Trust and explainability in machine learning

Steady growth in the machine learning market with an estimated value of US $20.83 billion globally by 2024 and (estimated) CAGR of 44.06% from 2017 to 2024 [Col20] reflects the brisk adoption of machine learning methods in the industry. Besides, the majority of companies equipped with these technologies are expected to use them in making major businesses decisions[Ant20].

While the hype surrounding this field is mostly justified, given the plethora of business and academic environments it effects - from risk management to protein-structure prediction, it is also precisely the grounds for debate over its trustworthiness.

We have witnessed cases where the relationship between machine and human judgement has been put to test. Citing the case of Stanislav Petrov - a Russian military officer in charge of monitoring a nuclear warning system during the cold war years - who ignored Soviet satellites' missile detection warnings owing to his hunch of a false alarm, saved several lives by not retaliating and further exacerbating political tensions between nations[see Fry18, Pg. 14, 15]. Moreover, looking at a much more recent incident concerning citizens of Idaho with disabilities - who did not receive adequate assistance by Medicaid program because of questionable modelling practices[Sta17, aclu] - it is clear how essential caution is in machine-assisted decision making.

Building trust, then, is the desideratum of the hour as we plunge deeper into the world of machine-powered prediction systems. Explainability is a crucial first step in the direction of trust with the goals of enabling subject matter experts to identify flaws in the model, deriving causality to improve results (eg.- improve sales), and making troubleshooting simpler[Gra19, Graaf]. The standard modelling methodology used today primarily focuses on optimizing evaluation metrics from predictions as evidenced in Kaggle competitions and machine learning research. The need for explainability arises as these metrics do not paint the full picture of the prediction process.

Herein lie the intricacies and traps of explainability. The reason model-building is performed in machines is because the complex nature of the task renders itself infeasible to be performed by humans. On account of the complicated computations under the hood, and development of growingly complex algorithms (transformers, GANs, etc.), giving a perfectly accurate and simple explanation of the process appears to be a futile effort[Koz18, Cassie]. It is also the case that the models generally considered interpretable (regression, SVMs, trees) may give a false sense of security by offering misleading explanations. For example, we may be easily deluded by the coefficients in a linear model with high multicollinearity among features.

Even so, some plausible and compelling ways to achieve the aforementioned goals of explainability are proposed by Zachary C. Lipton[see Lip17, Pg. 3, 4], which include generating hypotheses about causality that could be experimentally tested. [CRS05, Liu et al.] broaches Bayesian neural networks and regression trees for uncovering causal relationships, though these methods tend to rely on strong prior assumptions. Furthermore, rather than expounding on the lower-level workings of the model, post-hoc interpretations like visualizations of learned patterns (Google's Deepmind, Language Interpretability Tool) can help instil greater confidence in its inner workings. It is to be noted, however, that yielding a higher-level understanding does not equate to an absolute discernment of the complexity underneath, and thus there will always be gaps in explanations. Thirdly, the supervised model could be used to inform the human decision-making instead of being the decision-maker itself[Kim15, Kim et al.][Huy+11, Huysmans et al.].

There are areas of machine learning that require special attention in interpretability research due to their unique goals and approach to training. [DLS13, Dragan et al.] In reinforcement learning, two main

properties of predictability and legibility are often contradictory. Predictability is related to how well the observer can predict the trajectory of the agent given the goal at hand. Legibility, on the other hand, is how well the goal(intent) can be inferred from the trajectory. These properties can be quantified using bayesian inference, and a shift from a currently prevalent focus on predictability to legibility could accomplish the double-sided goal of interpretability and accuracy.

Psychological studies on infants afford a peek into the human tendency of interpreting observed actions as goal-oriented[CG07, G. Csibra]. This may be indicative of the effectiveness of the above-stated approach. Interactive reinforcement learning systems like assistive teleoperation, with their potentially widespread service applications in assisting persons with special needs, will need the user to be a good judge of the actions taken by the machine, which again, could only be achieved through improved legibility.

Apart from explainability, we are faced with issues of fairness and accountability[, FATML]. Accountability means it should be easy to challenge the decisions and the rationale behind the decisions. As noted by the FATML community - increased reliance on machine learning models may diminish the culpability of an organization or individual making bad decisions, deliberate or not. Facebook's study of the effects of news feeds on their users' emotions reveals the influence our social-network friends can have on our own moods[KGH14, Kramer et al.]. With the mental health of such a large user-base under the sway of one big corporation, it is easy to see the seriousness of the issue of accountability. Not only social networks, something as ubiquitous as search engines are prone to manipulation for ends such as impacting the outcome of elections[ER15, pnas].

Fairness, as pointed above, is an essential ingredient of trust. It is perhaps also the most volatile in its meaning and applicability. There are a few generalized definitions[CG18, Sharad] that are all rooted in socio-political structures of the world. Attempts to de-bias models (eg. de-biasing word embeddings) are constantly being made in the machine learning community. The treatment of fairness as a purely algorithmic problem, however, is likely to never produce sufficiently satisfactory results due to the implicit socio-political influences that reside outside the domain of available data. Nevertheless, the biases introduced by models resulting in unfair recruitment patterns[Das18, reuters] or race-influenced criminal risk scores[JK16, propublica] carry a huge ethical and social cost that cannot be ignored. It is therefore imperative that disciplines of ethics, law, politics, mathematics and computer science join forces to tackle the problem of fairness[Ber19, Marco].

In the wake of heightening awareness regarding the popularity of machine learning systems and its possible repercussions, many initiatives have been launched that aim to engender trust. Explainable AI is creating a suite of machine learning techniques that produce explainable models without sacrificing prediction accuracy[Tur18, darpa], while the European Union's requirements of trustworthiness emphasize human oversight, privacy, technical competence, and societal and environmental well-being, promising a well-rounded regulatory environment[AI19, europa]. Doubts have been cast on the validity of these goals, citing the incomplete nature of explainability[Koz18, Cassie]. Even with the weight of oft-times nebulous ideas of explainability and fairness, these developments reflect the growing and much-needed steps towards a holistic and modulated unlocking of machine learning's vast potential so that it benefits us at a minimal degree of danger.

References

[] URL: https://www.fatml.org/.

[AI19] High-Level Expert Group on AI. *Ethics guidelines for trustworthy AI*. 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[Ant20] James Anthony. *55 Notable Machine Learning Statistics: 2020 Market Share & Data Analysis*. 2020. URL: https://financesonline.com/machine-learning-statistics/.

[Ber19] Marco Angel Bertani-Økland. *What is FATML and why should you care*. 2019. URL: https://medium.com/grensesnittet/https-medium-com-mab-55055-what-is-fatml-and-why-should-you-care-dfb36e51f2f4.

[CG07] Gergely Csibra and Gergely György. "'Obsessed with Goals': Functions and Mechanisms of Teleological Interpretation of Actions in Humans". In: *Acta psychologica* 124 (Feb. 2007), pp. 60–78. DOI: 10.1016/j.actpsy.2006.09.007.

[CG18] Sam Corbett-Davies and Sharad Goel. *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. 2018. arXiv: 1808.00023 [cs.CY].

[Col20] Louis Columbus. *Roundup Of Machine Learning Forecasts And Market Estimates,2020*. 2020. URL: https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machine-learning-forecasts-and-market-estimates-2020/?sh=5f9892165c02.

[CRS05] Changchun Liu, P. Rani, and N. Sarkar. "An empirical study of machine learning techniques for affect recognition in human-robot interaction". In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2005, pp. 2662–2667. DOI: 10.1109/IROS.2005.1545344.

[Das18] Jeffrey Dastin. *Amazon scraps secret AI recruiting tool that showed bias against wome*. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[DLS13] A. D. Dragan, K. C. T. Lee, and S. S. Srinivasa. "Legibility and predictability of robot motion". In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2013, pp. 301–308. DOI: 10.1109/HRI.2013.6483603.

[ER15] Robert Epstein and Ronald E. Robertson. "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections". In: *Proceedings of the National Academy of Sciences* 112.33 (2015), E4512–E4521. DOI: 10.1073/pnas.1419828112.

[Fry18] Hannah Fry. *Hello World: Being Human in the Age of Algorithms*. 2018.

[Gra19] Robert de Graaf. *Are All Explainable Models Trustworthy?* 2019. URL: https://towardsdatascience.com/are-all-explainable-models-trustworthy-4378c5b0c1a5.

[Huy+11] Johan Huysmans et al. "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models". In: *Decision Support Systems* 51 (Apr. 2011), pp. 141–154. DOI: 10.1016/j.dss.2010.12.003.

[JK16] Surya Mattu Julia Angwin Jeff Larson and Lauren Kirchner. *Machine Bias*. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[KGH14] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks". In: *Proceedings of the National Academy of Sciences* 111.24 (2014), pp. 8788–8790. ISSN: 0027-8424. DOI: 10.1073/pnas.1320040111.

[Kim15] Been Kim. "Interactive and interpretable machine learning models for human machine collaboration". In: (Jan. 2015).

[Koz18] Cassie Kozyrkov. *Explainable AI won't deliver. Here's why.* 2018. URL: https://medium.com/hackernoon/explainable-ai-wont-deliver-here-s-why-6738f54216be.

[Lip17] Zachary C. Lipton. *The Mythos of Model Interpretability*. 2017. arXiv: 1606.03490 [cs.LG].

[Sta17] Jay Stanley. *Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case*. 2017. URL: https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case.

[Tur18] Dr. Matt Turek. *Explainable Artificial Intelligence (XAI)*. 2018. URL: https://www.darpa.mil/program/explainable-artificial-intelligence.