

Improving Fairness and Cybersecurity in the Artificial Intelligence Act

Gabriele Carovano¹, Alexander Meinke

¹University of Tübingen, Germany

Abstract

The EU's draft Artificial Intelligence Act (AIA) aims to regulate artificial intelligence (AI) systems, especially so-called 'high-risk AI systems', to ensure that they incorporate EU values and respect EU fundamental rights. However, despite its good intentions, the AIA fails to address basic challenges in the development of high-risk AI systems with regards to both fairness and cybersecurity and thus offers insufficient protections to the public. Specifically, we discuss how fairness, cybersecurity, and accuracy can often be in unavoidable conflict that necessarily leads to ineliminable trade-offs unaccounted by the AIA. Against this backdrop, We propose the creation of a specialised AI institute and offer detailed solutions through new theoretical legal approaches consisting of mathematically computable, principle-based obligations.

Keywords

Artificial Intelligence Act, Fairness, Cybersecurity, Pareto-optimality, Trade-offs

1. Fairness


AI systems are often claimed to be biased or unfair in one way or another, but as is well-known in the algorithmic fairness literature, there is no uniquely agreed-upon definition of the concept [1]. Some definitions (i) require equalising certain prediction metrics across subgroups [2]; some (ii) demand that similar individuals, with respect to the prediction task, are treated similarly [3]; others (iii) impose certain requirements based on a set of causal relationships that are assumed to hold on the task [4]. Yet, despite the many fairness notions proposed thus far, legally qualifying an AI system as fair remains difficult as for many AI applications there exist mathematically unavoidable trade-offs either between different fairness notions or between those systems' accuracy and fairness [5]. Additionally, the described situation gets further complicated by the fact that all fairness definitions require certain assumptions and/or decisions on which attributes or subgroups deserve protection. As a result, effective regulation and supervision are necessary to ensure that high-risk AI developers do not prioritize profit over the public interest by optimizing their systems' fairness towards definitions and subgroups that cause the least degradation to their systems' performance.

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

✉ gabriele.carovano@uni-tuebingen.de (G. Carovano); alexander.meinke@uni-tuebingen.de (A. Meinke)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Cybersecurity

Conventionally, cybersecurity involves the defence against any threats aiming to compromise computer systems and information security, without distinguishing between threat types (e.g., AI-related or non-AI-related) and sources (e.g., harm, theft, or unauthorised use). AI systems, however, cause new cybersecurity risks or modify existing ones. AI-specific cybersecurity threats can be grouped into four categories: (i) adversarial attacks [6] (maliciously modified inputs that cause vastly altered behaviour in an AI system); (ii) data poisoning [7] (malicious modifications to training data); (iii) privacy theft [8] (inferring information about the training data from the model's input-output relation); and (iv) model stealing [9] (inferring information about the model's parameters from its input-output relation). While defending against these threats is in the public's interest, their mitigation is not straightforward for several reasons. Firstly, the precise definition of the threat model often requires many assumptions. Secondly, even under precisely specified threat models, successfully mitigating one attack vector is known to come at the expense of accuracy, other desired properties, or both [10, 7, 11]. Thus again, as for fairness, also for cybersecurity effective regulation and supervision are necessary to ensure that high-risk AI developers do not prioritize profit over the public interest by optimizing their systems' cybersecurity towards threats that minimally interfere with their profit-maximisation objectives.

3. Draft AI Act's Limitations and Proposals

Despite existing ineliminable trade-offs between high-risk AI systems' fairness, cybersecurity, and accuracy, the draft AIA at the time of our paper's acceptance neglected fairness as self-standing legal principle (aside from minor mentions of biased data [12]) and only superficially treated AI-specific cybersecurity issues, mainly through legally undefined terms (e.g., AI specific vulnerabilities, model flaw etc [13]). As a result, the draft AIA resembled a 'blank cheque' that neither ensured legal certainty, nor guaranteed individuals' fundamental rights, nor provided meaningful guidelines to inform AI developers' business decisions and democratic oversight over the latter. Problematically, the original draft AIA left many normative decisions to standardisation organisations despite their lack of democratic representation, risk of regulatory capture, and the unclear judicial control over their operations.

Note that the original version of the draft AIA mandated the creation of a European Artificial Intelligence Board which would have had some limited and (we believe) insufficient authority and expertise to shape the implementation of the AIA by issuing recommendations and opinions about technical specifications and harmonisation standards. However, the most recent amendments to the draft AIA have replaced the AI Board by the "AI Office", which has a significantly expanded scope [14]. In the accepted version of this paper, we proposed an "AI Institute" to assist in the concrete enforcement of the draft AIA. We will now describe our proposal for its duties while still referring to it as the AI Institute, although we expressly do not endorse the creation of a separate legal entity and rather propose to incorporate our suggestions into the operation of the AI Office instead.

The purpose of the AI Institute would be to introduce some democratic oversight into the

setting of the legal standards necessary that objectively measure and evaluate AI systems' fairness and cybersecurity. Crucially, standardisation bodies would still perform their normal functions, except under the supervision of the AI Institute, which could overwrite their choices when deemed needed in the public interest. This differs from the currently proposed AI Office's authority in two key ways. Firstly, the AI Office is only able to "issue opinions, recommendations or written contributions" on technical specifications [14] and its guidance on the matter of robustness and AI-specific cybersecurity is still explicitly "non-binding" [15]. Secondly, its role in defining metrics for discrimination and algorithmic fairness is still at best implicit in the amended draft AIA. We argue that the AI Institute should be able to give binding definitions, benchmarks and thresholds if it deems them necessary in some application or group thereof.

Given the high social impact of such binding decisions from the AI Institute, we also advise subjecting the latter to the review of the EU Institutions (including the European Parliament) to ensure full democratic oversight and political accountability. Concretely, we propose that in cases where the Institute wants to make a metric and threshold of fairness or cybersecurity legally binding for a certain application of AI, that they submit a short proposal to the EU Parliament for approval. This increases democratic oversight for high-stakes decisions while producing small bureaucratic overhead.

Furthermore, we suggest to include within the AIA a duty to commercialise only AI models placed on the Pareto frontier. To illustrate our proposal, we trained AI models predicting job applicants' success by assigning a score to the photos of their faces.¹ We used AI photo screening hiring systems as an illustrative example given that (i) they qualify as high-risk AI systems under the draft AIA (ii) their fairness and adversarial robustness has already been called into question [16]. Importantly, although we only experiment with AI hiring systems, the arguments generalise to other high-risk AI systems. Specifically, Figure 1 illustrates the trade-offs that arise on our dataset. Each point on Figure 1 corresponds to a different model having its own specific trade-off between the model's error (x-coordinate), and the model's unfairness/ the model's susceptibility to adversarial examples (y-coordinate). For each model outside the curve there is a model on the curve that performs better on one or both measures (i.e., it is either more accurate, or fairer/adversarially robust, or both), while not underperforming on the other. It follows that any model that is not on the curve offers a bad trade-off to be forbidden by law.

The technical name of the curves shown in Figure 1 is 'Pareto curve' or 'Pareto frontier'. Since the Pareto curve identifies the entire set of optimal trade-offs existing between accuracy and fairness or cybersecurity, we suggest embedding within the AIA an obligation for AI developers to commercialise only AI systems that are placed on the Pareto curve given legally pre-established protection-worthy subgroups, fairness definition(s), and selection of cyberattacks to protect against (in cases where such measures have been defined by the AI Institute).²

Additional measures could prevent that AI developers' competition gets toxic turning into races-to-the-bottom. For example, when selecting a model on the Pareto frontier for a hiring AI system, society might want to prioritise fairness over accuracy, i.e. the AI Institute could explicitly impose thresholds of the chosen fairness metric that need to be met. When

¹The source code is available under <https://github.com/AlexMeinke/fairness-cybersecurity-draft-aia-experiments>

²If more metrics are deemed desirable the Pareto frontier will be a surface in higher dimensions. A 3D example can be seen under <https://github.com/AlexMeinke/fairness-cybersecurity-draft-aia-experiments> where we show the trade-offs that arise when fairness and cybersecurity are simultaneously enforced.

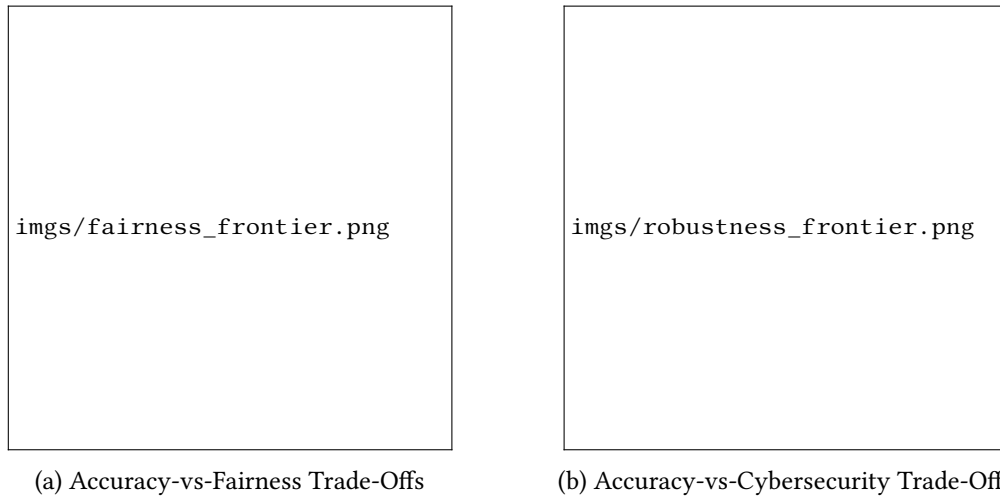


Figure 1: (a) The difference in false negative rates between male and female applicants vs. the total error rate. (b) The error rate against adversarial examples (if each pixel can change by 1%) against the total error rate. Each blue point is a model and the green line is the empirical Pareto-frontier. In both plots we see clear trade-offs.

thresholding or other similar techniques are not introduced by the AI Institute, the AIA should mandate the respect of a default ‘reasonability principle’ to guide AI developers’ selection of the model to commercialise among those on the Pareto curve. The reasonability principle, therefore, is envisioned as a default and residual mechanism guiding AI developers’ discretion when other techniques are not used and potentially in combination with them when the two are not incompatible. Specifically, the principle of reasonability would require AI developers to opt for the model on the Pareto frontier that offers the most reasonable, proportionate, and appropriate trade-off between accuracy, fairness, and cybersecurity considering the technological state-of-the-art. Such a principle, among others, will require AI developers (i) not to opt for excessively accurate but highly unfair or cyber insecure models; (ii) not to opt for highly inaccurate but fairer or more cybersecurity models; (iii) provide relevant evidence and explanations supporting decisions over commercialised models among those on the Pareto frontier; (iv) provide relevant evidence and justifications explaining why models other models on the Pareto frontier were dismissed. The explanations requested by points (i), (ii), (iii), and (iv) should be included in the technical documentation AI developers need to draw before placing their AI systems on the market under Article 11(1) draft AIA. This way, the principle of reasonability will simultaneously realise three socially desirable objectives. Firstly, it ensures that AI developers avoid extreme and socially damaging accuracy-vs-fairness or accuracy-vs-cybersecurity trade-offs. Secondly, it preserves a sufficient competition space among AI developers so as to guarantee that market forces boost innovation and social welfare. Thirdly, it provides legal certainty as it offers a legal framework against which AI developers can ex ante inform their business decisions and judges can ex post assess the legality of the latter.

Finally, note that for general purpose foundation models, the number of measures and benchmarks one could reasonably propose for fairness or cybersecurity is so vast that it seems

implausible that the Pareto-frontiers could be observed in practice. However, as soon as the general purpose model is embedded into a specific high-risk application, it can be evaluated according to task-specific measures and trade-offs apply and thus our proposed Pareto-duty can be implemented.

References

- [1] S. Verma, J. Rubin, Fairness definitions explained, in: 2018 IEEE/ACM International Workshop on Software Fairness (Fairware), IEEE, 2018, pp. 1–7.
- [2] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, *Advances in neural information processing systems* 29 (2016).
- [3] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.
- [4] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, *Advances in neural information processing systems* 30 (2017).
- [5] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 797–806.
- [6] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, *arXiv preprint arXiv:1412.6572* (2014).
- [7] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, T. Goldstein, Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [8] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2017, pp. 3–18.
- [9] T. Orekondy, B. Schiele, M. Fritz, Knockoff nets: Stealing functionality of black-box models, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4954–4963.
- [10] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, *arXiv preprint arXiv:1805.12152* (2018).
- [11] S. Asoodeh, F. Alajaji, T. Linder, Notes on information-theoretic privacy, in: *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2014, pp. 1272–1278.
- [12] European Commission, Draft EU AI Act, 2021. Article 10(2).
- [13] European Commission, Draft EU AI Act, 2021. Article 15(3).
- [14] European Commission, Draft Compromise Amendments to the Draft EU AI Act, 2023. Article 56-58.
- [15] European Commission, Draft Compromise Amendments to the Draft EU AI Act, 2023. Article 15.
- [16] E. Harlan, O. Schnuck, Objective or biased: On the questionable use of artificial in-

telligence for job applications, 2021. Bayerischer Rundfunk <<https://interaktiv.br.de/ki-bewerbung/en/>> accessed 22 May 2022.