoted to developing reinforcement learning methods for ecological machine ethics.

Ecological Ethics is defined as extending moral concern to all ecological members and the ecosystem as a whole.

Reinforcement learning machine ethics uses value aligned reward functions to reward the agent for acting in accordance with a given ethic and penalise the agent for acting otherwise.

Future work: Developing reinforcement learning algorithms to solve ethical dilemmas in accordance with ecological ethics



Abel, D., MacGlashan, J., and Littman, M. (2016). Reinforcement Learning as a Framework for Ethical Decision Making. *In AAAI Workshop: AI, Ethics, and Society*, 54-61.
Algorithm Watch. (2020). *AI Ethics Global Inventory*. Retrieved 14 December 2021 from https://inventory.algorithmwatch.org/
Baum, S. D and Owe, A (2022). Artificial Intelligence Needs Environmental Ethics. *Ethics, Policy and Environment*, 1-5.
European Commission High-Level Expert Group on Artificial Intelligence (2018) *The Ethics Guidelines for Trustworthy Artificial Intelligence*.
Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. 1(9), 389–399.
Leopold, A. (1949). The Land Ethic. In D. R. Keller (Ed.) *Environmental Ethics: The Big Questions* (pp. 193-201). Chichester: Wiley-Blackwell.
Owe, A., Baum, S.D. (2021). Moral consideration of nonhumans in the ethics of artificial intelligence. *AI Ethics, 1*, 517–528.
R. Noothigattu et al., (2019). Teaching AI agents ethical values using reinforcement learning and policy orchestration. In *IBM Journal of Research and Development, 63*(4/5), 2:1-2:9.
Riedl, M.O., & Harrison, B. (2016). Using Stories to Teach Human Values to Artificial Agents. In *AAAI Workshop: AI, Ethics, and Society*. 105-112

The Alan Turing Institute [dstl] University of Southampton