AI SAFETY OF NEURAL NETWORKS

Gomozov Y. P., Mats V. I.

National Technical University "Kharkiv Polytechnic Institute", Kharkiv

Recent advances in artificial intelligence and machine learning intensified discussions regarding AI safety and the possible risks of technological advances. There are a few main directions where these discussions take place:

Security: What can a malicious adversary do to an ML system? How do we prevent the misuse of ML systems to attack or harm people? How we should limit access to AI systems and their research and development? For example, a strong language model might spam millions of users in seconds with personalised emails, a risk which wasn't present in the past, or at least not nearly on the same scale. [1]

Privacy: How can we ensure privacy when applying machine learning to sensitive data sources such as medical data? How should these regulations differ from those on humans using the same data?

Economic and social consequences of ML/AI development: what areas of human life will be partially or fully taken by ML/AI systems in the nearest future? What consequences might it have for people's level of life, economic activity, and the economic system in general? For a long time, it was thought that the main risk to be replaced by machines have people doing physical work, such as factory workers, builders, and drivers, but in recent years AI suddenly took a large portion of "simple smart work", such as translation, advertising, consulting, which has profound economic consequences. [2]

Fairness: How can we make sure ML systems don't discriminate? For example, if using ML systems to make decisions about hiring people for some job, applying for credit, or scoring people in university, ML system often might learn wrong patterns from the data, interchanging correlation with the causation, and leading to unjustified discrimination in the final behavior. We don't want ML systems to make a judgement based on race/gender/age, even if data show some correlations there. [2, 3]

Transparency: How can we understand what complicated ML systems are making decisions and what can affect these? How to make systems which remain safe from both random and malicious adversarial attacks? [3]

This work analyses these topics and overviews possible solutions, risks, and ways to reduce the negative impact of the fast development of AI and ML systems on society.

References:

1. Marco Barreno et al. "The security of machine learning". In: Machine Learning 81.2 (2010), pp. 121–148.
2. Ifeoma Ajunwa et al. "Hiring by algorithm: predicting and preventing disparate impact". In: Available at SSRN 2746078 (2016).
3. Julius Adebayo, Lalana Kagal, and Alex Pentland. The Hidden Cost of Efficiency: Fairness and Discrimination in Predictive Modeling. 2015.