The Ethics of Going Deep: Challenges in Machine Learning for Sensitive Security Domains

Aliai Eusebi
UCL
London, UK
aliai.eusebi.16@ucl.ac.uk

Marie Vasek
UCL
London, UK
m.vasek@ucl.ac.uk

Ella Cockbain
UCL
London, UK
e.cockbain@ucl.ac.uk

Enrico Mariconti
UCL
London, UK
e.mariconti@ucl.ac.uk

***Abstract*—Sometimes, machine learning models can determine the trajectory of human life, and a series of cascading ethical failures could be irreversible. Ethical concerns are nevertheless set to increase, in particular when the injection of algorithmic forms of decision-making occurs in highly sensitive security contexts. In cybercrime, there have been cases of algorithms that have not identified racist and hateful speeches, as well as missing the identification of Image Based Sexual Abuse cases. Hence, this paper intends to add a voice of caution on the vulnerabilities pervading the different stages of a machine learning development pipeline and the ethical challenges that these potentially nurture and perpetuate. To highlight both the issues and potential fixes in an adversarial environment, we use Child Sexual Exploitation and its implications on the Internet as a case study, being 2021 its worst year according to the Internet Watch Foundation.**

***Index Terms*—machine learning, ethics, security, online child sexual abuse**

1. Introduction

Ethical concerns mount as machine learning solutions embrace a more influential decision role in sensitive domains. Examples abound, from hate speech detection algorithms operating at the expense of ethnic minorities to flawed child sexual abuse imagery detection tools on the web.[1,2]

Distortions, and thus, unaccountability, can arise from several phases of the pipeline characterising these systems, as for instance, feeding 'dirty data' into model training, selecting opaque instead of transparent models without a significant increase in predictive power, relying on a misleading rationalisation of the model behaviour, or reducing control over the model whose performance is gradually degrading in the operational environment.

Hence, this work tries to identify the different aspects and challenges to be considered when implementing AI-based systems and contributes to the discussion using a case study related to Online Child Sexual Abuse:

- We explore what are poorly suited diagnostics or design constraints that can impact the definition of ethically-neutral algorithm of Morley et al. (2021). They define an ethically-aligned algorithm as to which is: "(a) beneficial to, and respectful of, people and the environment, (b) robust and secure, (c) respectful of human values, (d) fair, and (e) explainable, accountable and understandable."
- We summarise the main issues that need to be considered throughout the different stages of the design and implementation of the machine learning-based systems with a particular mention to the topics of privacy and explainability that are crucial towards the implementation of these systems in extremely sensitive fields.
- We use Online Child Sexual Abuse as case study to understand the challenges that the community has to face when designing machine learning systems in sensitive environments as this poses important ethical questions.
- We conclude our analysis indicating the risks as well as the considerations on how to integrate machine learning systems in such delicate environments.

1. https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/
2. https://www.nytimes.com/interactive/2019/11/09/us/internet-child-sex-abuse.html

2. Related Work

The fight back against the ethical challenges of machine learning has already begun. There are many studies that have been published at the theoretical intersection of machine learning and ethics, with empirical applications in multiple high-stakes domains.

Most notably, Mittelstadt et al. (2016) review the existing discussion of ethical aspects of algorithms to propose a prescriptive map to organise the debate. The map demonstrates that "solving problems at one level does not address all types of concerns; a perfectly auditable algorithmic decision, or one that is based on conclusive, scrutable and well-founded evidence, can nevertheless cause unfair and transformative effects" (Mittelstadt et al., 2016). Moving from concerns to resolutions, Morley et al. (2021) construct an interesting typology with the aim of guiding the application of ethics at each stage of the machine learning development pipeline. The authors ultimately find that existing techniques are either unresponsive to context or vulnerable to ethics washing. Ethics washing is described as "the malpractice of making misleading claims about, or implementing superficial measures in favour of, the ethical values and benefits of digital processes [...] in order to appear more digitally ethical than one is" (Floridi, 2021).

Within medical and judicial applications, Rudin and Ustun (2018) propose two highly interpretable optimised scoring systems (SLIM and RiskSLIM) that do not lose accuracy over black-box models, ultimately questioning the need of complicated and uninterpretable systems. This work suggests that interpretable models can perform competitively if the representation of the problem is adequate, and their behaviour, if unethical, more open to scrutiny and debate. In contrast to model-based interpretability, Mishra et al. (2021) address the question of how explainability can effectively be operationalised within abusive language detection, in view of the ethical challenges of incorporating user and community information into the machine learning development process. Important ethical vulnerabilities are raised, like privacy invasion, demographic bias, knowledge profiling, and faulty generalisation of personal traits at the population level.

3. Machine Learning for Sensitive Operational Environments

Several steps are required in the learning procedure of a machine learning algorithm. Data, which is the food for machine learning systems, need to be prepared to use it in model training. During the model training phase, suitable machine learning algorithms are selected, trained, and optimised to tune data into knowledge. The performance of the machine learning algorithm is then evaluated before possible deployments in the wild. Each of these stages can potentially be susceptible to errors that may tremendously impact the safety of the model in its interactions with human lives.

Table 1 shows a summary of these issues according to the different stages of a machine learning system pipeline.

Stages	Operational Issues
Data Preparation	• Poor data quality • Tradeoff between fairness definitions
Model Selection and Training	• Lack of interpretability • Trade-off between interpretability and predictive power • Biased parameters selection
Model Evaluation	• Misleading XAI explanations • Inappropriate evaluation metrics
Model Deployment	• Domain shift • Feedback loop • Sensitive information leakage

TABLE 1: Operational issues affecting a machine learning pipeline

3.1. Data Preparation

Data represents the critical aspect of any machine learning model, largely impacting its performance, scalability, and fairness. Paradoxically, data is also "the most under-valued and de-glamorised aspect" of the machine learning life-cycle and, as a consequence, its quality is often inadequate because of potentially corrupt, biased, or unlawful practices (Sambasivan et al., 2021). Through the lens of biased data, fairness has been pushed into machine learning as a steadily growing body of research which aims at debiasing machine learning systems to control and mitigate discriminatory behaviours. While formal and mathematical definitions of fairness are found to be controversial and often mutually incompatible (Rudin et al., 2018), they can be broadly categorised into anti-classification, classification parity, and calibration (Corbett-Davies and Goel, 2018). Anti-classification avoids the use of protected and sensitive attributes (e.g. race, gender, or their proxies) in model training. Classification parity requires instead that the model equalises the misclassification metrics between different protected attributes groups. Ultimately, calibration estimates the misclassification risks before defining an outcome that is independent from the protected attributes.

However, while it can feel reasonable to exclude sensitive attributes from model training to safeguard against the explicit dependency between predictions and group membership, Corbett-Davies and Goel (2018) have demonstrated that "anti-classification or classification parity can, perversely, harm the very groups they were designed to protect; and calibration, though generally desirable, provides little guarantee that decisions are equitable." For example, gender-neutral recidivism algorithms can lead to a risk overestimation for women, if women and men have different risk distributions but similar criminal histories - a problem known in statistics as infra-marginality.

Another interesting finding has been recently published by Wachter et al. (2021). Through the analysis of EU non-discrimination law and jurisprudence of the European Court of Justice (ECJ) and national courts, the authors found "a critical incompatibility between European notions of discrimination and existing work on algorithmic and automated fairness." This finding implies that the automation of fairness or non-discrimination in Europe is not feasible because there is "no static or homogenous framework suited to testing for discrimination in AI systems" (Wachter et al., 2021).

Sampling bias and prejudice-based bias are also two other common concerns. Sampling bias arises from a lack of representation for a particular group, whereas prejudice-based bias happens when human bias is reflected in data labels in supervised machine learning. There are multiple diagnostic and mitigation techniques that can be considered, such as adversarial debiasing, data augmentation, or reporting checklists (e.g. PROBAST). Generally, however, there is no one-shot solution and the balance between the notions of fairness is often a difficult tradeoff. Importantly, mathematics can rarely overcome prejudice, especially when the mathematics underlying the model becomes inherently complicated, as we will see in the following section.

3.2. Model Selection and Training

Determining whether a machine learning model is exposed to failures, such as gender or racial bias, for example, can be largely settled by its level of introspection (Rudin, 2019). Machine learning algorithms can be either transparent - i.e., constrained to provide a better visibility on the reasoning underlying predictions - or black-box - i.e., involving functions that are either too complicated for any human to understand or those that are proprietary

(Rudin, 2019). Deep neural networks are classic examples of black-box algorithms, while linear and logistic regression, rules-based learning (Belle and Papantonis, 2021; Rudin, 2019), and decision lists (Rudin, 2019), among others, are typically considered transparent models.

Recent years have seen the proliferation of highly predictive yet complex models leading to the widespread belief that the implementation of a black-box algorithm is imperative in order to achieve a strong predictive performance. Contesting the assumption that black-box models and high predictive power are intertwined, Rudin has instead demonstrated, across multiple domains, that equivalent accuracy can be achieved with interpretable algorithms. Hence, in certain scenarios, interpretable models are not only technically equivalent, but are also a more ethical alternative (Rudin and Radin, 2019).

The fairness of interpretable models can indeed be more easily debated (Rudin, 2019). That because the inherent complex nature of black-box models renders opaque how and why variables interact with each other so as to influence the final prediction. Moreover, black-box models are susceptible to the Clever Hans effect, meaning that the learned model produces correct predictions based on the wrong features (Lapuschkin et al., 2019).

An example of how interpretability of black-box models can "spiral out of control" is the investigation conducted by ProPublica journalists on the COMPAS model for recidivism prediction (Angwin et al., 2016, as cited in Rudin and Radin, 2019). According to Rudin and Radin (2019), ProPublica have erroneously concluded that COMPAS depended on race because its approximation with a linear model depended on race, age, and criminal history. However, when COMPAS is approximated using a nonlinear model, "the explicit dependence on race vanishes leaving dependence on race only through age and criminal history" (Rudin and Radin, 2019).

Hence, the allure of the predictive power of black-box models, rather than their interpretability, becomes questionable and can be dangerous, especially in the security realm, where interpretable models should be preferred as ethical alternatives to black-box models, unless a substantial increase in performance justifies the use of the latter.

Ultimately, distorted results can be caused by tuning hyperparameters or calibrating thresholds on the test data instead of the training data (Arp et al., 2022). Therefore, the parameter selection is biased as learning parameters indirectly depend on the test set. In contrast to other problems considered throughout this paper, biased parameter selection is relatively easy to mitigate by using a separate validation set for model selection and hyper-parameters tuning (Arp et al., 2022).

3.3. Model Evaluation

As a response to the limited interpretability of certain machine learning models, there is a fast-growing area of research referred to explainable artificial intelligence (XAI), which aims to acquire explanations underlying machine learning behaviours. This intent is key in order to evaluate a model's reliability, and consequently, be able to decide whether and how much to trust it. Post-hoc, model-agnostic local explanation tools (e.g. feature attributions, rule lists, and counterfactuals) are at the forefront of XAI (Watson et al., 2021). Several authors have, however, demonstrated inconsistencies between popular XAI methods (Krishna et al., 2022; Kommiya Mothilal et al., 2021; Ramon et al., 2020), questioning their reliability on targeted use-cases and underscoring the importance of evaluating explanation usefulness and actionability to avoid misleading or false characterisations.

We recommend caution also in relation to the choice of the performance metrics, which is highly application-specific. For example, if there is sampling bias present in the data under consideration, then precision and recall may be misleading. In this case, for instance, the Matthews Correlation Coefficient is a more reliable measure to evaluate the model's performance (Arp et al., 2022).

3.4. Model Deployment

In the deployment stage, machine learning models are integrated into a production environment to enhance, if not even drive, decision-making processes. Distortions can result in situations where there is a change in the data distribution between an algorithm's training dataset, and a dataset it encounters when deployed, which is known as a domain shift (Vokinger et al., 2021).

Generally, strategies to control domain shift have relied on performance monitoring, model updating and model calibration, either requiring degradation in the model to be detected or depending upon explicit knowledge or assumptions about the nature of the shift over certain time intervals (Davis et al., 2019, 2020; Siregar et al., 2019). There are other complementary approaches often developed for domain generalisation and unsupervised domain adaptation that instead "learn robust models by using data from multiple environments to identify invariant properties" (Guo et al., 2022).

Sometimes, bias begets bias, which results in a second systematic failure known as a feedback loop. Feedback loops occur when machine learning outcomes influence end-users' practice, so that bias is self-reinforcing itself. Such vulnerabilities ultimately underscore the need to better understand the dynamics of algorithmic decision making in order to develop machine learning solutions that are aligned with their intended scope and remain traceable, accountable, and, to the greatest extent, explainable to people whose lives may be affected by their decisions.

Ultimately, machine learning models can be subject to attacks to privacy, resulting in serious private and/or sensitive information leakage. Recent attempts to defend against such attacks include advanced privacy-enhancing technologies like cryptography and differential privacy as well approaches (like dropout, weight normalization, and dimensionality reduction) used as part of the learning process (mainly, training) to reduce the information available to the adversary (De Cristofaro, 2021).

4. Machine Learning for Social Good: Online Child Sexual Abuse

Noticeably, ethical concerns are set to increase in this space, in particular when the injection of algorithmic forms of decision-making occurs in extremely sensitive security domains - such as child protection and safeguarding

(Keddell, 2019). Machine learning has indeed stepped into the protection of children in the context of online abuse (e.g. CEASE.ai, Childsafe.ai, among others) to support independent clearinghouses and law-enforcement in managing reports of abusive images on a scale that increasingly exceeds their capabilities to take action (Bursztein et al., 2019).

Machine learning applied for the detection of child abuse episodes in the online environment has a direct impact on the individuals under assessment, child safeguarding decision-making process, and resource allocation. Machine learning entering this dimension brings both opportunities and increased risks of unintended consequences. Biased training data is a widely spread issue when talking about machine learning-based systems; this area is subject to this issue as well. In the past, biased training data caused large numbers of false positive in protected categories on which the system had unbalanced training in favour of the positive class. On the other hand, such biases can make some children invisible as the training set does not contemplate these cases.[3]

An issue related to data in cases of Online Child Sexual Abuse is that data silos from safeguarding agencies may inhibit a well-rounded picture of a child's risk.[4] This problem may happen frequently also due to model opaqueness. Model opaqueness causes obscure decision-making that, although may lead to correct decisions, does not allow to understand why the decisions are made and identify any issue causing unidentified risks. Model opaqueness may indeed lead to a lack of oversight and accountability around questions of fairness and discrimination. The significant reliance on highly sensitive data may also pose substantial risks in terms of maintaining individuals' privacy. Where there is little human involvement or even oversight in the automated detection process, the system's accountability and robustness can corrode over time, which can in turn lead to a false or missed detection of child abuse episodes. Moreover, even if a temporary fix can be issued for a problem, adversaries find workarounds that lead to the same risks for children, as for instance, intentionally creating adversarial inputs to elude model detection. Most importantly, "exploitation is a continuum of experiences" and embracing a dichotomous approach to the study of the phenomenon may "take us further away from a more nuanced understanding" (Kjellgren et al., 2022).

Unless we understand how exploitation is operationalised within the machine learning system, "decontextualised numbers are not very informative" (Kjellgren et al., 2022) if not even dangerous. The considerations made in Section 3 can serve as a source of ethical and evidence-informed reflection for the generation and contextualisation of those numbers. Certain machine learning models can provide a more transparent understanding on exploitation indicators underlying the machine learning narrative. This narrative can be also cautiously assessed with XAI strategies, controlled over time with domain adaptation and generation methods, or debiased through data-centric approaches to make more visible, otherwise invisible, children. De Vries and Cockbain (2023) have already identified structural biases and empirical challenges underlying human trafficking indicators in their attempt to simplify a complex problem into quantifiable categories. By reflecting upon indicators' selectiviness and skewness towards specific aspects of human trafficking, the authors find that they risk of perpetuating power imbalances and inequities. This problem significantly influences the reliability of machine learning models if their feature selection procedure is far from being neutral, context-specific and deeply-nuanced.

When treating sensitive data, we cannot consider the use of machine learning as an independent system that we trust and do not understand nor question. Machine learning is an extremely powerful tool that allows to advance quicker and more accurately than what the human can do, however, it cannot be perfect and, therefore, it is necessary to investigate and understand the potential for inaccurate risk evaluations and predictions.

This last aspect highlights even more how the explainability of these systems is crucial as when they are implemented and take decisions related to actual human lives, there must be the opportunity for people using the systems to intervene and be informed by the system. This leads towards seeing the machine learning systems as a useful and trusted support that accelerates the process and identifies in a more efficient way the cases where children may be at risk, but does not replace the human. At the same time, this requires humans to be able to interpret the machine outputs as part of the decision making process.

5. Conclusion

In this work we have looked into machine learning based system security and the implications of defective settings. We have pointed out the main issues that have to be considered when working towards these systems before focusing on when socio-technical distortions can affect ethically sensitive topics. 'Dirty data', black-box models, misleading model explanations, and performance degradation, among others, have caused issues in this field in the past and they can all give rise to ethical challenges, that should be tackled proactively rather than reactively (Morley et al., 2019). This case study confirms what has been seen in other fields as well: the community must tackle these challenges along all the different stages of a machine learning system development and deployment pipeline. Moreover, the systems cannot sacrifice the explainability aspect towards the accuracy one as a tradeoff is necessary for the process to be safeguarded and to understand whether an assessment made by the machine could have been wrong. We therefore proposed considerations to ensure that the convergence between machine learning and online child abuse is ethically aligned to support its potential flourish in full respect with human values.

3. https://www.theguardian.com/technology/2018/sep/19/dont-trust-algorithms-to-predict-child-abuse-risk

4. London Hackney Council abandoned machine learning pilots in child care because of difficulties in matching information across databases: https://www.theguardian.com/society/2019/nov/18/child-protection-ai-predict-prevent-risks

References

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016), Machine bias, *in* 'Ethics of Data and Analytics', Auer-

bach Publications, pp. 254–264.

Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L. and Rieck, K. (2022), Dos and dont's of machine learning in computer security, *in* 'USENIX Security Symposium'.

Belle, V. and Papantonis, I. (2021), 'Principles and practice of explainable machine learning', *Frontiers in big Data* p. 39.

Bursztein, E., Clarke, E., DeLaune, M., Elifff, D. M., Hsu, N., Olson, L., Shehan, J., Thakur, M., Thomas, K. and Bright, T. (2019), Rethinking the detection of child sexual abuse imagery on the internet, *in* 'The world wide web conference', pp. 2601–2607.

Corbett-Davies, S. and Goel, S. (2018), 'The measure and mismeasure of fairness: A critical review of fair machine learning', *arXiv preprint arXiv:1808.00023* .

Davis, S. E., Greevy Jr, R. A., Fonnesbeck, C., Lasko, T. A., Walsh, C. G. and Matheny, M. E. (2019), 'A nonparametric updating method to correct clinical prediction model drift', *Journal of the American Medical Informatics Association* **26**(12), 1448–1457.

Davis, S. E., Greevy Jr, R. A., Lasko, T. A., Walsh, C. G. and Matheny, M. E. (2020), 'Detection of calibration drift in clinical prediction models to inform model updating', *Journal of biomedical informatics* **112**, 103611.

De Cristofaro, E. (2021), 'A critical overview of privacy in machine learning', *IEEE Security and Privacy* **19**(4), 19–27.

De Vries, I. and Cockbain, E. (2023), 'Governing through indicators: Structural biases and empirical challenges in indicator-based approaches to human trafficking research and policy'.

Floridi, L. (2021), Translating principles into practices of digital ethics: Five risks of being unethical, *in* 'Ethics, Governance, and Policies in Artificial Intelligence', Springer, pp. 81–90.

Guo, L. L., Pfohl, S. R., Fries, J., Johnson, A. E., Posada, J., Aftandilian, C., Shah, N. and Sung, L. (2022), 'Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine', *Scientific reports* **12**(1), 1–10.

Keddell, E. (2019), 'Algorithmic justice in child protection: Statistical fairness, social justice and the implications for practice', *Social Sciences* **8**(10), 281.

Kjellgren, R. et al. (2022), 'Good tech, bad tech: Policing sex trafficking with big data', *International Journal for Crime, Justice and Social Democracy* **11**(1), 149–166.

Kommiya Mothilal, R., Mahajan, D., Tan, C. and Sharma, A. (2021), Towards unifying feature attribution and counterfactual explanations: Different means to the same end, *in* 'Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society', pp. 652–663.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S. and Lakkaraju, H. (2022), 'The disagreement problem in explainable machine learning: A practitioner's perspective', *arXiv preprint arXiv:2202.01602* .

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.-R. (2019), 'Unmasking clever hans predictors and assessing what machines really learn', *Nature communications* **10**(1), 1–8.

Mishra, P., Yannakoudakis, H. and Shutova, E. (2021), 'Modeling users and online communities for abuse detection: A position on ethics and explainability', *arXiv preprint arXiv:2103.17191* .

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016), 'The ethics of algorithms: Mapping the debate', *Big Data & Society* .

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2021), From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices, *in* 'Ethics, Governance, and Policies in Artificial Intelligence', Springer.

Morley, J., Machado, C., Burr, C., Cowls, J., Taddeo, M. and Floridi, L. (2019), 'The debate on the ethics of ai in health care: a reconstruction and critical review', *Available at SSRN 3486518* .

Ramon, Y., Martens, D., Provost, F. and Evgeniou, T. (2020), 'A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c', *Advances in Data Analysis and Classification* **14**(4), 801–819.

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215.

Rudin, C. and Radin, J. (2019), 'Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition'.

Rudin, C. and Ustun, B. (2018), 'Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice', *Interfaces* **48**(5), 449–466.

Rudin, C., Wang, C. and Coker, B. (2018), 'The age of secrecy and unfairness in recidivism prediction', *arXiv preprint arXiv:1811.00731* .

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L. M. (2021), "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai, *in* 'proceedings of the 2021 CHI Conference on Human Factors in Computing Systems'.

Siregar, S., Nieboer, D., Versteegh, M. I., Steyerberg, E. W. and Takkenberg, J. J. (2019), 'Methods for updating a risk prediction model for cardiac surgery: a statistical primer', *Interactive cardiovascular and thoracic surgery* **28**(3), 333–338.

Vokinger, K. N., Feuerriegel, S. and Kesselheim, A. S. (2021), 'Mitigating bias in machine learning for medicine', *Communications medicine* **1**(1), 1–3.

Wachter, S., Mittelstadt, B. and Russell, C. (2021), 'Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai', *Computer Law & Security Review* **41**, 105567.

Watson, D. S., Gultchin, L., Taly, A. and Floridi, L. (2021), Local explanations via necessity and sufficiency: unifying theory and practice, *in* 'Uncertainty in Artificial Intelligence', PMLR, pp. 1382–1392.