Computational Philosophy

Hubert Etienne
Facebook AI Research
Paris, France
hae@fb.com

ABSTRACT

The critical value of interdisciplinarity is increasingly accepted not only for promoting the responsible use of machine learning models, but also for increasing their performance. Social scientists can then help engineers better understand the population they are gathering data from. To be successful, however, interdisciplinary collaborations require more than just gathering researchers from various fields around a table. They call for addressing challenges such as developing a common language, understanding different ways of reasoning, and addressing epistemic controversies to agree on shared criteria upon which can be assessed the validity of the co-produced knowledge. Controversies bring about relevant epistemic questions to promote an ongoing reflection on scientific methods. In contrast, when a discipline leverages its own methods to approach a topic traditionally associated with another discipline, the unsolicited interference is often followed by a backlash which undermines the possibility of collaboration (e.g., [1], [2], [3]).

New environments emerge in academic centers willing to welcome interdisciplinary research and a key challenge is creating practices and methodologies that could enable such research. Computational philosophy is the approach I develop to serve this goal. I present here two examples of such collaborations I have led, as a philosopher, alongside machine learning engineers. These collaborations resulted in great outcomes and significantly advanced the state of the art in the domain of online social interactions.

The first example is an empirical study on misinformation based on an analysis of user-generated reports from Facebook and Instagram [4]. The mixed approach I developed with Onur Çelebi allowed us to identify meaningful variations in the volumes and types of false news, as well as in the manipulative strategies developed to spread misinformation among countries and platforms. Thanks to an original typology we created to classify content, we were able to identify four distinct types of behaviors for users reporting content to moderators. This allowed us to propose explanations for up to 55% of the inaccuracy in user reports, suggest solutions to improve the overall signal by taking action on the different sources of inaccuracy, and build a classifier capable of distinguishing credible user reports from others to support misinformation detection.

The second example is an empirical study of social influencers on Instagram [5]. François Charton and I developed an original typology to classify Instagram influencers based on their source of legitimacy. This allowed us to identify different kinds of audiences characteristic of the different types of influencers, which we analyzed through René Girard's mimetic desire theory [6]. This research enriched Girard's philosophical theory while it helped us advance the understanding of social interactions between influencers and their followers by superimposing a communication system inspired by Marshall McLuhan's media theory onto them [7]. Leveraging different signals, we were then able to identify for each category of influencer which kinds of posts were most likely to generate a positive response and negative feedback.

I am now expanding my approach to hate speech, bringing together the psychological mechanisms of hate conceptualized by Girard [8] to better understand the manifestations of hate on Instagram. These three blocks should then allow me to present a holistic approach to online interactions upon which an ethical system for the moderation of problematic online interactions could be erected.

KEYWORDS: Philosophy of AI, AI Ethics, Content Moderation, Behavioral sciences, Misinformation

ACM Reference format:
Hubert Etienne. 2022. Computational philosophy. In *Proceedings of 2022 AAAI/ACM Conference on Artificial intelligence, Ethics and Society (AIES'22), 1-3 August, Oxford,* ACM, New York, NY, USA, DOI: https://doi.org/10.1145/3514094.3539562

AIES '22, August 1–3, 2022, Oxford, United Kingdom.

ACM ISBN 978-1-4503-9247-1/22/08. https://doi.org/10.1145/3514094.3539562

REFERENCES

[1] Hubert Etienne. 2021. The dark side of the 'Moral Machine' and the fallacy of computational ethical decision-making for autonomous vehicles. In *Law, Innovation and Technology* 13, 1 (2021), 85–107, DOI: https://doi.org/10.1080/17579961.2021.1898310
[2] John Harris. 2020. The immoral machine. *Cambridge Quarterly of Healthcare Ethics* 29, 1 (2020), 71–79.
[3] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. A Word on Machine Ethics: A Response toJiang et al. arXiv:2111.04158 (2021).
[4] Hubert Etienne and Onur Celebi. 2022. Listen to what they say: better understand and detect misinformation with user feedback. in review.
[5] Hubert Etienne and François Charton. 2022. Computational philosophy enlightens Social influencers on Instagram. in review.
[6] René Girard. 1961. *Mensonge romantique et vérité romanesque.* Grasset, Paris.
[7] Marshall McLuhan. 1994. *Understanding media: The extensions of man.* MIT press, Cambridge.
[8] René Girard. 1972. *La Violence et le sacré.* Grasset, Paris.