Essay: Discussion about future of Bias in AI

Hanna Foerster
Department of Computer Science
Durham University
Durham, UK
hanna.foerster@durham.ac.uk

I. Discussion of paper

Tesla's Elon Musk said that AI technology in military is "more dangerous than nukes." [1]. The Russian president Putin said the nation that leads in AI "will be the ruler of the world" [2]. The paper "Artificial Intelligence: A Threat to Strategic Stability" by J. Johnson [3] discusses more in detail how AI will affect the strategic stability between countries with great military power. He explains that combining AI with conventional weapons will multiply the impact of these weapons in terms of capacity, speed, and possibility of escalation. Countries will assess these AI-enhanced weapons predominantly by their possible destruction power rather than actual functionality. Since underestimation of destructive could be catastrophic and the speed of AI puts the attacked country at a disadvantage Johnson notes that incentives for pre-emptive strikes against powers with superior technology will rise. Additionally, there will be competition for superior military AI. Thereby, one of the biggest risks to stability will be the premature usage of AI technology which as of now are still more unreliable and unsafe than is often calculated by the head of states. The fallibility of AI systems is due to the unpredictive nature of machine learning, risk of 'data poisoning' and algorithmic biases. Johnson notes that this is especially dangerous due to the accelerating power of AI systems that outpace our human comprehension. Johnson ends with his warning note that 'human error' and 'machine error' is going to complement each other and cause inadvertent effects and possibly even accidental escalation.

Johnson highlights the dire effects of unconditionally relying in AI systems. Thereby the resolution of the sources of unreliability in machine learning models directly affect the stability of world peace. In his paper Johnson focuses on detailing possible AI technologies that could lead to escalation of conflict and leaves an impactful picture of potential disaster in the reader's mind. He goes as far as describing possible accidents when using AI in nuclear missile target detection. In combination with his arguments about the speed of AI and possible technical failures convinces the reader to support his views and realise the potential dangers of erroneous AI use.

In my opinion Johnson falls short of giving the reader a full picture on how problematic biases are in the context of military AI. Johnson mentions the bias ensuing from wrong assumptions made by the coders, but in the lectures, we have learned that a bias could have been produced at every stage of building the ML model. For example, data only shows a limited view of the world, so there will always be a dataset bias and selection bias. Moreover, prejudicial sentiment in the labelling and interpretation process could have affected the model. A highly problematic example of a bias in military context would be if an AI were to decide from which threshold of aggression a nuclear missile was launched, and this threshold was distinct for every country. Additionally, Johnson fails to address a military future in which the nuclear, bio-chemical, and cyber weapons that have been developed are deployed by human decisions rather than the decision of AI in the case the bias of AI could not fully be resolved. This could be equally disastrous, and rash based on subjective views of a decision-maker. It then becomes the question which of the two alternatives is more unsafe if both methods remain imperfect.

II. Discussion about future of bias in AI

Bias in AI will become the pivotal cornerstone in the future use of AI solutions. If bias issues are satisfactorily dealt with, AI decisions could help humanity to make fairer decisions. To mitigate the bias risk all layers of the creation process of AI models need to be examined. Barocas and Selbst argue that the data underlying a model often is imperfect with models often incorporating existing prejudices [4]. They further address the importance of pinpointing the source of discrimination and finding methods to deal with it in a 'fair' way. One such example would be dealing with historic biases in datasets such as the bias from century-long discrimination of black people in the US.

There have been attempts at dealing with bias in AI, such as the disparate impact remover from Feldman et al., which equalizes the distribution of groups by editing the values in the dataset in the pre-processing stage [5]. Other methods include reweighting, in which sample weights are recalculated until the ML model is bias-free [6] or the prejudice remover regularizer [7] which directly changes the ML model by adding a regularization term based on discrimination. Yet, it is important to note that the methods for dealing with biases are not all-purpose solutions. Suresh and Guttag elaborate in their paper that when working on fixing the bias of a ML model it should always be concretely stated which bias is being fixed, since different biases require different resolution methods. Moreover, choices in terms of discrimination are always being made in the ML creation process, hence, the concrete fairness for the AI application should always be rethought of based on the broad notion of fairness before attempting a repair [8].

If there is no one answer to fairness in AI solutions, creating a fair AI solution will remain an ethical gamble, which will continuously evolve. In my opinion, an AI application should therefore always be used with a tag about bias issues that were dealt with. Dealing with bias issues will not be easy since vested interests of companies and institution will prevent transparent discussions. Therefore, the right legislation and international consensus on a 'fairness' base will be key for dealing with bias in AI and the future of AI as a technology. The EU has now made a first step towards this by coming up with the first draft for an "Artificial Intelligence Act". This would ban the use of AI in the most dangerous areas and force upon rules of transparency and data quality for other high-risk areas, such as facial recognition and credit scoring [9]. The important question for the future will be how well these rules will be formulated and imposed on AI and whether an international standard will follow.

REFERENCES

[1] C. Clifford, "Elon Musk: 'Mark my words — A.I. is far more dangerous than nukes'", cnbc, 2018.

[2] G. Allen, "Putin and Musk are right: Whoever masters AI will run the world", cnn, 2019.

[3] J. Johnson, "Artificial Intelligence: A Threat to Strategic Stability", Strategic Studies Quarterly, vol. 14, no. 1, pp. 16-39, 2020.

[4] S. Barocas and A. Selbst, "Big Data's Disparate Impact", California Law Review, vol. 104, no. 3, pp. 671-732, 2016. Available: 10.2139/ssrn.2477899.

[5] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger and S. Venkatasubramanian, "Certifying and Removing Disparate Impact", Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268, 2015. Available: 10.1145/2783258.2783311.

[6] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination", Knowledge and Information Systems, vol. 33, no. 1, pp. 1-33, 2011. Available: 10.1007/s10115-011-0463-8.

[7] T. Kamishima, S. Akaho, H. Asoh and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer", Machine Learning and Knowledge Discovery in Databases, pp. 35-50, 2012. Available: 10.1007/978-3-642-33486-3_3.

[8] H. Suresh and J. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning", CoRR, vol. 190110002, 2019. Available: http://arxiv.org/abs/1901.10002.

[9] The Economist, "The Brussels effect The EU wants to become the world's super-regulator in AI", 2021. Available: https://www.economist.com/europe/2021/04/24/the-eu-wants-to-become-the-worlds-super-regulator-in-ai. [Accessed 2 May 2021].