From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices

Authors: Jessica Morley, Luciano Floridi, Libby Kensey & Anat Elhalal

Published:
11 December 2019

Presentation by:
Sharukh Afridi

23 February 2023

ABSTRACT

- Debate on ethical implications of Artificial Intelligence dates back from the 1960s, but it wasn't actively discussed. However, in recent times with the emergence of Deep Neural Networks and Machine Learning techniques, the impact of these techniques has massively increased on the society, which has resulted in this debate going mainstream.
- Such debates primarily focus on **'What ?'** of the AI ethics rather than '**How ?'**. Through this paper, the authors aim to close this gap between the principles of AI ethics and practices of AI ethics by constructing a typology that can help practically minded developers to apply ethics at every stage of Machine Learning development pipeline and to signal to researchers where further work is needed.

GOALS OF THE STUDY

The research aims to mitigate the below stated ethical concerns of AI:

Ethical Concern	Explanation
Inconclusive evidence	Algorithmic conclusions are probabilities and hence not infallible. For instance, even if a credit worthiness algorithm is 99% accurate, it still means that 1 in 100 capable applicants would be denied credit wrongly.
Inscrutable evidence	Commonly known as the black-box issue, a lack of transparency can lead to algorithmic systems that are hard to control, monitor and correct.
Misguided evidence	Predictions can only be as reliable as the data the model is trained on. So, if the data presented to the model is biased, it is then bound for the model to be bias in its future predictions as well.
Unfair outcomes	An action could be found to be discriminatory if it has a disproportionate impact on one group of people. Predictive policing tools are leading to more people of color being arrested, jailed or physically harmed by the cops.
Transformative effects	Unethical AI algorithms can pose challenges towards informational privacy. A study on the level of access to personal data required by 1000 'free' and paid apps indicated that, the 'free' apps requested for significantly more data, indicating their business model is to sell the personal data of the users.
Traceability	It is difficult to assign responsibility in algorithmic harms, leading to issues in moral responsibility. For instance, it would be unclear as to who or what is responsible for self-driving car fatalities.

METHODOLOGY

- **Step 1-** The first task towards how to implement the AI ethics was to design a typology that would match the existing tools and methods to corresponding ethical principles that it targets such as beneficence, non-maleficence, autonomy, justice and explicability. To create this typology, the ethical principles were combined with the different stages of algorithmic development for Artificial Intelligence and its core components, as shown below.

'Applied AI Ethics' typology comprising the stages of algorithmic development and ethical principles' :

Business and use-case development	**Design phase**	**Training & test data procurement**	**Building**	**Testing**	**Development**	**Monitoring**
Beneficence	Non-Maleficence		Autonomy	Justice	Explicability	

METHODOLOGY

- **Step 2 -** The next step was the identification of the tools and methods and subsequently the company or the individuals researching and producing them. This involved seeking answer to the question, 'what practices, tools or methods, if any, do industry professionals utilize to implement ethics into AI design and development?' by conducting interviews at companies that develop AI systems in different fields. The analysis revealed that though the developers were aware of the AI ethics, the companies seemed to provide them with no tools or methods for implementing the AI ethics.
- Based on a hypothesis that these findings did not imply the non-existence of applied-ethics tools and methods, but rather a lack of progress in the translation of available tools and methods from academic literature or early-stage development and research, to real-life use, this study used the traditional approach of providing an overarching assessment of a research topic.
- Scopus, arXiv and PhilPapers, as well as Google search were explored. Every result (of which there were originally over 1000) was checked for *relevance*—either in terms of theoretical framing or in terms of the use of the tool, which is *actionability* by AI developers, and *generalizability* across industry sectors. In total, 425 sources were reviewed. They provided a practical or theoretical contribution to the answer of the question: 'how to develop an ethical algorithmic system.

METHODOLOGY

- **Step 3 –** The third and final step was to review the recommendations, theories, methodologies, and tools outlined in the reviewed sources, and identify where they may fit in the typology. In order to achieve this, each of the principles (beneficence, non-maleficence, autonomy, justice and explicability) were translated into tangible system requirements.
- This is the approach taken by the EU's High Level Ethics Group for AI which offers guidance on the implementation and realization of trustworthy AI, via a list of seven requirements that should be met. Such a translation aims at gradually reducing the uncertainty of abstract norms to produce 'Minimum-Viable-Ethical-Product' (MVEP), that can be used by people who have various disciplinary backgrounds, interests and priorities

KEY ASPECTS OF THE RESEARCH DESIGN

- The goal of the developed typology is to provide a synthesis 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 and to be an online searchable database 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 pro-ethical design approach. Interpretation of the research and the corresponding literature review is bound to be subjective, where people with different disciplinary backgrounds (engineering, moral philosophy, sociology etc.) will see different patterns, and different meanings in these patterns. The outcomes can be outlined under 3 headings:

Explicability as the All-Encompassing Principle

An Individual Focus

Usability

FUTURE WORK – A WAY FORWARD

- Bridging together multi-disciplinary researchers into the development process of pro-ethical design tools and methodologies will be essential. A multi-disciplinary approach will help the ethical AI community overcome obstacles concerning social complexity and embrace uncertainty.
- In a digital context, ethical principles are not simply either applied or not, but regularly re-applied or applied differently, as algorithmic systems are developed, deployed, configured, tested, revised and re-tuned. This approach of regular reflection and application will rely heavily on –
  I. The creation of more tools – to cover the less focused areas of ethical significance and to cater to larger audience rather than to specific group of people.
  II. Acceleration of the tools' maturity level from research labs into production environments.
- To achieve the above stated couple of points, the society needs to come together in communities comprised of multidisciplinary researchers, including innovators, policy makers, citizens, developers and designers.

EVALUATION - STRENGTH

- The first and foremost strength of the research has to be the identification of the root cause of non ethical AI and dealing with it at the source, as they say, 'prevention is better than cure'. The AI developers are presented with the typology listing out the tools and methods that they can follow while developing the model at different stages. If the model is developed considering the ethical tools and methods, there is minimum possibility of the final product being unethical.
- Next major strength of the study is the utilization of existing tools and methods. Instead of devising new methods or tools, by making use of existing methods, it makes the credibility better, rather than having people to try new un-tested methods. This creates a sense of trust among people towards this process.
- Lastly, the study evaluates the similar existing methods and puts together a list of curated methods and tools. Sometimes, with a lot of options comes greater confusion. A curated database makes the selection of methods easy and further increases the use of such methods, which would previously be ignored just because of the lack of clarity on which method is better over the other.

EVALUATION - WEAKNESS

- The first limitation is that "What resources and techniques are accessible to AI developers to "apply" ethics at each stage of the AI system design?" has a wide scope. Due to the lack of precision, the amount of literature available was overwhelming and constantly expanding, compromising what is realistically necessary. It also very likely that a number of proprietary applied ethics tools and methods being developed by private companies for internal purposes would not be publicly available, hence they might be missing from the study and thus the typology.
- The tools and methods are identified for distinct AI development stages. In practice, such theoretical distinction between different stages of technological development does not always exist. Hence by categorizing the tools by stages of development, reduces their usability as the development process might involve multiple stages running in parallel.
- There is lack of clarity how the tools and methods have been selected. Governments are increasingly setting standards for ethical AI. Due to the lack of clarity on how these methods have been identified, it might not be adopted legally as it might not withstand the scrutiny and cross countering of pro AI activists who do not care for ethics.

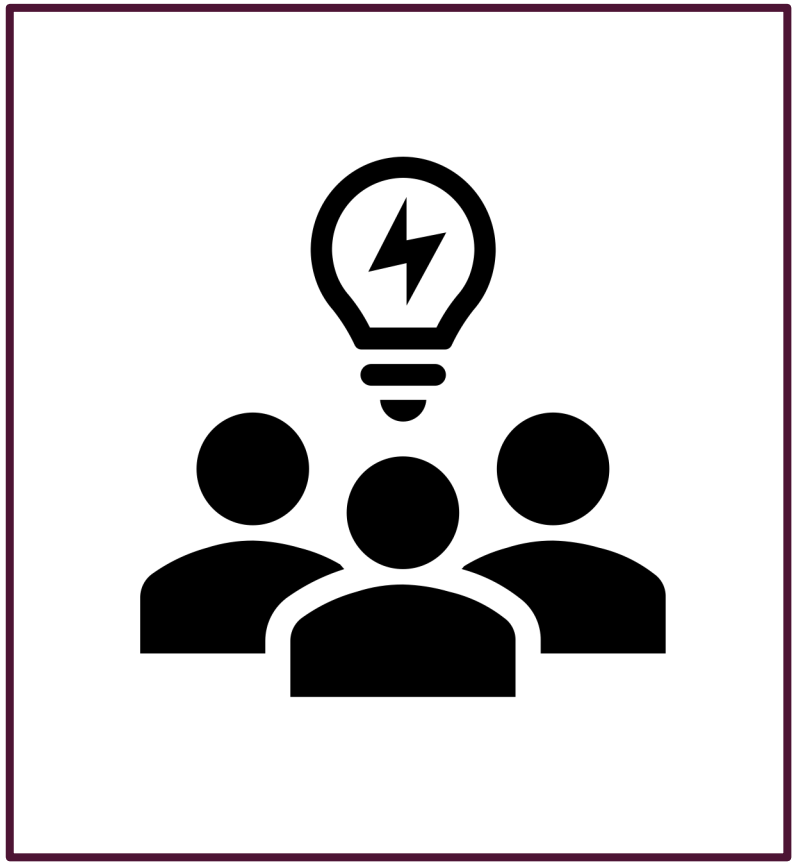
CONCLUSION

- The realization and discussions on ethical considerations in the design of AI algorithms is not new, however with the increase in the complexity and reach of the AI algorithms in the day-to-day life of the people, it becomes necessary to be more critical and convert the theoretical principles of AI ethics into practical conclusion.

- The need for ethical development of AI algorithm by solving the problem at the root has raised from the witnessed consequences such as bias against people of color in judicial prosecutions, denying credit to capable applicants, etc.

- There must be support and patience from the ethical AI community for this approach, as filling the gap between ‘what’ and ‘how’ can not be quick. Adopting the approach mentioned in this study may not be a 100% efficient, but it is rather much better than dealing with consquences of doing nothing at all.

PERSONAL THOUGHTS & POINTS OF DISCUSSION



- The introduction of methods and tools for ethics in the AI algorithm development process will result in cost and time overhead. This overhead might not be commercially viable for a lot of organizations in the short-term. Long term benefits and sustainability, where citizens are now more aware of the consequences of unethical AI, must be focused.
- A common set of AI standards and methods must be agreed upon globally. In this age of internet and connectivity across the world, it becomes extremely essential to have uniform protocols, as the software developed in one country will not just be utilized in that country, but across the world.
- If current methods are not acceptable by the majority, new tools and methods of implementing ethical AI development must be encouraged, where all stakeholders – developers, company, government and citizens have meaningful opportunity to have their concerns addressed.
- Black-boxing of AI algorithms must be eliminated or minimized. All stakeholders must be aware of what is happening behind the 'predictions' made by this algorithm wherever possible.

THANK YOU!

Questions, Concerns ?