DSSV 2023
Data Science, Statistics & Visualisation

ECDA 2023
European Conference on Data Analysis

Reinforcement learning from human feedback and AI safety

Dmitrii Krasheninnikov

University of Cambridge, Computational and Biological Learning Lab

Reinforcement learning from human feedback (RLHF) was shown to be useful in finetuning large language models (LLMs) to be easier to interact with, more honest, and less toxic [1]. In our upcoming publication with a working title “RLHF is Not All You Need”, my collaborators and I provide an overview of the various issues associated with RLHF. My talk will cover several of these issues, from straightforward ones like flawed feedback resulting in bad reward models that lead to reward hacking [2], to less obvious ones like goal misgeneralization and the incentive to manipulate the human evaluators. I will also present potential solutions and directions for addressing these challenges. Finally, I will connect these issues to broader trends in AI development and the potential risks posed by advanced AI systems.

Keywords: RLHF, reward modeling, large language models, AI safety, AI alignment

[1] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, *et al.*, “Training a helpful and harmless assistant with reinforcement learning from human feedback,” *arXiv preprint arXiv:2204.05862*, 2022.

[2] J. M. V. Skalse, N. H. Howe, D. Krasheninnikov, and D. Krueger, “Defining and characterizing reward gaming,” in *Advances in Neural Information Processing Systems*, 2022.