The urgent need for accountability in predictive AI

Arvind Narayanan
Princeton University
Oct 30, 2023

Leader Schumer, Senator Rounds, Senator Heinrich, and Senator Young,

Thank you for the opportunity to speak to you on high-impact AI. I am a professor of computer science at Princeton University and the director of the Center for Information Technology Policy. This document is written in my personal capacity.[1]

Much of high-impact AI is predictive AI. I will explain why this is a distinct type of AI that is inherently failure-prone and currently used in unaccountable ways, and offer policy recommendations.

Predictive AI is pervasive in high-impact domains

AI is used to make consequential decisions about people's lives: how long a patient should stay in a hospital before being discharged; whether a loan application should be approved; which job candidates should be hired; whether a criminal defendant should be released before their trial, and many others.[2]

These automated decision making systems use predictive logic. They use machine learning to predict how likely it is that a patient will be readmitted if charged, or whether a credit applicant will pay back the loan if approved, or whether a job candidate will be a good employee if hired, or whether a defendant will go on to commit another crime if released. These systems are trained by learning patterns from the behavior of past patients, applicants, etc.

Predictive AI stands in contrast to generative AI, such as ChatGPT, and other types of AI. By and large, predictive AI uses simple regression models using century-old principles rather than cutting-edge deep neural networks. Not all AI used in high-impact domains is predictive AI. For example, generative AI is being used to transcribe dictated medical notes or conversations into electronic health records. But when it comes to decision making systems in these domains, they largely use predictive logic. Predictive AI is important because its harms are being felt here and

[1] I'm grateful to Sayash Kapoor for collaboration in preparing this statement.
[2] Wang, Angelina, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. "Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy." ACM Conference on Fairness, Accountability and Transparency (FAccT) 2023.

now. Even when generative AI is used in a decision-making pipeline, such as résumé analysis to predict job success, it is still subject to the limits to predictability, which I will now discuss.

Predictive AI is error prone because it is hard to predict the future

Questions such as whether someone will pay back a loan depend on an array of factors that are unknown, and unknowable, at the time of decision making. So there is an intrinsic limit to the accuracy of predictive AI. At best, the technology can offer broad statistical generalizations. This is different from other applications of AI, say language translation, where there is nothing unknown; in contrast, there is a consensus answer that the system can learn to generate.

The empirical evidence is clear. Epic developed a tool for predicting which hospitalized patients are at risk of developing sepsis. It was deployed by hundreds of hospitals before an independent evaluation found that the accuracy of the tool was only 63%,[3] barely better than the flip of a coin.[4] Criminal risk prediction tools similarly have an accuracy of around 70%, which can be matched by a simple formula with just two variables: age and number of prior arrests.[5] The majority of defendants predicted to be at high risk of committing violent crime do not in fact go on to recidivate. Turning to hiring, there does not appear to be any peer-reviewed validation of the performance of any automated hiring tool. There have been bias audits of two of the leading tools but these were carefully scoped to exclude the more fundamental question of whether the tools even work.[6]

Unfortunately, companies have exploited public confusion around the disparate types of technologies that fall under the umbrella term “AI”. While some types of AI are rapidly advancing, predictive AI is not. This fact is not widely appreciated, whether by those who buy these tools or those subject to its decisions.

The predictive AI industry has evaded accountability

Predictive AI is an industry. In most cases, hospitals, court districts, or employers do not develop predictive tools in house — tools that might be tailored to their specific needs and

---

[3] Wong, Andrew, et al. "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients." JAMA Internal Medicine 181.8 (2021).
[4] In this paragraph, accuracy is measured using AUC-ROC (Area Under the Curve, Receiver Operating Characteristic), a standard way to measure the performance of a binary classifier.
[5] Dressel, Julia, and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism." Science advances 4.1 (2018).
[6] Costanza-Chock, Sasha, Inioluwa Deborah Raji, and Joy Buolamwini. "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022; Wilson, Christo, et al. "Building and auditing fair algorithms: A case study in candidate screening." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

those of the populations that they serve. Rather, they purchase or license one-size-fits-all tools from AI vendors. This exacerbates the problem. The predictive AI industry is built on an inherently limited technology that has been overhyped, but avoids transparency to obscure this fact. Its incentives aren't aligned with those of society.

For example, Epic originally claimed that its sepsis prediction AI had an accuracy of 76–83%, far higher than its actual accuracy. The hundreds of hospitals that adopted it did not challenge this claim. Performance evaluation of machine learning is notoriously tricky because of problems like data leakage.[7] This allows vendors to get away with false or misleading claims.

Vendors sell these tools based on the promise of full automation and elimination of jobs, but when the tools perform poorly, they retreat to the fine print which says that the tool shouldn't be used on its own. For example, Toronto recently used an AI tool to predict when a public beach will be safe. It went horribly awry: On a majority of the days the water was declared safe to swim in, it was actually unsafe.[8] Although the tool was not intended to be used on its own, it turned out that city officials never questioned its recommendations.

This incident illustrates diffusion of responsibility. Another example comes from Optum's tools to predict patients' future healthcare costs. Hospitals used it to prioritize patients for intervention. But it turned out that since hospitals had a history of spending less on Black patients, the tool baked in this bias and was less likely to prioritize a Black patient even if they had the same health conditions as a White patient.[9] Hospitals blamed Optum, but Optum said that the tool accurately predicted costs as designed, implying that the use of the tool in a way that resulted in disparate impact was the hospitals' responsibility.[10]

When predictive AI is used by governments for decision making, the AI system essentially enacts policy. This results in a form of policy making that is outsourced to private vendors without public participation.[11] The individual decisions made by these systems also tend to not be contestable by citizens, as vendors claim that the logic of the tool is a trade secret.[12]

---

[7] Kapoor, Sayash, and Arvind Narayanan. "Leakage and the reproducibility crisis in machine-learning-based science." Patterns 4.9 (2023).
[8] Martineau, Paris. Toronto Tapped Artificial Intelligence to Warn Swimmers. The Experiment Failed. The Information (2022). https://www.theinformation.com/articles/when-artificial-intelligence-isnt-smarter
[9] Obermeyer, Ziad, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366.6464 (2019).
[10] Gawronski, Quinn. Racial bias found in widely used health care algorithm. NBC News (2019). https://www.nbcnews.com/news/nbcblk/racial-bias-found-widely-used-health-care-algorithm-n1076436
[11] Mulligan, Deirdre K., and Kenneth A. Bamberger. "Procurement as policy: Administrative process for machine learning." Berkeley Tech. LJ 34 (2019): 773.
[12] Pasquale, Frank. Secret Algorithms Threaten the Rule of Law. MIT Technology Review (2017). https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/

Five recommendations for predictive AI accountability

The recommendations in this section are informed by the limitations of predictive AI described above, but they are aimed at the slightly broader category of automated decision making systems.

1. Strengthen efforts to inventory automated decision making systems. Citizens must know which automated decision making systems they are subject to, and who deploys them. This is a prerequisite to many of the rest of the recommendations such as contestability. Executive order 13960, from 2020, requires federal agencies to publish lists of current and planned uses of AI, but compliance has thus far been poor.[13] New York City's effort to catalog its uses of automated decision making also ran into difficulty.[14] Such efforts must be renewed and strengthened.

Beyond government use of AI, transparency can be helpful even for private sector systems, especially as many of these systems make life-altering decisions. Given the proper authority, regulatory agencies can enable such transparency regarding automated decision making systems within their respective purview.[15]

2. Establish standards for efficacy. Auditing and impact assessment are vital tools for identifying and remedying AI harms. While auditing has largely focused on discriminatory impacts, measuring the efficacy (or performance) of AI systems is equally important. Transparency around efficacy (or lack thereof) may help weed out snake oil products from the market. Beyond transparency, policy makers should consider minimum performance requirements for particularly consequential decisions. The details will depend on the sector and application.

The rules will have to account for the fact that meaningful auditing is hard and best practices are still emerging. Internal auditing not sufficient, and ongoing evaluation of deployed products by external auditors is needed.[16] An agency such as the National Institutes of Standards and Technology could lead the development of auditing best practices.

---

[13] Heilweil, Rebecca and Madison Alder. The government is struggling to track its AI. And that's a problem. FedScoop 2023. https://fedscoop.com/the-government-is-struggling-to-track-its-ai-and-thats-a-problem/

[14] Richardson, Rashida. "Confronting black boxes: A shadow report of the New York City automated decision system task force." AI Now Institute (2019).

[15] Engler, Alex. "A comprehensive and distributed approach to AI regulation." Brookings (2023). https://www.brookings.edu/articles/a-comprehensive-and-distributed-approach-to-ai-regulation/

[16] Raji, Inioluwa Deborah, et al. "Outsider oversight: Designing a third party audit ecosystem for ai governance." Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 2022.

3. Establish requirements for explanation and contestability. Explanation and contestability have come to be seen as core elements of procedural justice for automated decision making.[17] Explanation must encompass the overall logic of the system as well as how a particular decision was rendered. The adverse action notice requirements in the Fair Credit Reporting Act and the Equal Credit Opportunity Act are notable examples of explanation requirements in federal law, but generally, explanation has been under-utilized as a policy lever.

In requiring explanation and contestability, policy makers must reject the myth that AI is a black box and cannot be understood or explained. In many cases, predictive AI uses simple statistical models. In other cases, simple, interpretable models can be just as effective as more complex ones; policy can incentivize their use. More importantly, even without opening the black box, we can understand the aspects that matter: how it was built and how it behaves.[18]

4. Revamp AI procurement. Procurement processes should incorporate requirements for transparency and impact assessments.[19] Beyond improving the quality of AI products purchased by the government, this will have a ripple effect that will raise standards across the AI industry. Recognizing that AI evaluation requires specialized expertise, the government must create a centralized pool of experts that agencies can draw upon during procurement. The revamped procurement policies must recognize the stakes involved in procuring automated decision making systems. They must include public engagement and promote contestable design.[20]

5. The above recommendations can be carried out within the existing sectoral approach. None of the above recommendations require structural changes to AI oversight, such as a new regulatory agency.[21] Maintaining the existing oversight regime is not only more feasible; it will also lead to better outcomes, as it allows regulators armed with knowledge of specific domains and the respective industries to fine tune their rule-making effectively.

Conclusion. Predictive AI has proliferated both in the public and private sector. It is error-prone and unaccountable. Government agencies, both as procurers and regulators of automated decision making, are up to the task of changing the status quo, but policy can provide the impetus and authority for them to act.

---

[17] Selbst, Andrew, and Julia Powles. ""Meaningful information" and the right to explanation." conference on fairness, accountability and transparency. PMLR, 2018.
[18] Kroll, Joshua A. "The fallacy of inscrutability." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376.2133 (2018): 20180084.
[19] Improving AI procurement is an item in the recent Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.
[20] Mulligan and Bamberger 2019.
[21] Engler (2023).