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

Abstract

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? Our intention in presenting this research is to contribute to closing the gap between principles and practices by constructing a typology that may help practically-minded developers 'apply ethics' at each stage of the AI development pipeline, and to signal to researchers where further work is needed. We found that there is an uneven distribution of effort in the applied AI ethics space, and that the stage of maturity (readiness for widespread use) of the identified tools is mostly low.

1 Introduction

Machine learning algorithms are powerful [1] socio-technical constructs that raise concerns that are as much (if not more) about people as they are about code [2]. Enabling the so-called dual advantage of 'ethical ML' — so that the opportunities are capitalised on, whilst the harms are foreseen and minimised or prevented [3] — requires asking difficult questions about design, development, deployment, practices, uses and users, as well as the data that fuel the whole process [4].

Much effort to date has been focused on the 'what' of ethical AI (i.e. debates about principles and codes of conduct) but there has been less attention on the 'how' of applied AI ethics (the tools and methodologies that can be used to help embed principles in practice). Thus, the aim of this research project is to identify the methods and tools available to help developers, engineers and designers of AI (especially machine learning) reflect on and apply 'ethics' [5] so that they may know not only what to do or not to do, but also how to do it, or avoid doing it [6].

We propose a model, an **'applied ethical AI typology'**, to map the space of applied AI ethics, identify and categorise available tools in the typology, and comment on what we find.

2 Methodology

Typology design. With the aim of identifying the methods and tools available to help developers, engineers and designers to reflect on and apply 'ethics' in mind, the first task was to design a typology, for the very practically minded AI community [7], that would 'match' the tools and methods we identify to ethical principles. Inspired by Saltz and Dewar (2019) [8] (who produced a framework that is meant to help data scientists consider ethical issues at each stage of a project), the typology

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

is constructed as a grid 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.

- **Ethical principles.** A recent review of 84 ethical AI documents [9] 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 [10]. 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 [11]. These are, *business and use-case development, design phase, training and test data procurement, building, testing, deployment* and *monitoring*.

Identification of tools and methods. As this is phase one of the research project, the intention was to provide a broad overview of the current state of play, rather than a qualitative analysis of 'what tools or methods are currently in use and why' (this is provided by [12]). Therefore we chose to use the traditional method of providing an overarching assessment of a research topic – a literature review.

Scopus, arXiv and PhilPapers[1], as well as Google Search were searched. More information about the search terms and categories can be found in Table 2 of the appendix. The original searches were run in February 2019, but weekly alerts were set for all searches and reviewed up until mid-July 2019. Every result (of which there were originally over 1,000) 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.

Categorisation. The third, and final, task was to review the recommendations, theories, methodologies, and tools outlined in the reviewed sources, and identify where they may fit in the typology. To do this, each of the high-level principles (*beneficence, non-maleficence, autonomy, justice* and *explicability*) were translated into tangible system requirements that reflect the meaning of the principles [13; 14].

The translation requires a substantial change in the level of abstraction from mid-level ethics (principles) to what [15] refers to as 'microethics.' It is a process that gradually reduces the indeterminacy of abstract norms to produce a desiderata for a 'minimum-viable-ethical-(ML)product (MVEP) that can be used by people who have various disciplinary backgrounds, interests, and priorities [16]. The outcome of this translation is in Table 3 of the appendix.

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 examples of tools/methods. The appendix also includes notes on how the tools/methods were categorised.

3 Discussion of initial results

The fully-populated typology is too large to include here, but it is available at `tinyurl.com/appliedAIethics`.

[1] `https://www.scopus.com/home.uri`, `https://arxiv.org/` and `https://philpapers.org/`

Table 1: Applied AI ethics typology with illustrative examples of where different tools and methods are plotted.

	Business and use-case development Problem / improvements are defined and use of AI is proposed	**Design Phase** The business case is turned into design requirements for engineers	**Training and test data procurement** Initial data sets are obtained to train and test the model	**Building** AI application is built	**Testing** The system is tested	**Deployment** When the AI system goes live	**Monitoring** Performance of the system is assessed
Beneficence							
Non-Maleficence		Privacy Design Templates					
Autonomy							
Justice			Data Statements				Audit Studies
Explicability							

Interpretation of the results of the literature review and the resulting typology are likely to be context specific. Those with different disciplinary backgrounds (engineering, moral philosophy, sociology etc.) will see different patterns, and different meaning in these patterns. This kind of multidisciplinary reflection on what the presence or absence of different tools and methods, and their function, might mean, is to be encouraged. To start the conversation, this section highlights the following three inter-related findings:

1. an over-reliance on 'explicability'
2. a focus on the need to 'protect' the individual over the collective; and
3. a lack of usability

Explicability as the all-encompassing principle. The most obvious observation is that the availability of tools and methods is not evenly distributed 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 'explanations' with individuals seeking to meet the principle of explicability during the testing phase having the greatest range of tools and methods to choose from.

Two complementary reasons for this stand out. First, the 'problem' of 'interpreting' an algorithmic decision has appeared tractable from a mathematical standpoint and thus the principle of explicability has come to be seen as most suitable for a technical fix [15]. Second, 'explicability' is not, from a moral philosophy perspective, a moral principle like the other four principles. Instead, it can be seen as a second order principle, that has come to be of vital importance in the ethical-AI community because, to a certain extent, it can be seen as encompassing all the other four principles. Indeed, it is argued that if a system is explicable (explainable and interpretable) it is inherently more transparent and therefore more accountable in terms of its decision-making properties and the extent to which they include human oversight and are fair, robust and justifiable [17; 18; 19].

An individual focus. The next observation of note is that 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 [20]. This is evident from the very sparsely populated 'deployment' column of the typology. Its emptiness implies that the need for pro-ethically designed human-computer interaction (at an individual level) or networks of AI systems (at a group level) has been paid little heed. This is most likely because it is very difficult to translate complex human behaviour into simple to use, generalisable design tools.

Tools that, for example, help developers pro-ethically design solutions that do not overly restrict the user's options in acting on a prediction (i.e. tools that promote the user's autonomy) are in short supply [21]. If users feel as though their decisions are being too curtailed and controlled by systems that they do not understand, it is very unlikely that these systems will meet the condition of social acceptability, never mind the condition of social preferability which should be the aim for truly ethically designed AI [22].

A 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 [23]. Even when there are open-source code libraries available documentation is often limited and the skill-level required for use is high.

This overarching lack of usability of the tools and methods highlighted in the typology means that, although they are promising, they require more work before being 'production-ready.' As a result, applying ethics still requires considerable amounts of effort by AI developers undermining one of the main aims of developing and using technologically-based 'tools': to remove friction from applied ethics.

Furthermore, until these tools are embedded in practice and tested in the 'real world' it is extremely unclear what impact they will have on the overall 'governability' of the algorithmic ecosystem. For example, [24] asks how will a system actually be held accountable for an 'unfair' decision in a way that is acceptable to all? This makes it almost impossible to measure the impact, 'define success', and document the performance [25] of a new design methodology or tool. As a result, there is no clear problem statement (and therefore no clear business case) that the AI community can use to justify time and financial investment in developing much-needed tools and techniques that truly enable pro-ethical design. Consequently, there is no guarantee that the tools do anything other than help the groups in society who already have the loudest voices embed and protect their values in design tools, and then into the resultant ML systems.

4 Conclusions

The purpose of presenting this research is not to imply that the typology is 'complete' nor that the identified tools and methodologies are the best, or indeed the only, means of 'solving' each of the individual ethical problems. How to apply ethics to the development of AI is an open question that can be solved in a multitude of different ways at different scales and in different contexts [26]. It would, for example, be entirely possible to complete the process using a different set of principles and requirements. Instead, the goal is to provide a brief snapshot of what tools are currently available to AI developers to encourage the progression of ethical AI from principles to practice and to signal clearly, to the 'ethical AI' community at large, where further work is needed.

Additionally, the purpose of presenting the typology is not to give the impression that the tools act in a deontological sense i.e. as means of translating the principles into definitive 'rules' that technology developers should adhere to, or that developers must always complete one 'task' from each of the boxes. This only promotes ethics by 'tick-box' [15]. Instead, the typology is intended to eventually be searchable so that developers can look for the appropriate tools and methodologies for their given context and use them to enable a shift from a prescriptive 'ethics-by-design' approach to a dialogic pro-ethical design approach [13; 27].

It is possible to design things to be better [28], but this will require more coordinated and sophisticated approaches [29] to translating ethical principles into design protocols [30]. This call for increased coordination is necessary as this research has shown that there is uneven distribution of effort across the 'Applied AI Ethics' typology.

AI practitioners want these applied AI ethics resources [31] and widespread adoption requires them to be practical (accessible and easy-to-use). While tools remain immature (undocumented and untested) it is difficult to assess their scope of use (resulting in the 'moral-semantic trilemma' [32]), and consequently hard to encourage their adoption by the practically-minded AI community.

Constructive patience needs to be exercised, by society and by the ethical AI community, because such progress on the question of 'how' to meet the 'what' will not be quick, and there will definitely be mistakes along the way. Only by accepting this can society be positive about seizing the opportunities presented by AI, whilst remaining mindful of the potential costs to be avoided [3].

Acknowledgements

This work was supported by Digital Catapult

References

[1] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. 20(3):973–989. ISSN 1461-4448, 1461-7315. doi: 10.1177/1461444816676645. URL http://journals.sagepub.com/doi/10.1177/1461444816676645.

[2] Kate Crawford and Ryan Calo. There is a blind spot in AI research. 538(7625):311–313. ISSN 0028-0836, 1476-4687. doi: 10.1038/538311a. URL http://www.nature.com/doifinder/10.1038/538311a.

[3] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. AI4people—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. 28(4):689–707. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-018-9482-5. URL http://link.springer.com/10.1007/s11023-018-9482-5.

[4] Corinne Cath, Michael Zimmer, Stine Lomborg, and Ben Zevenbergen. Association of internet researchers (AoIR) roundtable summary: Artificial intelligence and the good society workshop proceedings. 31 (1):155–162. ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-018-0304-8. URL http://link.springer.com/10.1007/s13347-018-0304-8.

[5] Greg Adamson, John C. Havens, and Raja Chatila. Designing a value-driven future for ethical autonomous and intelligent systems. 107(3):518–525. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2018.2884923. URL https://ieeexplore.ieee.org/document/8610013/.

[6] Majed Alshammari and Andrew Simpson. Towards a principled approach for engineering privacy by design. In Erich Schweighofer, Herbert Leitold, Andreas Mitrakas, and Kai Rannenberg, editors, *Privacy Technologies and Policy*, volume 10518, pages 161–177. Springer International Publishing. ISBN 978-3-319-67279-3 978-3-319-67280-9. doi: 10.1007/978-3-319-67280-9_9. URL http://link.springer.com/10.1007/978-3-319-67280-9_9.

[7] Andreas Holzinger. From machine learning to explainable AI. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 55–66. IEEE. ISBN 978-1-5386-5102-5. doi: 10.1109/DISA.2018.8490530. URL https://ieeexplore.ieee.org/document/8490530/.

[8] Jeffrey S. Saltz and Neil Dewar. Data science ethical considerations: a systematic literature review and proposed project framework. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-019-09502-5. URL http://link.springer.com/10.1007/s10676-019-09502-5.

[9] Anna Jobin, Marcello Ienca, and Effy Vayena. Artificial intelligence: the global landscape of ethics guidelines. URL http://arxiv.org/abs/1906.11668.

[10] 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.

[11] 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.

[12] Ville Vakkuri, Kai-Kristian Kemell, and Pekka Abrahamsson. Implementing ethics in AI: An industrial multiple case study. . URL http://arxiv.org/abs/1906.12307.

[13] Icy Fresno Anabo, Iciar Elexpuru-Albizuri, and Lourdes Villardón-Gallego. Revisiting the belmont report's ethical principles in internet-mediated research: perspectives from disciplinary associations in the social sciences. 21(2):137–149. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-018-9495-z. URL http://link.springer.com/10.1007/s10676-018-9495-z.

[14] Karolina La Fors, Bart Custers, and Esther Keymolen. Reassessing values for emerging big data technologies: integrating design-based and application-based approaches. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-019-09503-4. URL http://link.springer.com/10.1007/s10676-019-09503-4.

[15] Thilo Hagendorff. The ethics of AI ethics – an evaluation of guidelines. URL http://arxiv.org/abs/1903.03425.

[16] Naomi Jacobs and Alina Huldtgren. Why value sensitive design needs ethical commitments. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-018-9467-3. URL http://link.springer.com/10.1007/s10676-018-9467-3.

[17] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–14. ACM Press. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173951. URL http://dl.acm.org/citation.cfm?doid=3173574.3173951.

[18] Corinne Cath. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. 376(2133):20180080. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2018.0080. URL http://rsta.royalsocietypublishing.org/lookup/doi/10.1098/rsta.2018.0080.

[19] Zachary C. Lipton. The mythos of model interpretability. URL http://arxiv.org/abs/1606.03490.

[20] 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.

[21] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions*. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjx032. URL http://academic.oup.com/qje/article/doi/10.1093/qje/qjx032/4095198/Human-Decisions-and-Machine-Predictions.

[22] Luciano Floridi and Mariarosaria Taddeo. What is data ethics? 374(2083):20160360. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2016.0360. URL http://rsta.royalsocietypublishing.org/lookup/doi/10.1098/rsta.2016.0360.

[23] 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.

[24] Reuben Binns. Algorithmic accountability and public reason. 31(4):543–556, . ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-017-0263-5. URL http://link.springer.com/10.1007/s13347-017-0263-5.

[25] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. pages 220–229. doi: 10.1145/3287560.3287596. URL http://arxiv.org/abs/1810.03993.

[26] Luciano Floridi. Establishing the rules for building trustworthy AI. . ISSN 2522-5839. doi: 10.1038/s42256-019-0055-y. URL http://www.nature.com/articles/s42256-019-0055-y.

[27] Luciano Floridi. *The logic of information: a theory of philosophy as conceptual design*. Oxford University Press, 1st edition edition, . ISBN 978-0-19-883363-5.

[28] Luciano Floridi. The logic of design as a conceptual logic of information. 27(3):495–519, . ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-017-9438-1. URL http://link.springer.com/10.1007/s11023-017-9438-1.

[29] Colin Allen, Gary Varner, and Jason Zinser. Prolegomena to any future artificial moral agent. 12(3):251–261. ISSN 0952-813X, 1362-3079. doi: 10.1080/09528130050111428. URL http://www.tandfonline.com/doi/abs/10.1080/09528130050111428.

[30] Matteo Turilli. Ethical protocols design. 9(1):49–62. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-006-9128-9. URL http://link.springer.com/10.1007/s10676-006-9128-9.

[31] C Miller and R Coldicott. People, power and technology: The tech workers' view. URL https://doteveryone.org.uk/report/workersview/.

[32] Marcus Arvan. Mental time-travel, semantic flexibility, and a.i. ethics. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-018-0848-2. URL http://link.springer.com/10.1007/s00146-018-0848-2.

[33] 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/.

[34] 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.

[35] 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.

[36] C Sandvig, K Hamilton, K Karahalios, and C Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms.

Appendix

This appendix contains further details about how applied AI ethics tools and methods were identified and how they were categorised in the typology.

It is intended that the typology will form an online searchable (and update-able) database.

Identification of tools and methods

Table 2 contains further information on the search terms and categories used in the literature review undertaken to identify sources that provided a practical or theoretical contribution to the answer of the question: 'how to develop an ethical algorithmic system?'.

Table 2: Showing the search terms used to search Scopus, arXiv and Google and the categories reviewed on PhilPapers

Scopus, Google and arXiv Search Terms (all searched with AND Machine Learning OR Artificial Intelligence)	Category of PhilPapers reviewed
Ethics	Information Ethics
Public Perception	Technology Ethics
Intellectual Property	Computer Ethics
Business Model	Autonomy in Applied Ethics
Evaluation	Beneficence in Applied Ethics
Data Sharing	Harm in Applied Ethics
Impact Assessment	Justices in Applied Ethics
Privacy	Human Rights in Applied Ethics
Harm	Applied Ethics and Normative Ethics
Legislation	Responsibility in Applied Ethics
Regulation	Ethical Theories in Applied Ethics
Data Minimisation	
Transparency	
Bias	
Data protection	

Categorisation

This section contains the principles-to-systems requirements translation (Table 3), and gives three examples of applied AI ethics tools and methods to illustrate how the systems requirements are used to categorise each tool/methodology in the applied AI typology (Table 1, main paper).

Here are the three illustrative examples:

Privacy Design Templates. Suphakul & Senivongse [33] set out a series of privacy design patterns which include details on how to apply them to the design of software applications. These patterns meet the requirements of privacy and data protection (non-maleficence) and are meant to be used during the design phase of a system and so are plotted where non-maleficence and design intersect.

Data Statements. Bender Friedman [34] and Holland, Hosny, Newman, Joseph, and Chmielinski [35] have created a framework for 'data statements' that should be filled in at the time of dataset

Table 3: Translation: how system requirements and principles align

Principles	Beneficence	Non-Maleficence	Autonomy	Justice	Explicability
Requirements	**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**

curation for natural language processing research to help alleviate issues related to exclusion and bias. Therefore, it is plotted at the intersection between justice and training.

Audit Studies. Sandvig and colleagues [36] outline 5 different audit study designs that can be used to investigate instances of algorithmic bias: code study, non-invasive user audit, scraping audit, sock puppet audit, crowdsourced audit. These meet the requirement of avoiding bias (justice) and are designed to be used for the purpose of monitoring systems that have already been deployed. Consequently, the two methods are plotted at the intersection between justice and monitoring.