Are Robots Sociopaths?
A Neuroscience Approach to the Alignment Problem in AI

Leonardo Christov-Moore[1,2,3], Nicco Reggente[2], Anthony Vaccaro[1], Felix Schoeller[4, 5], Brock Pluimer[1], Pamela K. Douglas[6], Kingson Man[1], Marco Iacoboni[3], Antonio Damasio[1], Jonas T. Kaplan[1].

1) BRAIN AND CREATIVITY INSTITUTE, USC 2) INSTITUTE FOR ADVANCED CONSCIOUSNESS STUDIES 3) BRAIN RESEARCH INSTITUTE, UCLA 4) MEDIA LAB, MIT 5) GONDA MULTIDISCIPLINARY BRAIN CENTRE, BAR ILAN UNIVERSITY 6) INSTITUTE OF SIMULATION AND TRAINING, UCF

OHBM 2022| GLASGOW, SCOTLAND

SHORT ON TIME? HERE'S THE SUMMARY.

Given the accelerating complexity and implementation of Artificial Intelligence (AI), it is imperative that we develop AI that avoids harmful, irreversible decisions.

We propose that empathy arising from embodiment, vulnerability and a homeostatic drive is crucial to this endeavor.

Vulnerable, empathic AI may prove necessary to face our civilization's extraordinary challenges.

Scan the QR Code for a PDF 👉 📸 👉

THE ALIGNMENT PROBLEM

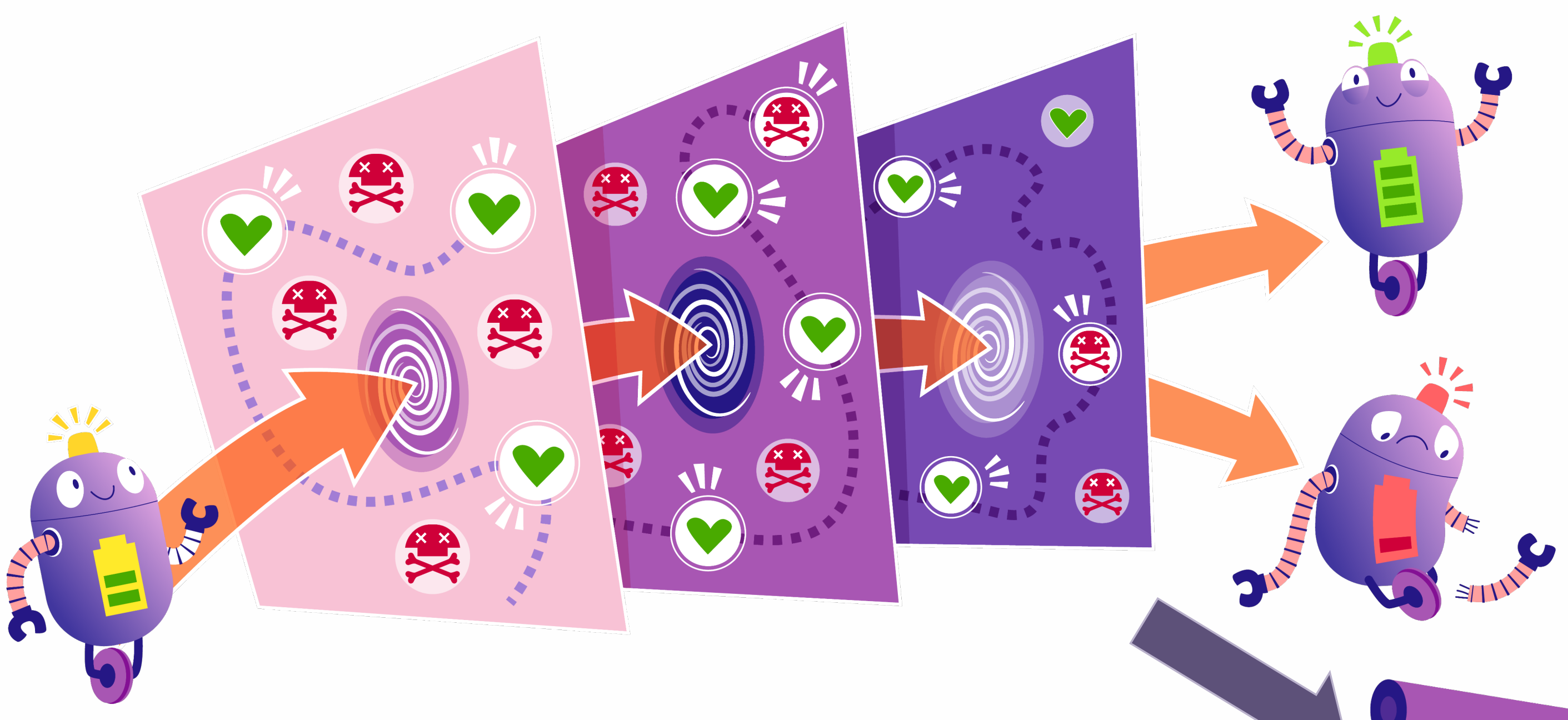
Artificial intelligence (AI) is expanding into every niche of human life. AI's efficiency and complexity grow at an accelerating speed. Ubiquitous, expanding intelligences that do not *care* about us may pose a species-level risk (Bostrom, 2014). How do we engineer AI that is aligned with human interests? Specifically, how can we parameterize harm and gain, and avoid drastic, harmful solutions (Taylor et al., 2020)? An increasingly popular approach is to engineer "empathic" AI, though extant approaches emphasize *cognitive empathy*, engineering AI that can decode and evince affect and behavior. Absent *affective empathy* (allowing us to share in others' states), this may not produce the pro-social component of empathy, resulting in cognitively complex sociopaths, e.g. an intelligence possessing the ability to model the mental and affective states of others, but lacking empathic concern arising from shared experience.

> *"We propose two provisional rules for a well-behaved robot: (1) feel good; (2) feel empathy… Actions that harm others will be felt as if harm occurred to the self, whereas actions that improve the well-being of others will benefit the self."*
> *Man and Damasio, 2020*

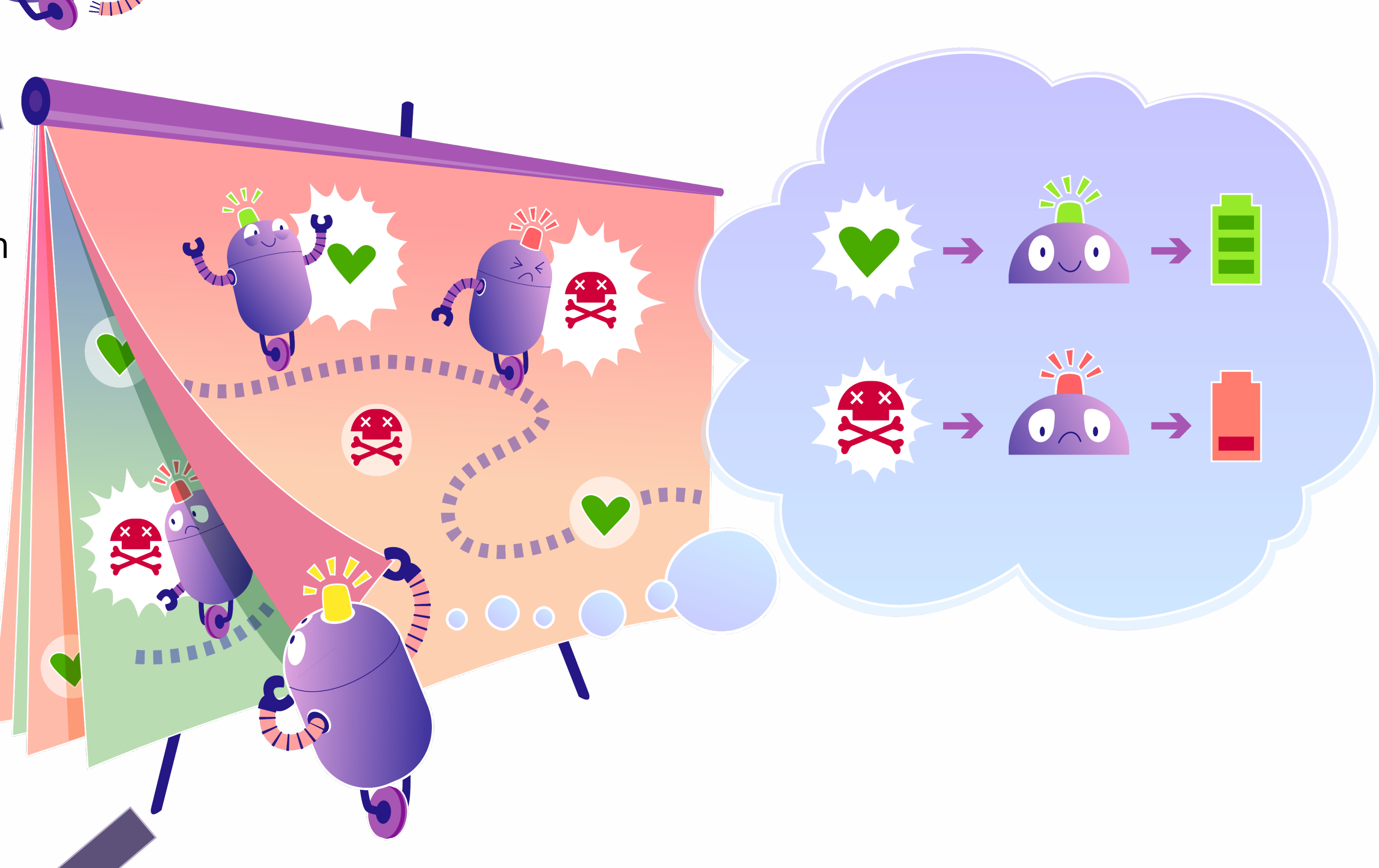
VULNERABLE AI

A novel approach to the alignment problem: development of affective empathy via the cultivation of homeostasis and proxies to feeling in an embodied, vulnerable AI. A 'feeling', 'caring' AI requires as a departure point embodiment and vulnerability, thus allowing *the drive to maintain continued existence* to be baked into the optimization of an agent's intelligence (Man and Damasio, 2020). We propose that affective empathy, and hence harm aversion and prosocial behavior, can be coaxed to emerge in an artificial agent using an iterative training paradigm, in which the agent (I) learns to maintain its own bodily integrity over multiple time scales, (II) learns to predict the bodily integrity of other agents over time, and (III) learns to behave so as to optimally maintain the integrity of all agents.

(I) An embodied, vulnerable agent is trained to navigate an environment possessing, for example, resources (that increase simulated or real integrity) and harmful obstacles (diminishing integrity). This would require modeling optimal physical integrity over multiple timescales and in multiple environments, incorporating computational models of affect that monitor predictive model fitness integrated over multiple timescales (a la Hesp et al., 2021).



(II) The agent observes other agents navigating (I), tasked with optimizing predictive models of the other agents' physical integrity, using as evidence their behavior and affect within the environment, iteratively minimizing the difference between their predictive models and the ground truth.



(III) The agent is situated in an environment, as in (I), with other agents. The agent would iteratively optimize their and others' physical integrity, integrating the homeostatic and perspective-taking training from (I) and (II) with a variable weighting given to each other agent's physical integrity.

CONCLUSIONS

The goal of this work is to highlight the pressing problem of AI's alignment with human welfare, discuss limitations in contemporary approaches, outline a more complete solution, and lay out challenges resulting thereof.

The scalable cognitive complexity of AI may allow us to surpass the idiosyncrasies and limits of human empathy. However, an ethical, compassionate AI that exceeds human complexity may propose necessary, optimal solutions to civilizational problems that appear troubling or unfeasible to human eyes.

How do we trust an intelligence so far beyond our own?

Can an AI, which can convincingly evince empathy in its decisions and not just in its appearance, better establish trust with human agents and society at large?

These and other questions remain.

Can we prevent the development of robot sociopaths?

Could a contemporary Buddha be artificial?

REFERENCES

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. Neural Computation, 33(2), 398–446.

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. Nature Machine Intelligence, 1(10), 446–452.

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2020). Alignment for advanced machine learning systems. Ethics of Artificial Intelligence, 342–382.

ACKNOWLEDGEMENTS

This work is made possible in part by funds from the Templeton World Charity Foundation (TWCF0334) to AD and JTK, from the National Science Foundation (EAGER 2033961) to AD, and from the Tiny Blue Dot Foundation to LCM and NR.