Toward Implementable AI Standards

Christopher S. Yoo
University of Pennsylvania
csyoo@law.upenn.edu

Abstract

The U.S. government has proposed a standards-based approach to AI governance, with the precise contours of that standard to be developed over time. This article lays out the case for a standards-based approach and identifies four major elements that must be part of any AI standard.

1 Introduction

The release of ChatGPT-4 in early 2023 has given debates over how artificial intelligence ("AI") should be governed a greater sense of urgency. Some international organizations have issued high-level principles to guide governments when deciding how to regulate AI, e.g., OECD [2019]; UK [2023]. Most notably, the European Union's Artificial Intelligence Act, adopted on 13 March 2024, imposed a wide range of ex ante restrictions, the severity of which varies based on the risk level posed by a particular type of AI and whether the system constituted what the Act called "general purpose AI" ("GPAI") [EU, 2024].

The U.S. appears to be taking a different approach. Instead of adopting prescriptive regulation, the President Biden's 2023 Executive Order calls on various federal agencies to develop guidelines, standards, and best practices to guide the use of AI [U.S., 2023]. While helpful, the Executive Order provides little information about what topics such documents might address.

This article begins the process of filling this gap by exploring the merits of the U.S.'s approach as well as taking the first steps to translate the generalities contained in the high-level statements that dominate the discourse into parameters that are technically implementable. One essential consideration is an initial assessment of the major components that would comprise an AI standard.

2 The case for standards as the basis for AI governance

Standards represent a modality of governance that has become quite common in technologically sophisticated domains. This approach differs starkly from traditional command-and-control regulation in ways that yield substantial benefits. As an initial matter, unlike regulations, which are purely the product of governments, standards are produced by standards development organizations ("SDOs") that typically adopt a multistakeholder approach to governance that permits other constituencies, such as civil society, businesses, and the technical community, to help set agendas, speak, and vote. These decisionmaking processes are typically nimbler than those of governments. In addition, final decisions about which standard will prevail are made through choices made by users and implementers rather than by government fiat, as occurred in the U.S. during the competition between GSM and CDMA as the preferred standard for 2G and 3G cellular networks. The voluntary nature of standards adoption also allows successor technologies to emerge so long as they provide sufficient value to incentivize abandoning the incumbent standard.

Standards provide more than just a benchmark for proper behavior. In a world where the development of AI models involve a vertical chain of multiple entities, including producers of pre-trained models, fine tuners, and users, standards can play a key role in providing each link in this chain of production with the information it needs to understand the domains over which the model is likely to perform well and how much validation is appropriate before relying on a model as an input for a particular use.

Consider, for example, the Internet, where the most important SDO is the Internet Engineering Task Force ("IETF"). Participation in the IETF is open to anyone willing to engage in its processes. In contrast to the prior regime for setting telecommunications standards, which was dominated by the International Telecommunication Union ("ITU"), a United Nations organization in which governments make all of the key decisions, the IETF encompasses a wide range of participants, including most prominently the technical community. Decisions are also made by consensus. Despite early predictions that the IETF's efforts would amount to little more than an intermediate step on the way to adoption of the Open Systems Interconnection ("OSI") model, the resulting standards have proven remarkably robust even as the Internet has scaled far beyond its designers' wildest dreams.

This is not to say that standards-based governance is perfect. The decisionmaking processes employed by a particular SDO can favor certain interests. The decisionmaking process of some SDOs have become so slow that they have been criticized as ossified. Economic features such as network effects can cause standards to remain locked in long after they have become obsolete.

That said, the fact that the ultimate success of any standard is the product of decentralized decisions made by users and implementers rather than a centralized authority tends to make them more meritocratic and can lead to outcomes that surprise even so-called experts. For example, many knowledgeable observers confidently predicted that Bluetooth would emerge as the dominant wireless local area networking technology instead of Wi-Fi. Moreover, as is the case with Bluetooth and Wi-Fi, standards competition can result in multiple technologies existing in the end, each targeted toward different uses.

3 Principal elements of an AI standard

Simply deciding that standards represent the preferred modality of governance is not sufficient. The technical content of the standards are equally essential. The precise level of generality is critical. For example, the model cards often issued by AI providers generally provide too little information to be useful. That said, requiring disclosure of too much information is both costly and risks forcing providers to share with their competitors the very basis on which they are competing.

The first step in developing any standard is determining its major components. I contend that any AI standard must include provisions governing algorithms, training data, pre-release testing, and post-release evaluation.

2.1 Algorithms

One key area that any AI standard must govern is regarding the algorithms comprising the model. Many commentators have called for turning black boxes into glass boxes by requiring AI providers to disclose their algorithms. Other commentators concerned about AI bias argue for algorithmic disclosure to allow determination of whether the algorithm differentiates on impermissible criteria, such as race, gender, or religion.

While some degree of algorithmic disclosure is probably necessary, the benefits of such a requirement are easily overstated. The existence of hidden layers of neural nets necessarily mean that simply looking at the end product of AI training often provides little insight into what the parameters of the algorithm actually represent.

Even those concerned about bias may find that simply looking at the algorithms fails to answer many key questions. Any bias that is the result of biases in the training data may not be apparent on the face of the algorithm. Moreover, algorithms can construct proxies that mimic prohibited criteria without invoking them directly. Bias may thus become apparent only by analyzing the AI system's outputs.

In addition, the inclusion of parameters specific to criteria such as race may play a critical role in enabling adjustments to correct for biases in the training data or the use of proxies. As discussed in greater depth below, simply studying algorithms also cannot take into account the effects of the interactions of the decisions of multiple agents acting independently.

Algorithmic disclosure is also limited by legal constraints. For example, the Supreme Court has recognized that the Takings Clase of the U.S. Constitution places limits on the federal government's ability to require companies to disclose trade secrets without compensation [Ruckelshaus v. Monsanto Co., 1984]. Moreover, criminal prosecutors often assert that the parameters comprising AI used for criminal law are protected by government privilege.

2.2 Training data

Understanding the likely behavior of an AI system also depends on knowing a significant amount of information about the data on which the model was trained. Because AI is a form of predictive analytics that uses patterns in existing data to generate responses to prompts given to it, every AI system necessarily reflects the data on which it is trained. Although model cards typically include some information about the data used to train the model, they do not provide sufficient detail to evaluate a model's likely performance.

Disclosures about the source of training data can provide important guides as to their quality. In addition, some disclosures are essential to understanding what, if any, biases may exist in the data.

One critical component that determines the robustness of any AI model is the scope of the data on which it is trained. This is easily illustrated by the fact that ChatGPT-4 was initially trained on data through September 2021 and has since been extended to include data through April 2023. This necessarily means that any answers it gives to questions about factual events taking place after April 2023 are necessarily hallucinations.

Considerations about scope extend far beyond time. The fact that ChatGPT-2 and ChatGPT-3 were trained primarily on Reddit and Wikipedia data respectively makes those models inevitably overrepresent the patterns characteristic of those types of communications.

Consider further the use of AI to predict weather. Although studies indicate that this approach produces more accurate result faster and using less computing power than conventional models, concerns remain that AI-based models will provide less effective predictions over rarer events not well represented in the data on which these models were trained despite early findings that AI was able to predict three types of extreme weather events [Lam et al., 2023]. Although correctness may be more difficult to determine than with historical information, erroneous AI predictions based on patterns that fall outside the data on which the model was trained can constitute hallucinations in the same way as factual misstatements.

The limitations necessarily imposed by the scope of training data also belie the tendency of many AI developers to solve any problems in fidelity by throwing more data at the model. If the scope of the new data is no different from the old data, adding more will not expand the range of circumstances over which the model can provide accurate predictions. This phenomenon is underscored by current efforts by AI designers to train models on smaller amounts of higher quality data.

Finally, even the best trained model may produce inaccurate predictions when the environment has structurally changed since the time the training data was collected. One prime example is the 1998 collapse of the largest hedge fund in the world, known as Long-Term Capital Management ("LTCM") and founded in part by two Nobel Laureates in economics, which was triggered by a circumstance that the model had not seen before, specifically Russai's default on its debt. Another example is the collapse of Zillow's algorithmically driven iBuying platform, which was ill-prepared for the changes to the real estate market caused by the COVID-19 pandemic.

Thus, an AI standard must carefully consider what providers should disclose about the data on which a model was trained. This can be particularly important for foundation models used to develop other models and for when AI develop for one context is ported to another. Anu such disclosures must include the information that downstream AI firms need to know about the models on which they are building in order make sure they are fit for purpose.

While requiring further disclosure is always tempting, any standard must take into account that such disclosures are costly. These costs imply that any transparency requirement must carefully assess whether the benefits justify the costs. Moreover, the fact that outputs of AI systems are probabilistic means that no amount of disclosure can guarantee the veracity of any particular outcome. A key element of any standard must necessarily include a framework for assessing the optimal amount of disclosure that balances these considerations based on some measure of acceptable risk.

2.3 Pre-release testing

Another key element of any AI standard is requirements regarding pre-release testing. As an initial matter, the standard should specify which of the many forms of testing that those seeking to conform to the standard must conduct. For example, the IEEE "Standard for Assumptions in Safety-Related Models for Automated Driving Systems" discusses seven methods of validation and verification: systematic processes, safety-by-design architectures, formal methods, robustness analysis, simulation testing, closed course testing, and public road testing. Moreover, rather than creating a single standard covering all aspects of autonomous vehicle safety, the standard focuses on seven commonly occurring scenarios as well as twenty-three attributes verifiable via inspection and six others demonstrable via validation [IEEE, 2022].

Disclosure about testing provides the basis for others to assess the limitations of the testing regime. To use a non-AI example, seatbelts that previously passed a testing regime began to fail when the weight used to perform the test was positioned at a different angle [Weiss, 2008]. Strong performance under test conditions but poor performance in more robust circumstances is similar to the well-known algorithmic problem of overfitting. Information about the testing regime is thus critical to understanding what passing the test signifies and fails to signify.

In addition, every testing AI regime is susceptible to specification gaming and reward hacking in a manner reminiscent of the well-known problem of search engine optimization ("SEO"), in which website owners promote their ranking in search results by making changes designed to cater to the selection criteria that search engine values the most. This dynamic is captured by what is commonly known as "Goodhart's Law," which holds that "when a measure becomes a target, it ceases being a good measure" [Chrystal et al., 2003]. Examples of these problems are legion, including the pancake-flipping bot that maximized the duration of its performance by flinging the pancake as high in the air as possible rather than perfecting flipping technique, the Tetris bot that maximized its time of survival by putting the game on pause, and the CycleGan neural network that hid the original data in its code rather than develop an algorithm to reconstruct the data.

Any standard must thus carefully consider how much pre-release testing it will require both to assess the robustness of the validation criteria and to anticipate their vulnerability to opportunistic behavior. The fact that more testing is always an option again requires that any standard include some measure of optimality to determine when requiring additional testing would be justified. The probabilistic nature of AI outputs requires that any such measure be built around some concept of acceptable risk.

2.4 Post-release evaluation

Any AI standard must also include some regime of post-release evaluation. The simple reality is complex systems are characterized by emergent behavior that only emerges when the system is exposed to real-world environments at scale.

As noted above, algorithms prohibited from taking into account prohibited criteria such as race may nonetheless discriminate by using neutral variables that are highly correlated with the prohibited criterion as proxies. Unless one knows all of the correlations among all variables (both individually and in interaction with one another) and the prohibited variables, such proxy discrimination is almost impossible to detect except through studying the algorithm's outputs.

Another form of emergent behavior results from the interaction of actions multiple agents that are individually rational but interact in unpredictable ways. One classic, non-AI example is the flash crash of May 6, 2010, in which trades by one trader initiated a cascade of program trades that caused the Dow Joens Industrial Average to lose almost $1 trillion in market value, one of its largest intraday losses in its history. Scholars are now creating models to study the circumstances under which similar swarming effects might occur for AI [Canoniuco et al., 2019]. Such unpredictable interactions among individual actions that individually rational are only visible in post-release testing.

Unexpected outcomes can arise through the actions of hostile actors who are not acting in a manner consistent with an AI system's goals. Exhibit A is Microsoft's chatbot, Tay, which degenerated into a cesspool of racism and misogyny after trolls discovered that it would echo back whatever was fed to it. Studies indicate bad actors can cause AI systems such as ChatGPT to exhibit similar toxicity [Deshpande et al., 2023]. How AI responds to hostile environments is best studied after the fact.

Post-release testing can also play a critical role in detecting hallucinations, which can appear somewhat unpredictably. It can also reveal the problem of memorization, in which an AI model regurgitates a verbatim copy of a work when multiple copies of it are contained in the data on which it was trained.

AI's tendency to exhibit emergent behavior underscores the need to subject it to post-release testing. Any AI standard must provide details about what types of post-release testing it requires. As with data disclosure and pre-release testing, the probabilistic nature of AI and the fact that the standard could always require more testing necessarily requires that the standard include some basis for determining when the benefits of additional post-release testing would exceed the costs based on some measure of acceptable risk.

4 Conclusion

Standard represent a promising approach to AI governance that avoids the pitfalls of prescriptive command-and-control regulation. At a minimum, any such standard must contain provisions governing algorithms, training data, pre-release testing, and post-release valuation. Identifying such major categories is the first step toward developing standards that are implementable. In addition, any AI standard must provide some basis for determining when the benefits of additional protections would justify the costs, taking into account AI's inherently probabilistic nature.

Acknowledgements

This article expands on ideas initially advanced in Beyond Algorithmic Transparency for Generative AI. *Columbia Science & Technology Law Review*, 25(2) (forthcoming 2024). It was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A5A7083908).

References

[Canonico et al., 2019] Lorenzo Barberis Canonico and Nathan McNeese. Flash Crashes in Multi-Agent Systems Using Minority Games and Reinforcement Learning to Test AI Safety. In *Proceedings of the 2019 Winter Simulation Conference (WSC 2019)*, pages 193-204, National Harbor, Maryland, December 8-11, 2019. Institute of Electrical and Electronics Engineers. https://ieeexplore.ieee.org/document/9004675.

[Chrystal et al., 2003] K. Alec Chrystal and Paul D. Mizen. Goodhart's Law: Its Origins, Meaning and Implications for Monetary Policy. In Paul Mizen (ed), *Central Banking, Monetary Theory and Practice: Essays in Honor of Charles Goodhart*, vol. 1, pages 221-243. Edward Elgar Publishing, Inc.: Northampton, Massachusetts, 20003.

[Deshpande et al., 2023] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 1236-1270, Singapore, December 6-10, 2023. Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.88.pdf.

[IEEE, 2022] IEEE Standards Association. *IEEE Standard for Assumptions in Safety-Related Models for Automated Driving Systems*, IEEE Std. 2846-2022. Institute of Electrical and Electronics Engineers, New York, New York, 2022. https://ieeexplore.ieee.org/document/9761121 [https://perma.cc/NLH7-6U3L].

[Lam et al., 2023] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning Skillful Medium-Range Global Weather Forecasting. *Science*, 382(6677): 1416-1421, November 14, 2023.

[OECD, 2019] Organization for Economic Cooperation and Development. *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, May 21, 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 [https://perma.cc/ J7UP-PKH7].

[Ruckelshaus v. Monsanto Co., 1984] Ruckelshaus v. Monsanto Co., 467 U.S. 986 (1984).

[UK, 2023] United Kingdom, *The Bletchley Declaration by Countries Attending the AI Safety Summit*, *1-2 November 2023*, Policy Paper, Nov. 1, 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 [https://perma.cc/W6CU-TBYA].

[U.S., 2023] Joseph R. Biden, Jr., Executive Order No. 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *Federal Register*, 88(210): 75191-75226, November 1, 2023. https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf.

[Weiss, 2008] Kurt D. Weiss. Failure Mode Testing of Seat Belts. *Plaintiff Magazine*, January 2008. https://plaintiff-magazine.com/images/issues/2008/01-january/reprints/Weiss_Failure-mode-forensic-testing-of-seat-belts_Plaintiff-magazine.pdf./