

AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values

GOPAL P. SARMA ^{1*}, NICK J. HAY ^{2†‡}, AND ADAM SAFRON ^{3§}

1. School of Medicine, Emory University, Atlanta, GA USA

2. Vicarious AI, San Francisco, CA USA

3. Department of Psychology, Northwestern University, Evanston IL USA

Abstract

We propose the creation of a systematic effort to identify and replicate key findings in neuropsychology and allied fields related to understanding human values. Our aim is to ensure that research underpinning the value alignment problem of artificial intelligence has been sufficiently validated to play a role in the design of AI systems.

I. ANTHROPOMORPHIC DESIGN OF SUPERINTELLIGENT AI SYSTEMS

There has been considerable discussion in recent years about the consequences of achieving human-level artificial intelligence. In a survey of top researchers in computer science, an aggregate forecast of 352 scientists assigned a 50% probability of human-level machine intelligence being realized within 45 years. In the same survey, 48% responded that greater emphasis should be placed on minimizing the societal risks of AI, an emerging area of study known as “AI safety” [1].

A distinct area of research within AI safety concerns software systems whose capacities substantially exceed that of human beings along every dimension, that is, superintelligence [2]. Within the framework of superintelligence theory, a core research topic known as the *value alignment problem* is to specify a goal structure for autonomous agents compatible with human values. The logic behind the framing of this problem is the following:

*Email: gopal.sarma@emory.edu

†Email: nnickhay@gmail.com

‡The views expressed herein are those of the author and do not necessarily reflect the views of Vicarious AI.

§Email: adamsafron@u.northwestern.edu

Current software and AI systems are brittle and primitive, showing little capacity for generalized intelligence. However, ongoing research advances suggest that future systems may someday show fluid intelligence, creativity, and true thinking capacity. Defining the parameters of goal-directed behavior will be a necessary component of designing such systems. Because of the complex and intricate nature of human behavior and values, an emerging train of thought in the AI safety community is that such a goal structure will have to be inferred by software systems themselves, rather than pre-programmed by their human designers. Russell summarizes the notion of indirect inference of human values by stating three principles that should guide the development of AI systems [3]:

1. The machine’s purpose must be to maximize the realization of human values. In particular, it has no purpose of its own and no innate desire to protect itself.
2. The machine must be initially uncertain about what those human values are. The machine may learn more about human values as it goes along, but it may never achieve complete certainty.
3. The machine must be able to learn about hu-

man values by observing the choices that we humans make.

In other words, rather than have a detailed ethical taxonomy programmed into them, AI systems should infer human values by observing and emulating our behavior [3–5]. The value alignment perspective on building safe, superintelligent agents is a natural extension of a broader set of questions related to the moral status of artificial intelligence and issues related to the architectural transparency and intelligibility of such software-based agents. Many of these questions are important for systems whose capabilities fall well short of superintelligence, but which can nonetheless have significant impact on the world. For instance, medical diagnostic systems which arrive at highly unusual and difficult to interpret diagnostic plans may ultimately do great harm if patients do not respond the way the AI system had predicted. In the medical setting, intelligible AI systems can ensure that healthcare workers are not subsequently forced to reason about circumstances that would not have ordinarily arisen via human diagnostics. Many researchers believe that similar situations will arise in industries ranging from transportation, to insurance, to cybersecurity [6–9].

A significant tension that has arisen in the AI safety community is between those researchers concerned with near-term safety concerns and those more oriented towards longer-term, superintelligence-related concerns [10]. Are these two sets of issues fundamentally in opposition to one another? Does researching safety issues arising from superintelligence necessarily entail disregarding more contemporary concerns? Our firm belief is that the answer to this question is “no.” We are of the viewpoint that there is an organic continuum extending between contemporary and long-term AI safety issues and that individuals and research groups can freely pursue both sets of issues without tension. One of the purposes of this article is to argue that not only can research related to superintelligence be grounded in contemporary concerns, but moreover, that there is a wealth of existing work across a wide variety of fields that is of direct relevance to superintelligence. This perspective should be reassuring to researchers

who are either skeptical of or have yet to form an opinion on the intellectual validity of long-term issues in AI safety. As we see it, there is no shortage of concrete research problems that can be pursued within a familiar academic setting.

To give a specific instance of this viewpoint, in a recent article, we argued that ideas from affective neuroscience and related fields may play a key role in developing AI systems that can acquire human values. The broader context of this proposal is an inverse reinforcement learning (IRL) type paradigm in which an AI system infers the underlying utility function of an agent by observing its behavior. Our perspective is that a neuropsychological understanding of human values may play a role in characterizing the initially uncertain structure that the AI system refines over time. Having a more accurate initial goal structure may allow an agent to learn from fewer examples. For a system that is actively taking actions and having an impact on the world, a more efficient learning process can directly translate into a lower risk of adverse outcomes. Moreover, systems built with human-inspired architectures may help to address issues of transparency and intelligibility that we cited earlier [6, 8], but in the novel context of superintelligence. As an example, we suggested that human values could be schematically and informally decomposed into three components: 1) *mammalian values*, 2) *human cognition*, and 3) *several millennia of human social and cultural evolution* [11]. This decomposition is simply one possible framing of the problem. There are major controversies within these fields and many avenues to approach the question of how neuroscience and cognitive psychology can inform the design of future AI systems [12]. We refer to this broader perspective, i.e. building AI systems which possess structural commonalities with the human mind, as *anthropomorphic design*.

II. FORMAL MODELS OF HUMAN VALUES AND THE REPRODUCIBILITY CRISIS

The connection between value alignment and research in the biological and social sciences intertwines this work with another major topic in contemporary scientific discussion, the repro-

ducibility crisis. Systematic studies conducted recently have uncovered astonishingly low rates of reproducibility in several areas of scientific inquiry [13–15]. Although we do not know what the “reproducibility distribution” looks like for the entirety of science, the shared incentive structures of academia suggest that we should view all research with some amount of skepticism.

How then do we prioritize research to be the focus of targeted replication efforts? Surely all results do not merit the same level of scrutiny. Moreover, all areas likely have “linchpin results,” which if verified, will increase researchers’ confidence substantially in entire bodies of knowledge. Therefore, a challenge for modern science is to efficiently identify areas of research and corresponding linchpin results that merit targeted replication efforts [16]. A natural strategy to pursue is to focus such efforts around major scientific themes or research agendas. The Reproducibility Projects of the Center for Open Science, for example, are targeted initiatives aimed replicating key results in psychology and cancer biology [17,18].

In a similar spirit, we propose a focused effort aimed at investigating and replicating results which underpin the neuropsychology of human values. Artificial intelligence has already been woven into the fabric of modern society, a trend that will only increase in scope and pace in the coming decades. If, as we strongly believe, a neuropsychological understanding of human values plays a role in the design of future AI systems, it essential that this knowledge base is thoroughly validated.

III. DISCUSSION AND FUTURE DIRECTIONS

We have deliberately left this commentary brief and open-ended. The topic is broad enough that it merits substantial discussion before proceeding. In addition to the obvious questions of which subjects and studies should fall under the umbrella of the reproducibility initiative that we are proposing, it is also worth asking how such an effort will be coordinated. Furthermore, this initiative should also be an opportunity to take advantage of novel scientific practices aimed at improving research

quality, such as pre-prints, post-publication peer review, and pre-registration of study design. The specific task of replication is likely only applicable to a subset of results that are relevant to anthropomorphic design. There are legitimate scientific disagreements in these fields and many theories and frameworks that have yet to achieve consensus. Therefore, in addition to identifying those studies that are sufficiently concrete and precise to be the focus of targeted replication efforts, it is also our aim to identify “linchpin” controversies that are of high-value to resolve, for example, via special issues in journals, workshops, or more rapid, iterated discussion among experts.

We make a few remarks about possible starting points. One source of candidate high-value linchpin findings would be those used by frameworks for understanding the nature of emotions. The extent of innate contributions to emotions is hotly debated, with positions ranging from emotions having their origins in conserved evolutionary programs [19, 20] to more recent suggestions that emotions are for the most part constructed through social inference [21, 22]. For example, Barrett suggests that the existing affective neuroscience and ethological literature may be based on questionable interpretations of studies of limited generalizability and uncertain reliability of research methods [22,23]. A related discipline is contemplative neuroscience, a field aimed at correlating introspective insights with a neuroscientific understanding of the brain. Highly skilled meditators from the Tibetan Buddhist tradition and others have claimed to have significant insight into human emotions [24,25], an understanding which is likely relevant to developing a rigorous characterization of human values. Other frameworks worth considering in depth are models of social-emotional learning based on predictive coding and Bayesian inference [26]. In these models, uniquely human cognition and affect arises from factors such as extensive early dependency for homeostatic regulation (e.g. fine-CT fibers contributing to analgesia through vagal stimulation [27,28]). It has been proposed that this dependence leads to models of self that are strongly shaped by the need to predict the minds

of others with whom the developing individual interacts. These reciprocal relationships may be the basis for the kind of joint attention and joint intentionality emphasized by Tomasello and others as a basis for uniquely human social cognition [29].

In terms of strategies for organizing this literature, we favor an open science or wiki-style approach in which individuals suggest high-value studies and topics to be the focus of targeted replication efforts. Knowledgeable researchers can then debate these proposals in either a structured (such as the RAND Corporation’s Delphi protocol [30]) or unstructured format until consensus is achieved on how best to proceed. As we have discussed in the previous section, The Center for Open Science has demonstrated that reproducibility efforts targeting large bodies of literature are achievable with modest resources [17, 18].

Our overarching message: *From philosophers pursuing fundamental theories of ethics, to artists immersed in crafting compelling emotional narratives, to ordinary individuals struggling with personal challenges, deep engagement with the nature of human values is a fundamental part of the human experience. As AI systems become more powerful and widespread, such an understanding may also prove to be important for ensuring the safety of these systems. We propose that enhancing the reliability of our knowledge of human values should be a priority for researchers and funding agencies concerned about AI safety and existential risks. We hope this brief note brings to light an important set of contemporary scientific issues and we are eager to collaborate with other researchers in order to take informed next steps.*

ACKNOWLEDGEMENTS

We would like to thank Owain Evans and several anonymous reviewers for insightful discussions on the topics of value alignment and reproducibility in psychology and neuroscience.

ORCID

Gopal P. Sarma  0000-0002-9413-6202
 Nick J. Hay  0000-0002-8037-5843
 Adam Safron  0000-0002-3102-7623

REFERENCES

- [1] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, “When Will AI Exceed Human Performance? Evidence from AI Experts,” *ArXiv e-prints*, May 2017.
- [2] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.
- [3] S. Russell, “Should We Fear Supersmart Robots?,” *Scientific American*, vol. 314, no. 6, pp. 58–59, 2016.
- [4] O. Evans, A. Stuhlmüller, and N. D. Goodman, “Learning the Preferences of Ignorant, Inconsistent Agents,” *arXiv:1512.05832*, 2015.
- [5] O. Evans and N. D. Goodman, “Learning the Preferences of Bounded Agents,” in *NIPS Workshop on Bounded Optimality*, 2015.
- [6] R. H. Wortham, A. Theodorou, and J. J. Bryson, “What does the robot think? Transparency as a fundamental design requirement for intelligent systems,” in *IJCAI 2016 Ethics for AI Workshop*, 2016.
- [7] J. P. Sullins, “When is a robot a moral agent?,” *International Review of Information Ethics*, vol. 6, 2006.
- [8] S. Wachter, B. Mittelstadt, and L. Floridi, “Transparent, explainable, and accountable AI for robotics,” *Science Robotics*, vol. 2, 2006.
- [9] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” *arXiv preprint arXiv:1606.06565*, 2016.
- [10] S. D. Baum, “Reconciliation between factions focused on near-term and long-term artificial intelligence,” *AI & Society*, pp. 1–8, 2017.

- [11] G. P. Sarma and N. J. Hay, "Mammalian Value Systems," *Informatica*, vol. 41, no. 3, 2017.
- [12] K. Sotala, "Defining human values for value learners," in *AAAI Workshop: AI, Ethics, and Society*, 2016.
- [13] M. R. Munafò *et al.*, "A manifesto for reproducible science," *Nature Human Behaviour*, vol. 1, p. 0021, 2017.
- [14] R. Horton, "What's medicine's 5 sigma?," *The Lancet*, vol. 385, no. 9976, 2015.
- [15] P. Campbell, ed., *Challenges in Irreproducible Research*, vol. 526, Nature, 2015.
- [16] G. P. Sarma, "Doing Things Twice (Or Differently): Strategies to Identify Studies for Targeted Validation," *ArXiv e-prints*, Mar. 2017.
- [17] The Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015.
- [18] T. M. Errington *et al.*, "Science forum: An open investigation of the reproducibility of cancer biology research," *eLife*, vol. 3, p. e04333, dec 2014.
- [19] J. Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford university press, 1998.
- [20] A. Damasio, *Self comes to mind: Constructing the conscious brain*. Vintage, 2012.
- [21] J. E. LeDoux and D. S. Pine, "Using neuroscience to help understand fear and anxiety: a two-system framework," *American Journal of Psychiatry*, vol. 173, no. 11, pp. 1083–1093, 2016.
- [22] L. F. Barrett, *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.
- [23] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face: Guide-lines for Research and an Integration of Findings*. Pergamon, 1972.
- [24] Harrington, Anne and Zajonc, Arthur and others, *The Dalai Lama at MIT*. Harvard University Press, 2006.
- [25] M. Solms and O. Turnbull, *The brain and the inner world: An introduction to the neuroscience of subjective experience*. Karnac Books, 2002.
- [26] V. Ainley *et al.*, "'Bodily precision': a predictive coding account of individual differences in interoceptive accuracy," *Phil. Trans. R. Soc. B*, vol. 371, no. 1708, 2016.
- [27] S. W. Porges and S. A. Furman, "The early development of the autonomic nervous system provides a neural platform for social behaviour: A polyvagal perspective," *Infant and Child Development*, vol. 20, no. 1, pp. 106–118, 2011.
- [28] M. Björnsdotter and H. Olausson, "Vicarious responses to social touch in posterior insular cortex are tuned to pleasant caressing speeds," *Journal of Neuroscience*, vol. 31, no. 26, pp. 9554–9562, 2011.
- [29] M. Tomasello, *The Cultural Origins of Human Cognition*. Harvard University Press, 1999.
- [30] B. B. Brown, "Delphi process: A methodology used for the elicitation of opinions of experts," tech. rep., Rand Corp Santa Monica CA, 1968.