A Typology of AI Ethics Tools, Methods and Research to Translate Principles into Practices

Jessica Morley*[1], Luciano Floridi[1], Libby Kinsey[2], Anat Elhalal[2]

(1) Oxford Internet Institute, University of Oxford, UK (2) Digital Catapult, UK

Introduction

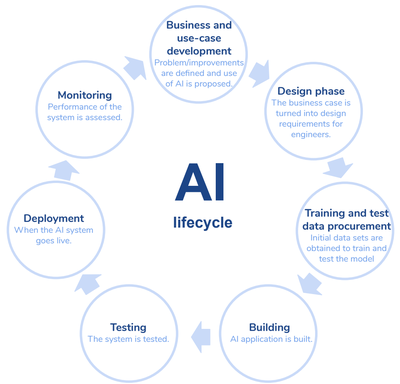
Awareness of the potential ethical issues arising from the development and deployment of machine learning applications is growing at a fast rate and has resulted in a number of AI ethics codes and principles. However, there's a gap between aspiration and viability, and between principle and practice. To fill this gap, methodologies, techniques and processes ('tools') are being developed that seek to operationalise and automate adherence to, and monitoring of, good ethical practices when developing and deploying AI-driven products and services. When should they be used, and what is (or is not) covered? We propose a model, an '**applied ethical AI typology**', on to which we map the tools that are available. Our intention is to help developers, engineers and designers of AI (especially machine learning) 'apply ethics' at each stage of the AI development pipeline, and to signal to researchers where further work is needed.

Methodology

1. Typology design. The first task was to design a typology to organise the tools we identify. The typology is constructed as a grid (Table 1) with 'ethical principles' on one axis and the stages of the 'AI application lifecycle' on the other, to encourage AI developers to go between design decisions and ethical principles regularly.

- **Choice of ethical principles**. A recent review of 84 ethical AI documents [1] found that although no single principle featured in all of them, the themes of transparency, justice & fairness, non-maleficence, responsibility and privacy appeared in over half. Similarly, a systematic review of the literature on ethical technology revealed that the themes of privacy, security, autonomy, justice, human dignity, control of technology and the balance of powers, were recurrent [2]. Taken together these themes 'define' ethically-aligned AI as that which is (a) beneficial to, and respectful of, people and the environment (*beneficence)*; (b) robust and secure (*non-maleficence*); (c) respectful of human values (*autonomy*); (d) fair (*justice*); and (e) explainable, accountable and understandable (*explicability*). Accordingly, these are the principles used in the typology.

- **AI application lifecycle**.The typology uses the seven stages of algorithmic development outlined in the UK's Information Commissioner's Office (ICO) auditing framework for artificial intelligence and its core components [3]. These are, *business and use-case development, design phase, training and test data procurement, building, testing, deployment* and *monitoring.*



Full typology

https://tinyurl.com/AppliedAIEthics (gdoc)

https://appliedaiethics.digicatapult.org.uk/

Table 1: Applied AI ethics typology with illustrative examples of where different tools and methods are plotted (and colour map indicating tool distribution)

	Business and use-case development	Design Phase	Training and test data procurement	Building	Testing	Deployment	Monitoring
Beneficence							
Non-Maleficence		Privacy Design Templates [4]					
Autonomy							
Justice			Data Statements [5, 6]				Audit Studies [7]
Explicability							

2. Identification of tools and methods. A literature review resulted in over *1,000* results each of which was was checked for relevance (either in terms of theoretical framing or in terms of the use of the tool), actionability by AI developers, and generalisability across industry sectors. In total *425* sources that provide a practical or theoretical contribution to the answer of the question: 'how to develop an ethical algorithmic system.' were reviewed.

3. Categorisation. The sources were categorised against the typology

- **Translation.** Each of the high-level principles were translated into tangible system requirements that reflect the meaning of the principles (Table 2).
- **Plotting.** Once this translation process and the literature review were complete, it was possible to plot each of the tools, or methods, reviewed onto the typology by identifying which requirement(s) the tool/methodology in question met and at what stage(s) in the AI application lifecycle it could be implemented or used. Table 1 shows the Applied AI typology containing three example tools.

Table 2: Translation: how system requirements and principles align

Beneficence	Non-Maleficence	Autonomy	Justice	Explicability
Stakeholder participation: to develop systems that are trustworthy and support human flourishing, those who will be affected by the system should be consulted **Protection of fundamental rights** **Sustainable and environmentally friendly AI**: the system's supply chain should be assessed for resource usage and energy consumption **Justification**: the purpose for building the system must be clear and linked to a clear benefit – system's should not be built 'for the sake of it'	**Resilience to attack and security**: AI systems should be protected against vulnerabilities that can allow them to be exploited by adversaries. **Fallback plan and general safety**: AI systems should have safeguards that enable a fallback plan in case of problems. **Accuracy**: for example, the ability documentation that demonstrates evaluation of whether the system is properly classifying results. **Privacy and Data Protection**: AI systems should guarantee privacy and data protection throughout a system's entire lifecycle. **Access to data**: there might be protocols in place governing data access **Reliability and Reproducibility**: does the system work the same way in a variety of different scenarios? **Quality and integrity of the data**: when data is gathered it may contain socially constructed biases, inaccuracies, errors and mistakes – this needs to be addressed. **Social impact**: the effects of system's on people's physical and mental wellbeing should be carefully considered and monitored	**Human agency:** users should be able to make informed autonomous decisions regarding AI systems **Human oversight**: may be achieved through governance mechanisms such as human-in-the-loop, human-on-the-loop, human-in-command.	**Avoidance of unfair bias** **Accessibility and universal design** **Society and democracy**: the impact of the system on institutions, democracy and society at large should be considered. **Auditability**: the enablement of the assessment of algorithms, data and design processes. **Minimisation and reporting of negative impacts**: measures should be taken to identify, assess, document, minimise and respond to potential negative impacts of AI systems. **Trade-offs:** when trade-offs between requirements are necessary, a process should be put in place to explicitly acknowledge the trade-off, and evaluate it transparently. **Redress**: mechanism should be in place to respond when things go wrong.	**Traceability**: the data sets and the processes that yield the AI system's decision should be documented **Explainability**: the ability to explain both the technical processes of an AI system and the related human decisions **Interpretability**

Initial results

We highlight three interrelated findings:

- **Uneven distribution**. The availability of tools and methods is not evenly distributed (Table 1) across the typology either in terms of the ethical principles or in terms of the stages in the application lifecycle. The most noticeable 'skew' is towards post-hoc (i.e. testing phase) 'explanations'.
- **Lack of usability**. The vast majority of categorised tools and methods are not actionable as they offer little help on how to use them in practice [8]. Even when there are open-source code libraries, the available documentation is often limited and the skill-level required for use is high.
- **An individual focus**. Few of the available tools surveyed provide meaningful ways to assess, and respond to, the impact that the data-processing involved in an AI algorithm has on an individual, and even less on the impact on society as a whole [9]. This is evident from the very sparsely populated 'deployment' column of the typology.

Taken together, it is clear that it is not possible for a practitioner to consult the typology and find the tools that he or she needs, or that society demands. Applying ethics still requires considerable amounts of effort, undermining one of the main aims of developing and using technologically-based 'tools': to remove friction from applied ethics.

Discussion

- **A snapshot**. The typology contains a snapshot of what tools are currently available (and which we were able to find) to AI developers to encourage the progression of ethical AI from principles to practice and to signal clearly where further work is needed. We do not claim that the typology is 'complete' nor that the identified tools are the best, or indeed the only, means of 'solving' each of the individual ethical problems. It would be entirely possible to complete the process using a different set of principles and requirements.
- **Use**. The typology is now searchable and updateable so that developers can look for the available tools and methodologies for their given context. It is not intended to act in a deontological sense i.e. as means of translating the principles into definitive 'rules' that technology developers should adhere to, nor to imply that developers must always complete one 'task' from each of the boxes.
- **Continued, coordinated effort is required**. Practitioners want these applied AI ethics resources [10] and widespread adoption requires them to be practical (accessible and easy-to-use). While tools remain immature (undocumented and untested) it is also difficult to assess their scope of use, and consequently, hard to encourage their adoption. Areas of the typology with few tools invite further work to translate ethical principles into design protocols.

References

[1] Anna Jobin, Marcello Ienca, and Effy Vayena. Artificial intelligence: the global landscape of ethics guidelines. URL https://arxiv.org/abs/1906.11668
[2] Lambèr Royakkers, Jelte Timmer, Linda Kool, and Rinie van Est. Societal and ethical issues of digitization. 20(2):127–142. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-018-9452-x. URL http://link.springer.com/10.1007/s10676-018-9452-x
[3] Reuben Binns. An overview of the auditing framework for artificial intelligence and its core components. URL https://ai-auditingframework.blogspot.com/2019/03/an-overview-of-auditing-framework-for_26.html
[4] Theeraporn Suphakul and Twittie Senivongse. Development of privacy design patterns based on privacy principles and UML. In 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pages 369–375. IEEE. ISBN 978-1-5090-5504-3. doi: 10.1109/SNPD.2017.8022748. URL http://ieeexplore.ieee.org/document/8022748/
[5] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. 6:587–604. ISSN 2307-387X. doi: 10.1162/tacl_a_00041. URL https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00041
[6] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. URL http://arxiv.org/abs/1805.03677
[7] C Sandvig, K Hamilton, K Karahalios, and C Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms.
[8] Ville Vakkuri, Kai-Kristian Kemell, Joni Kultanen, Mikko Siponen, and Pekka Abrahamsson. Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. . URL http: //arxiv.org/abs/1906.07946.
[9] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. URL http://arxiv.org/abs/ 1802.07810.
[10] C Miller and R Coldicott. People, power and technology: The tech workers' view. URL https: //doteveryone.org.uk/report/workersview/.

Full reading and resource list: https://medium.com/@jessicamorley/applied-ai-ethics-reading-resource-list-ed9312499c0a

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada