One and the Same: Ethical Attribution and Distributed Reasoning in ML-driven Systems

Jesse Josua Benjamin
Human-Centered Computing, Freie Universität Berlin, jesse.benjamin@fu-berlin.de

KEYWORDS

post-phenomenology; moral ecology; participatory design; value-sensitive design, interpretability.

INTRODUCTION

In this position paper, I propose that the technical, designerly as well as the ethical dimension of interpretability[1] for machine learning (ML) are irreducibly intertwined, and even commensurate. With ML-driven systems, engineers and designers wield considerable power in shaping the values of the artefacts that govern our access to the world. This statement in itself is neither radical or new, with Winner's article on the politics of technological artefacts [12] a ubiquitous reference, and the post-phenomenological stance of mediation theory [11] gaining ground in the ethical discussions of HCI. Additionally, design methodologies such as participatory (PD) or value-sensitive design (VSD) are well articulated and poised to enter the discourse on interpretability. As a caveat, however, I suggest that any according assessment and design attempts for ML-driven systems ought to consider two co-constitutive factors: distributed hybrid reasoning and emergent values.

DISTRIBUTED HYBRID REASONING AND EMERGENT VALUES

When considering the opacity of ML-driven systems, Burrell suggests that a major reason lies in the operation of algorithms at scale, prompting system-wide change from high-dimensional parameter spaces [2]. Research in the field of ML attempts to deal with the latter, frequently striving to materialize attributes or reasons for decision-making from these spaces in order to visually provide explanations

[1]Which, for the sake of brevity, I take to be a high-level concept subsuming Explainability from the XAI [3], as well as Fairness, Accountability and Transparency, from the FAT-ML discourse[5] .

DIS'19, June 2019, San Diego, California USA

(e.g., visualizations, textual explanations, saliency maps). However, current attempts at improving interpretability of ML-driven systems in such ways are predominantly focused on expert interfaces (e.g., use cases which are heavily laden with confirmation bias and *a priori* assumptions [1]); and as a consequence are overwhelmingly and exclusively concerned with formal model interpretability [6]. The latter is particularly troublesome, as the major ethical issues of opaque regulation, surveillance or commercialisation are not co-located with the model as such, but with the particular way in which ML systems are embedded within socio-technical, distributed cognitive [4] systems. It is in the latter where the particular modes of reasoning of ML systems [7] have their actual effect; i.e. where human and machine frames of reference are contingent on each other. Parallel to the lack of the theoretical grounding in ML interpretability research, I claim that VSD or PD methodologies are not yet theoretically grounded for the consequences of this contingency: as frames of reference shift in distributed cognitive systems, values both expressed and implicit shift as well. Therefore, the ethico-political quality of what is visible, sayable or doable by whom at which point [9] cannot be a stable assessment. VSD and PD, as a consequence, need to be radicalized as a constant companion to deployment rather than early stage methods. Furthermore, the post-phenomenological framework that often motivates design research in HCI equally necessitates an extension, to be able to account for diversified mediation in distributed cognitive systems. A promising theoretical augmentation lies in the environmental stance of 'moral ecologies', which considers machine systems not as a unified, monolithic actor but rather an ecosystem of 'evaluators', e.g. decision-makers with diverse styles of reasoning [10]. The strength of this approach lies in the premise that ethics, mediation and human as well as machine reasoning are inextricably linked, each mediating specific frames of reference and action according to their particular style and system position. An explicit ethico-political charge would be part and parcel of such an extension.[2] Consider, for example, the ethics of mediation in decision-making by both humans and technologies when crowdworkers label images for below-minimum-wage, without being aware of providing training data for illegal facial recognition or drone surveillance.

[2]In contrast to the flat, de-politicized ontology of comparable approaches such as actor-network theory or object-oriented ontology.

PROPOSAL

I propose to combine the post-phenomenological stance of mediation with the environmental focus of moral ecology. Such a combination may be used to avoid reifying the ethics of ML-driven systems to specific artefacts (e.g., a user interface on the one or training data on the other side). But there is a larger point to this proposal as well: as suggested by Parisi [8], the unexpected promise of automated reasoning technologies may lie in the potential, through means of design, to disclose in which ways societal biases and particular values are enacted via the technologies of our distributed cognitive systems. We could not explicate the capitalist assembly line from the car at the same time as we drove it. Now, there is no reason beyond ideology to not make explicit the contingency and uncertainty of values (e.g., truths) in any technological encounter.

ACKNOWLEDGMENTS

This work is supported by the German Federal Ministry of Education and Research, grant 03IO1633 ("IKON — Visualizing the potential for knowledge transfer in research museums").

REFERENCES

[1] Jesse Josua Benjamin and Claudia MÃijller-Birn. 2019. Materializing Interpretability: Exploring Meaning in Algorithmic Systems. In *Proceedings of the 2019 ACM Conference Companion Publication on Designing Interactive Systems (DIS '19 Companion)*. ACM, San Diego, CA, USA. https://doi.org/10.1145/3301019.3323900

[2] Jenna Burrell. 2016. How the machine âĂŸthinksâĂŹ: Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (Jan. 2016), 205395171562251. https://doi.org/10.1177/2053951715622512

[3] David Gunning. 2016. Broad Agency Announcement : Explainable Artificial Intelligence (XAI). https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

[4] Edwin Hutchins. 2000. Distributed cognition. *International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science* 138 (2000).

[5] Till Kohli, Renata Barreto, and Joshua A. Kroll. 2018. Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Vol. 81. PMLR, New York, NY, USA, 1–7. https://fatconference.org/static/tutorials/fatconf18_lexicon_tutorial.pdf

[6] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]* (Dec. 2017). http://arxiv.org/abs/1712.00547 arXiv: 1712.00547.

[7] Luciana Parisi. 2016. Automated Thinking and the Limits of Reason. *Cultural Studies âĘŤ Critical Methodologies* 16, 5 (2016), 471–481. https://doi.org/10.1177/1532708616655765

[8] Luciana Parisi. 2019. The alien subject of AI. *Subjectivity* 12, 1 (March 2019), 27–48. https://doi.org/10.1057/s41286-018-00064-3

[9] Jacques Rancière. 2006. *The politics of aesthetics : the distribution of the sensible* (paperback ed. ed.). Verso, London.

[10] Christopher Charles Santos-Lang. 2015. Moral Ecology Approaches to Machine Ethics. In *Machine Medical Ethics*, Simon Peter van Rysewyk and Matthijs Pontier (Eds.). Springer International Publishing, Cham, 111–127. https://doi.org/10.1007/978-3-319-08108-3_8

[11] Peter-Paul Verbeek. 2006. Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology, & Human Values* 31, 3 (May 2006), 361–380. https://doi.org/10.1177/0162243905285847

[12] Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus* 109, 1 (1980), 121–136. http://www.jstor.org/stable/20024652