Preserving Freedoms in Light of the AI Transparency Paradox

James Richard

PHL 215: Engineering Ethics

Mr. Miles Budimir

Apr. 13 2022

Over the past few decades, artificial intelligence (AI) has grown into a widely studied and widely applied field of technology; the scope of its applications has been a great source of optimism for the tech community. However, AI is not without its caveats. “The AI Transparency Paradox,” a term coined by Andrew Burt in a Harvard Business Review article of the same name, describes the inverse relationship between the openness or explainability of an AI algorithm and the security of that algorithm. In the world of classical software development—basically anything that isn’t AI—more transparency is generally better for security. David Wheeler makes a strong argument for this in his online book *Secure Programming HOWTO* (sec. 2.4). But in the world of machine learning and AI, increased transparency generally leads to an increased risk of being reverse engineered (Milli et al.) and therefore an increased vulnerability to adversarial examples. It also leads to an increased privacy risk for those whose data was used to train the model (Shokri et al.). This is a vital issue. Transparency is often the go-to solution for preventing ethical issues like discrimination and abuse of power, and it is necessary for the determination of responsibility. Some have also argued that transparency is required for “democratic decision-making,” and in one case transparency was even codified in an effort to protect consumers (Müller sec. 2.3). So, how should transparency and security be balanced? At some level this is a technical debate, but its results have broad ethical ramifications. And, until a technical solution is found, we must examine our current options from the perspective of engineering ethics.

Before these options can be quantified, we need to dive deeper into the ethical value of transparency. Although it is not always mentioned by name, transparency is actually a fundamental concept that undergirds the concept of informed consent, and by extension, autonomy. Without at least some degree of transparency, an ethical treatment of consumers and

the public at large could never be achieved. It is worth noting that complete explainability, the ability to explain *how* an algorithm reaches a decision, is not required. Most members of the general public could not completely understand a technical explanation of AI, let alone a technical explanation of a particular algorithm. This is also true in medicine, a field in which the concept of informed consent is more often considered. In most cases, patients do not have the skills needed to completely understand the procedures they may undergo, but doctors still have an obligation to communicate the risks and benefits of those procedures (*Informed Consent*). I would argue that, in the field of AI, engineers also have an obligation to communicate the risks and benefits of their algorithms to those who are impacted by them. This is supported by the IEEE Computer Society's Code of Ethics, which states that engineers are obligated to "disclose to appropriate persons or authorities any actual or potential danger to the user, the public, or the environment, that they reasonably believe to be associated with software or related documents" (sec. 1.04). Although transparency isn't mentioned by name, it is a core concept and a prerequisite for preserving public safety and autonomy.

In general, transparency is also necessary for democracy and a democratic way of living. In *Information, Democracy, and Autocracy: Economic Transparency and Political (In)Stability*, Hollyer et al. argue that transparency is one of the main causes for stability in democracies. "By improving economic outcomes and creating a better informed electorate, transparency appears to hold only benefits for a democratic society" (Hollyer et al. 319). They do not say the same for autocracies—it is possible for increased transparency to result in a regime change from autocracy to autocracy (321)—but it can also prompt and/or enable a transition to democracy (320). Transparency is also important in the corporate world. Transparency engenders trust between the company and the consumer and, from a utilitarian point of view, proper (Weeks et al.) levels of

transparency between management and employees improve productivity and strengthen company culture (*Transparency in the Workplace*). By an extension of these concepts, I would argue that some level of transparency (including a public, proven track record) is needed for AI to be successful. Without it, consumers in a free market economy would not trust it to do sensitive and high risk tasks and legislators would not allow it past the prototyping phase.

The ethical value of transparency is unquestionable. So how can it be achieved in the field of AI? AI and machine learning traditionally follow the "black-box" paradigm; we can see the inputs and outputs, but no useful insights can be gained by looking at the algorithm's inner workings. This is especially true for fully connected neural networks and convolutional neural networks. The parameters of these models are simply too interconnected and numerous to study. Or, at least, that was true until the field of neural network interpretability was born. Olah et al. outline the use of one of the most discussed concepts in neural network interpretability, feature visualization, in their article of the same name. Feature visualization seeks to determine which inputs optimally excite certain areas or even individual nodes of a neural network, and it is most often used on convolutional neural networks, the backbone of AI image processing (also called computer vision). In many regards feature visualization is successful. It can produce images of the patterns that excite various neurons and give insight into how various image classification results are reached. However, Olah et al. admit that it is not enough.

> By itself, feature visualization will never give a completely satisfactory understanding. We see it as one of the fundamental building blocks that, combined with additional tools, will empower humans to understand these systems.

Other commonly used methods like LIME (Ribeiro et al.) and SHAP (Lundberg and Lee) use surrogate models, which train simply on the inputs and outputs of a black-box model, and have

been quite successful. However, the efficacy of these models have been called into question (Slack et al.). For those interested, a broader taxonomy of AI interpretability methods can be found in "Explainable AI: A Review of Machine Learning Interpretability Methods," an article written by Linardatos et al.

Security, the other half of the "AI Transparency Paradox," is also an important ethical value, and it is near and dear to the engineering profession. Its near synonym, safety, is the term most often used, and can be found in sec. 1.03 of the IEEE Code of Ethics and in the first fundamental canon of the NSPE Code of Ethics. It is also found at the heart of Respect for Persons, an ethical framework which upholds the rights of the individual. Without an appropriate standard of safety and security, individuals' rights to life, health, and privacy (in the context of AI) could be unduly infringed upon.

Unfortunately, AI is not intrinsically secure. Andrew Lohn's report "Hacking AI," published by the *Center for Security and Emerging Technologies*, explores a variety of AI security risks and the inadequate defenses we have against them. One example from the report included a sticker designed by Tencent, a Chinese tech company, that had the ability to disrupt the self-driving capabilities of Tesla automobiles when placed on the road (v). When the sticker appeared in the frame of a camera mounted on the car, Tesla's self-driving AI misidentified it and reacted dangerously. This is a kind of adversarial example, an input that is engineered to fool AI. Generally, adversarial examples are created by adding perturbations to an example that the AI can correctly identify, but there are other methods (Song et al.), and they are very common. Other notable adversarial examples include a graffiti-like pattern that causes a stop sign to be read as a 45 mph speed limit sign (Eykholt et al.) and a specially designed, 3D printed turtle that AI identifies as a rifle (Athalye et al.). In some cases adversarial examples are used during the

development of an AI algorithm to improve performance (Goodfellow et al.). But, because the process is purely reactive, the primary algorithm tends to stay one step behind the algorithms generating adversarial examples. The presence of adversarial examples is a big issue for the progression of the field of AI. Although complete security is an unrealistic goal, issues of this magnitude call into question whether or not we are ready for AI to be applied to high risk scenarios (e.g. self-driving cars).

Some have also voiced concerns about AI's ability not only to get hacked, but to hack. In "The Coming AI Hackers," Bruce Schneier discusses several ways in which AI could hack or is hacking traditional software, other AIs, and even the human psyche. This includes the harmless but annoying consequences of machine learning—AIs are bound to stumble upon "hack" solutions because they explore randomly (Schneier 26)—and the more harmful tendencies—many recommendation systems tend to encourage extremist views (Schneier 31). There are a great number of ethical concerns related to the unintended consequences of AI, and engineers must navigate them all with care. However, as it stands, it seems that society has chosen to live with these particular risks.

When transparency is added into the mix, security risks for all types of AI increase. As I mentioned in the introduction, Milli et al. demonstrated that it is possible to reconstruct neural networks from their explanations. Even public access to the inputs and output of an AI (through an application programming interface, or API) may be enough to reconstruct it in some cases (Tramèr et al.). This is a big issue for the protection of intellectual property rights and for the privacy of those whose data trained the model. Other studies, including one authored by Fredrikson et al. and another authored by Shokri et al., speak directly to the privacy risks of transparency and liberal API access. It is conceivable that the ability to reconstruct AI algorithms

also increases the risks of adversarial examples; certainly, wholesale access to a copy of a neural network would not only infringe on the intellectual property rights of the creator but also make a variety of attacks on the original neural network possible. And if the neural network is used for anything vital or high-risk public safety could be seriously impacted.

In light of the conflict between transparency and security, I would submit that we are not ready for AI to be used in high-risk or mission-critical applications that require exposed inputs. There are too many vulnerabilities, and although partial solutions to this problem exist there are no complete and comprehensive solutions. In “The AI Transparency Paradox,” Andrew Burt recommends that legal teams be involved from the start of the development process so that “companies [can] thoroughly probe their models for every vulnerability imaginable without creating additional liabilities.” For the models being used today I would agree, but this is not sufficient for the models that many are hoping to create tomorrow. Hopefully, as the field of AI grows, more solutions to this issue will be discovered, but until then AI should be restricted to lower risk applications.

Works Cited

Athalye, Anish, et al. "Synthesizing robust adversarial examples." *International conference on machine learning*, PMLR, 7 Jun. 2018, https://doi.org/10.48550/arXiv.1707.07397. Accessed 15 Apr. 2022.

Burt, Andrew. "The AI Transparency Paradox." *Harvard Business Review*, Harvard Business School Publishing, 13 Dec. 2019, hbr.org/2019/12/the-ai-transparency-paradox. Accessed 16 Apr. 2022.

Code of Ethics. IEEE Computer Society, www.computer.org/education/code-of-ethics. Accessed 15 Apr. 2022.

Code of Ethics. National Society of Professional Engineers, www.nspe.org/resources/ethics/code-ethics. Accessed 15 Apr. 2022.

Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual classification." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 10 Apr. 2018, https://doi.org/10.48550/arXiv.1707.08945. Accessed 15 Apr. 2022.

Fredrikson, Matt, et al. "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures." *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, https://doi.org/10.1145/2810103.2813677. Accessed 15 Apr. 2022.

Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*, 10 Jun. 2014, https://doi.org/10.48550/arXiv.1406.2661. Accessed 15 Apr. 2022.

Hollyer, James R., et al. *Information, Democracy, and Autocracy: Economic Transparency and Political (In)Stability*. Cambridge University Press, 2018.

"Informed Consent: Code of Medical Ethics Opinion 2.1.1." *American Medical Association*, www.ama-assn.org/delivering-care/ethics/informed-consent. Accessed 15 Apr. 2022.

Linardatos, Pantelis, et al. "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy*, vol. 23, no. 1, 2020, p. 18., https://doi.org/10.3390/e23010018. Accessed 15 Apr. 2022.

Lohn, Andrew. "Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity." *Center for Security and Emerging Technology*, Dec. 2020, https://doi.org/10.51593/2020ca006. Accessed 15 Apr. 2022.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems*, 25 Nov 2017, https://doi.org/10.48550/arXiv.1705.07874. Accessed 15 Apr. 2022.

Milli, Smitha, et al. "Model Reconstruction from Model Explanations." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, https://doi.org/10.1145/3287560.3287562. Accessed 15 Apr. 2022.

Müller, Vincent C. "Ethics of Artificial Intelligence and Robotics." *Stanford Encyclopedia of Philosophy*, Stanford University, 30 Apr. 2020, plato.stanford.edu/entries/ethics-ai. Accessed 15 Apr. 2022.

Olah, Chris, et al. "Feature Visualization." *Distill*, 28 Aug. 2019, distill.pub/2017/feature-visualization/. Accessed 15 Apr. 2022.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. ""Why should I trust you?": Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 9 Aug 2016, https://doi.org/10.48550/arXiv.1602.04938. Accessed 15 Apr. 2022.

Schneier, Bruce. “The Coming AI Hackers.” *Belfer Center for Science and International Affairs*, Harvard Kennedy School, Apr. 2021, www.belfercenter.org/publication/coming-ai-hackers. Accessed 15 Apr. 2022.

Shokri, Reza, et al. “On the Privacy Risks of Model Explanations.” *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, https://doi.org/10.1145/3461702.3462533. Accessed 15 Apr. 2022.

Slack, Dylan, et al. “Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods.” *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 3 Feb. 2020, https://doi.org/10.1145/3375627.3375830. Accessed 15 Apr. 2022.

Song, Yang, et al. "Constructing unrestricted adversarial examples with generative models." *Advances in Neural Information Processing Systems*, 2 Dec. 2018, https://doi.org/10.48550/arXiv.1805.07894. Accessed 15 Apr. 2022.

Tramèr, Florian, et al. "Stealing Machine Learning Models via Prediction APIs." *25th USENIX security symposium (USENIX Security 16)*, 3 Oct. 2016, https://doi.org/10.48550/arXiv.1609.02943. Accessed 15 Apr. 2022.

“Transparency in the Workplace: Why It Matters and How to Practice It.” *Glassdoor*, 29 June 2021, www.glassdoor.com/employers/blog/transparency-in-the-workplace/. Accessed 15 Apr. 2022.

Weeks, Anne-Laure Fayard and John, et al. “The Transparency Trap.” *Harvard Business Review*, Harvard Business School Publishing, 28 Oct. 2014, hbr.org/2014/10/the-transparency-trap. Accessed 15 Apr. 2022.

Wheeler, David A. *Secure Programming HOWTO*. v3.72 ed., 2015, dwheeler.com/secure-programs/Secure-Programs-HOWTO/index.html. Accessed 15 Apr. 2022.