Ethical and Commercial Benefits of Transparency in AI

University of Oulu
Information Processing Science
Bachelor's Thesis
Artturi Heikkilä
2024

Abstract

The more the field of AI has advanced the opaquer and harder to understand the systems have become. While deep learning technology is more mainstream than ever thanks to services such as ChatGPT the understanding of how they work to even those who create them is poor. This lack of explainability is first and foremost an ethical issue but in this paper, I argue that it also hurts the economic potential of the field.

In this study I conduct a literature survey to find out how an increase of transparency could help the field of AI both ethically and economically. I seek to use sources in literature to present examples of problems and opportunities in both domains that could be fixed and claimed by improving the transparency.

Keywords: AI Ethics, AI Transparency, XAI

Supervisor:

Associate Professor
M3S Research Unit University of Oulu
Khan, Arif Ali

Contents

1. Introduction

1.1 AI Ethics and modern machine learning

AI ethics is not a new field. Its existence can be traced as far back as the fifties (Borenstein et. al. 2021). However, with the new wave of machine learning AI technology AI ethics has become hotter topic than ever. In, (Borenstein et. al. 2021) the authors have included a table (p. 97) that demonstrates that the academic interest in AI ethics has exploded in the past decade.

The new technologies that we people are now simply calling AI is progressing faster than ever and because of that a lot of parties such as researchers, private companies and lawmakers share a worry of the ethics of AI (Jobin et. Al. 2019). While some parties are more pessimistic about setting ethical guidelines for it (Floridi. 2019) (Munn. 2022), multiple different ones from private companies to researchers are calling for a set of ethical guidelines to guide the future development and usage of the technology (Jobin et. al. 2019). Obviously in order to set any kind of guidelines one has to first acknowledge the potential issues at hand.

1.2 The lack of transparency and research questions

In, (Khan et. al. 2021) the authors seek to determine the most pressing ethical issues of AI today using the systematic literature review. The most common issue they found was the lack of transparency. By this they mean the lack of understanding why AI provides us with the predictions that it does. This understanding is often lacking for all stakeholders and its lacking is caused by the popularity of so-called black box –models (Hassija et. al. 2024). Black box –model is a name given for AI models that are modelled after the brain to work as neural networks. These models are opaque and their "thinking" is hard to understand (Castlevecchi. 2016). Because of this opaqueness, predictions that they produce are often hard to explain and that leads into lack of trust. This leads us into my first RQ (Research question) (All abbreviations used in this paper can be found from Table 9.1 Abbreviations):

RQ 1: How would an increase in transparency help the field of AI ethically?

With my first RQ I sought to find ways in which an increase in transparency would improve the field of AI from an ethical point of view. I sought to justify them with literary sources and also take into the account the possible opposing viewpoints. If answered in a satisfying way the first RQ should give us examples of ethical issues currently present in AI that could be solved with more transparency.

My Second RQ was formed after reading couple of papers that were more sceptical about AI regulations (Floridi. 2019) (Munn. 2022). Their scepticism is rooted on the belief that organizations, companies and other stakeholders who wish to generate value with AI systems will seek to find any way to circumvent ethical guidelines and regulations in order to maximize profit. Their arguments clearly put ethics in an opposing role against economic progress. This made me

wonder whether an increase of transparency could in fact also have economic benefits. If that were the case, then economic and ethical issues could be tackled on at the same time. My RQ 2 became:

RQ 2: How would an increase in transparency help the field of AI economically?

With my second RQ I sought to answer whether an increase in transparency could also help AI from an economical viewpoint. I sought to find examples of economic gain that could be acquired with more transparent AI and justify them with literary sources. If answered in a satisfactory manner this RQ should help us understand whether such economic benefits exist or not.

My wish for this paper was that after answering the research questions we could have a better understanding of how an adoption of more transparent and explainable models would affect AI systems from ethical and economical viewpoints. With that we could better understand what kind of effect it would have on all stakeholders.

2. Background

2.1 Deep neural networks

Khan et. al. (2021) conducted a systematic literature review and explored various challenges of AI ethics. Number one issue with a slight margin is the lack of transparency. In a context of AI a lack of transparency means lack of information about why the algorithm arrives in a specific conclusion instead of a different one. This is caused by the way these machine-learning algorithms are being developed.

To put it simply most if not all advanced machine learning models are being developed using a three-part system. (Cao et. al. 2023) First the model goes through a phase called “pre-training”. In this phase the mostly random model is being trained with the dataset and is spitting out answers of varying quality. The answers then get scored based on accuracy and the algorithm gets fine-tuned based on the scores for the second phase which is called “reward-learning". After this the third phase called “fine-tuning with reinforcement learning” takes place. In this phase the algorithm gets fine-tuned one more time to maximize the learning generated from the dataset. Also, in the last phase possible unconventional answers that trick the reward-model are also being addressed if they pop up. (Cao et. al. 2023)

AIs created using this method are modelled after the human brain and are hence called deep neural networks (DNN) (Castlevecchi. 2016). This causes their “thinking” to be inherently hard for its creators to understand. It ends up on right conclusions because of sustained reinforcement learning, but it is impossible to say exactly why that is. This makes AIs predictions basically impossible to justify (Hassija & al. 2023).

2.2 The lack of trust

The inherent problem of “black box” AI has been around for much longer than the DNN models themselves. The earliest calls for explainability can be traced as far as in the seventies (Scott et. Al. 1977). In their paper “Explanation capabilities of production-based consultation systems” they adamantly argue for explainability in AI. Their reasoning is that an understandable AI is more likely to be accepted by experts on the field and also its mistakes are easier to spot and correct. Many researchers still support the same argument. For example, in (Hassija et. al. 2023) the authors claim that the lack of trust many modern AI systems suffer as consequence for their inability to explain themselves leads directly into issues with their adoption into real world decision making.

In order to give AI any kind of power over decision making it needs to be trustworthy. This does not only mean that we have a good reason to believe its prediction to be correct. In (Von Eschebach. 2021) the author argues that “Our relationship with technology often is one of reliance rather than trust”. They separate trust and reliability from each other by explaining that if we were to trust AI we would trust not only in the predictions that the AI forms, but also that the predictions

would be produced with the intent of helping us. This kind of trust is extremely hard to develop when the working of the AI is not clear even for those who have created it.

2.3 Explainable AI

While black box-models have advanced, and the field of AI has gone forward the worry for transparency has resulted into a new research field called XAI (Explainable AI). The term was invented by Van Lent et. al. (2004) in a paper in which they examine the potential use of AI that is programmed to be able to explain its own thinking in the context of an army training simulation "game" called Full Spectrum Command. The research field is simply focused on finding solutions that increase the transparency of the AI technology (Adadi et. Berrada. 2018). While the idea of increasing the credibility of AI made decisions by increasing the transparency and making the "thinking" process explainable is nothing new, the research field has changed a lot with the modern advances in AI (Xu et. al. 2019). With the introduction of DNNs unexplainable AI models are now everywhere. This has raised interest in XAI as a whole.

According to (Hassija & al. 2023) a lot of existing research done in the field of XAI seeks to answer three specific questions. Those are what is the definition of explainable AI, why is explainable AI important and how can we increase explainability in AI. These questions paint a good picture about XAI as a research field and what kind of studies it can produce. Out of these questions this study mostly seeks to answer why explainable AI is important and what it could provide for its stakeholders.

3. Research method

3.1 Choosing the research method

I chose to conduct a literature survey (LS). The method was chosen because it fit the scope of this paper and I felt that it could be used to answer my RQs in a satisfying manner.

3.2 Conducting the literature survey

Before starting my search, I started my survey by going through the references of (Khan et. al 2021). After that I conducted most of my search for sources by using Google Scholar. It became my tool of choice because of its cross-platform search capabilities. I also used my university account to access platforms such as IEE XPLORE when I found an article that had restricted access. I used multiple search strings such as “Transparency + AI + Ethics”, “AI + Ethics” and “AI + Black-box". I did not settle on a singular search string at any one point but modified it by including and excluding certain words to maximize the number of potential articles I found. I also went through the sources of the articles I found to find more potential references. The papers I ended up choosing were published in a variety of publications. These publications include most notably multiple papers from sources published by IEEE and SpringerLink, some from ACM and couple of conference papers.

3.3 The Criteria for Inclusion and Exclusion

I developed criteria for inclusion and exclusion of sources to refine my search. My criteria for inclusion were:

- Paper provides information about the RQs specifically.
- Paper is peer-reviewed and is fully available.
- Paper is written in English language.

My criteria for exclusion were:

- Paper provides no new information in the context of my RQs.
- Paper is grey literature.

3.4 The Final List of References

After conducting the literature survey, I had enough material to answer my RQs in what I felt to be a satisfying manner. I still had to seek and find some references to be able to write about the more technical aspects of ML systems and the history of XAI (Explainable AI) as a research field. In order to write about the economic aspects well enough I also had to include one source from the International Journal of Online Marketing (Haque et. al. 2020). In this way my complete list of references could be split in two groups. Group one would be references that are answering the research questions and group two are references that helped me provide enough context. Group one was collected through a survey process which took place before and during the writing process and the second one was expanded whenever more specific information was needed for context.

4. Results

4.1 RQ 1: How would an increase in transparency help the field of AI ethically?

“Trust is crucial for effective human interaction with machine learning systems, and that explaining individual predictions is important in assessing trust.” (Ribeiro et. al. 2016) In other words, when AI algorithms make more and more decisions that affect the lives of people, the explainability or lack of it becomes a major point in either making or breaking the trust people have for the AI. In (Milossi et. al. 2021) the authors write about the “GDPR approach to AI ethics”. GDPR (General Data Protection Regulation) is a European Union Regulation that tackles data privacy. While it doesn’t mention AI by name, it highlights the importance of transparency for the end user concerning how their data is being used. In (Milossi et. Al. 2021) the authors argue that GDPR could also be used to demand transparency from AI systems because of the way they use data and that such demands could improve AI ethics.

An AI itself cannot be held responsible for its decisions, but if its working is transparent enough the elements and/or individuals that caused certain outcome should be recognizable (Dastani et. Yazdanpanah. 2022). In this way a more transparent AI would again benefit all stakeholders by assigning blame more accurately thus being more trustworthy for all parties involved.

It is not a secret that image-generating AI systems often perpetuate offensive and harmful stereotypes of people (Bianchi et. al. 2023). This is caused by bias that is an inherent part of most ML systems. The reason for that is that datasets that are used to train AIs always include at least some form of historical bias (Hellström et. al. 2020). Historical bias in data means that while collected properly, the data still has some sort of unwanted bias because of the source of the data. This problem is obviously not exclusive to image-generation algorithms but exists in all DNN models. Transforming the black box into a glass box would not get rid of this bias, but it would allow us as users to peek inside the AIs decision-making process and better judge when a bias has wrongly altered a prediction and override its decision.

4.2 RQ 2: How would an increase in transparency help the field of AI economically?

There is a phenomenon in AI that the more difficult the AI is to explain the more accurate its predictions become (Xu et. al. 2019). This should in theory make more explainable algorithms less useful in business decisions. In the same paper Xu and the co-authors also argue that in reality in some situations the black box models could not be used at all and for that reason transparent models could provide more value. For example, they mention GDPR which grants the user the right of explanation that simply cannot be produced if the AI being used is an unexplainable model. This causes the AIs prediction to be unusable in some contexts, no matter how accurate or helpful it would prove to be.

Even when not being pressured by legal guidelines, being able to explain the decision-making processes to customers would certainly build trust between the customer base and the organization offering a service. This would cause economic benefits, because for example in the context of online shopping customer trust leads to customer loyalty (Haque et. Mazumder. 2020). Overall being able to offer transparency for the customers who seek it would definently lead to a more positive customer experience.

In their paper (Rai. 2020) Rai argues that transparency of the AI model provides value for all stakeholders including the developers and the owners of the system. They argue that in the long run transparent algorithms will be easier to benefit from because of their transparency: “Examining XAI utility holistically from the perspective of different stakeholders will provide a nuanced understanding about how to leverage XAI through the development and deployment lifecycle of a marketing AI application.” They argue that understanding the algorithm and its decision-making process will make it easier for the developers and other stakeholders to realize its potential and how to modify the system to achieve it.

5. Discussion

While some experts believe that ethical principles work so poorly that considering them is technically useless (Munn. 2022), I would argue that AI would gain both commercial and ethical benefits from addressing some of them. Especially the lack of transparency in the technology. An AI that lacks transparency is often called to be a “black-box”-algorithm. It is capable of making very accurate predictions, but it lacks the transparency and the ability to justify those predictions by showing its “thinking” process (Hassija et. al. 2023). They do not have to be the default, however. The alternative is called a “white box” or a “transparent” model and is seen as a solution to AIs transparency issue by some of the experts of the field (Rudin. 2019). This alternative is being researched in a research field called XAI.

In my findings I came to the conclusion that the increasing of transparency of AI systems would in fact help the field of AI advance from an ethical point of view. Transparency builds trust and from an ethical perspective trust is an important factor in AI systems (Von Eschenbach 2021). They argue that while AI systems can be relied on, reliance is just one aspect of trust. In order to fully trust these systems, we should be able to not only rely on them but be able to believe that they are working in our favor and produce predictions to benefit us. Transparency could also benefit us by making the system's errors easier to investigate and this way makes it easier to place blame on the correct party when something goes wrong (Dastani et. Yazdanpanah. 2022). Responsibility is also a major component in building trust in the systems and this kind of trust can be gained with transparency.

Transparency would also benefit the systems by combatting its biases. All datasets are biased in one way or another and while the systems inner workings are shrouded in a black box – model, these biases are harder to detect and correct (Hellström et. al. 2020). If a system could explain why, it classifies one thing as one and the other as the other it would be infinitely easier for the developer of the said system to correct it when needed. Although transparency would help us with problems of bias, it is also good to remember that in order to solve them we need a diverse group of people who are educated on the issue and willing to give their input in developing reliable trustworthy AI systems (Hassija et. al. 2024). While transparency is a good place to begin and one that needs to be improved, the ethical progress of AI does not end with just XAI.

It is clear that transparency and XAI can also provide economic benefits (Adadi et. Berrada 2018). This is partially because of guidelines such as GDPR which prohibit the use of unexplainable black box –models in certain contexts and partially because if AI systems are used to generate value, their predictions need to be trustworthy. This trustworthiness can be built at least partially by more transparent AI systems.

Some researchers also argue that a more transparent algorithm would in the long run also lead to economic benefits in the way of systems that would be easier to develop and benefit from (Rai 2020). They argue that AI systems, if more transparent, would reveal new opportunities to generate value for all stakeholders. From a purely economic standpoint, however, a lot of research in XAI is missing the calculations for the price versus reward for adopting a more transparent model. While some studies in XAI include a surprisingly precise framework or a guideline on how a more transparent system could be achieved they fail to calculate a potential cost for such an

operation. While some experts (Rai. 2020) are hopeful about the economic benefits in the form of new innovations and some (Adadi et. Berrada 2018) in the form of a more trustworthy algorithm the size of the economic benefit is hard to estimate without an estimate for the cost.

6. Threats to validity

6.1 The validity of the research method

The research method chosen for this paper was a literature survey and the search for sources in literature was conducted without a formal planned-out process. For this reason, it is possible that some articles that would have fit this paper and maybe even changed its conclusions could have been missed. This could make the findings of this study potentially questionable and hurt their generalizability. This questionability could be avoided or at least lessened for the future study of this topic by conducting a more rigorous, systematic and better planned out systematic literature review.

6.2 The validity of the author's bias

The sources picked for this study were chosen by me alone and were not reviewed by other parties. This paper is also the first time I have conducted a study of this kind and were I more experienced I would have probably conducted my search in a different way. For these reasons, we cannot for certainty assume that sources chosen don’t reflect the bias of the author. This would be a valid reason to question the findings of this study. In the future the bias of the author could be seriously lessened by including a group of peers whose work would be to also check the sources and confirm that they fit the chosen inclusion criteria. This group should be as diverse as possible in order to conduct as objective a study as possible.

7. Conclusions and future research

The lack of transparency in AI is widely recognized as a huge ethical issue (Khan et. Al. 2021). In this study I my aim was to find out how tackling this issue might affect the field not just ethically, but also from economic standpoint. My first RQ considered how an increase of transparency could affect the field of AI ethically and it reveals that while increasing the explainability of the models might not fix everything by itself it would have potentially massive ethical benefits in domains such as allocating the responsibility and detecting the bias of the system. While the improvements that an explainable AI would gain in the economical domain are more theoretical, many experts still believe that an increase of transparency would also at least have potential to generate economic value.

A shift to prioritize transparency would likely be a step towards a better future in both economic and ethical points of view. The possibilities of such future existing is being researched in a research field called XAI (Xu et. al. 2019). According to the experts of the field the adoption of more transparent models could in theory provide both ethic and economic benefits to all stakeholders. It is important to remember that a lot of the research in XAI is still very abstract and hypothetical and that the increase of transparency is not a fix-it-all solution that can alone fix the pressing ethical concerns of AI while simultaneously providing economic growth. Nevertheless, a move towards more transparency in AI would definently be a move forward.

I would argue that there are multiple topics within the ethics of XAI in which more research is needed. A lot more empirical studies about the adoption of glass box -models are needed so that we can have concrete examples of the effects that they have. Other more specific future research points are for example correcting the bias in XAI and finding out what kind of economic costs the owners of the systems would have to pay for adopting more transparent systems. While increasing the transparency is a right way to go it doesn't fix the issue of AI ethics by itself. Transparency builds trust (Ribeiro et. al. 2016), and guardrails are needed for that trust not to be abused. The future research in AI ethics needs to account for the fact that if ethical guidelines become too loose, they will end up as nothing, but a marketing strategy. (Hagendroff. 2020) Simply put while more transparency is strictly better, it should not be heralded as a quick fix for everything, for treating anything that way usually leads into bigger problems. (Floridi. 2019)

8. References

Adadi, A., Berrada, M. (2018) Peeking Inside the Black –Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, vol 6. pp. 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A. (2023) Easily Accessible Text-to-Image Generation Aplifies Demographic Stereotypes at Large Scale. FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 1493-1504. https://doi.org/10.1145/3593013.3594095

Borenstein, J., Grodzinsky, F. S., Howard, A., Miller, K. W., Wolf, M. J. (2021) AI Ethics: A Long History and a Recent Burst of Attention. https://doi.org/10.1109/MC.2020.3034950

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., Sun, L. (2023) A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. ACM 37, 4, Article 111. https://doi.org/10.48550/arXiv.2303.04226

Castlevecchi, D. (2016) Can we open the black box of AI? Nature 538. 20-23. https://doi.org/10.1038/538020a

Dastani, M., Yazdanpanah, V. (2022) Responsibility of AI Systems. AI & Society 38. P. 843-852. https://doi.org/10.1007/s00146-022-01481-4

Floridi, L. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. (2019) Philos. Technol. 32, 185–193 https://doi.org/10.1007/s13347-019-00354-x

Hagendorff, T. (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. Minds & Machines 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hassija, V., Chamola, V., Mahapatra, A. et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn Comput 16, 45–74 (2024). https://doi.org/10.1007/s12559-023-10179-8

Haque, U. N., Mazumder, R. (2020) A Study on the Relationship Between Customer Loyalty and Customer Trust in Online Shopping. International Journal of Online Marketing (IJOM) 10(2). https://doi.org/10.4018/IJOM.2020040101

Hellström, T., Dignum, V., Bensch, S. (2020) Bias in Machine Learning – What is it Good for? ArXiv:2004.00686v2. https://doi.org/10.48550/arXiv.2004.00686

Jobin, A., Ienca, M., Vayena, E. (2019) Artificial Intelligence: the global landscape of ethics guidelines. Nature Machine Intelligence. 389-399. https://doi.org/10.1038/s42256-019-0088-2

Khan, A. A., Badshah, S., Liang, P, Khan, B., Waseem, M., Niazi, M., Akbar, M. A. (2021) Ethics of AI: A systematic literature review of principles and challenges. EASE '22: Proceedings of the

26th International Conference on Evaluation and Assessment in Software Engineering. 383-392. https://arxiv.org/pdf/2109.07906.pdf

Milossi, M., Alexandropoulou-Egyptiadou, E., Psannis, K. E. (2021) AI Ethics: Algorithmic Determinism or Self-Determination? The GDPR Approach. IEEE Access, vol 9, pp. 58455-58466. https://doi.org/10.1109/ACCESS.2021.3072782

Munn, L. (2022). The uselessness of AI ethics. AI and Ethics 3, 869-877. https://doi.org/10.1007/s43681-0220-00209-w

Rai, A. (2020) Explainable AI: from black box to glass box. J. of the Acad. Mark. Sci. 48, 137–141. https://doi.org/10.1007/s11747-019-00710-5

Ribeiro, M. T., Singh, S., Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1135-1144. https://doi.org/10.1145/2939672.2939778

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. https://doi.org/10.1038/s42256-019-0048-x

Scott, A. C., Clancey, W. J., Davis, R., Shortlife, E. H. (1977) Explanation Capabilities of production-based consultation systems. American Journal of Computational Linguistics 62. P. 9

Van Lent, M., Fisher, W., Mancuso, M. (2004) An Explainable Artificial Intelligence System for Small-unit Tactical Behavior. Proc. 16th Conf. Innov. Appl. Artif. Intell., pp. 900-907

Von Eschenbach, W. J. (2021) Transparency and the Black Box Problem: Why We Do Not Trust AI. Philosophy & Technology 34. 1607-1622 https://doi.org/10.1007/s13347-021-00477-0

Xu F., Uszkoreit H., Du Y., Fan W., Zhao D., Zhu J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. https://doi.org/10.1007/978-3-030-32236-6_51

9.Tables

9.1 Abbreviations

AI	Artificial intelligence
DNN	Deep neural network
GDPR	General data protection regulation
LS	Literature survey
ML	Machine learning
RQ	Research question
XAI	Explainable AI