An Ethics Impact Assessment (EIA) for AI uses in Health & Care

The correlation of ethics and legal aspects when AI systems are used in health & care contexts

Elsa, EP, Papadopoulou

Department of Informatics, Ionian University, Corfu, Greece, elsapapa30@yahoo.fr

Themis, TE, Exarchos

Department of Informatics, Ionian University, Corfu, Greece, themis.exarchos@gmail.com

The field of automated-decision making systems using AI technologies has evolved rapidly over the past years. Alongside the technological evolution, the debate surrounding the principles from an ethics point of view that these systems should be based upon, has produced a prolific number of analyses and discussions. In the health & care context, AI systems serve a wide range of activities, from prevention and diagnosis to treatment and monitoring and a variety of users, be it health & care experts or personnel as well as healthy individuals or those requiring medical assistance or treatment. Some of the categories of uses and users, the opportunities and the challenges from a legal and ethics point of view, linked to each category, are discussed in this paper. A transdisciplinary approach is of primordial importance. Such approach should comprise stakeholders and elements from the technology design and development area, the legal and ethics fields and importantly, the users of such AI systems in the health & care context. The paper attempts to demonstrate the various notions that are entangled in the conception, design, development, deployment and use of such systems and introduce the idea of an 'Ethics Impact Assessment' ('EIA') for a more robust operationalization in a health & care context.

CCS CONCEPTS • Artificial Intelligence

Additional Keywords and Phrases: Artificial Intelligence, Ethics-by-Design, Ethics Impact Assessment, Health & Care, Liability

ACM Reference Format:

Elsa Papadopoulou. Themis Exarchos, 2022. An Ethics Impact Assessment (EIA) for AI uses in Health & Care: The correlation of ethics and legal aspects when AI systems are used in health & care contexts. In Proceedings of SETN conference (SETN 2022). Corfu, Greece, 4 pages.

https://doi.org/10.1145/1234567890

1 INTRODUCTION

The analysis of this paper presents a unique approach as it combines findings and principles drawn from theories, practices and methodologies such as design thinking, computational ethics, value sensitive design, game theory, while integrating aspects pertinent to the explainability, trustworthiness and accountability of AI systems used in health & care contexts. The underlying notion of the analysis is to examine whether and how such notions could be 'merged' with more 'strictu sensu' legal aspects such as 'liability' and 'accountability', so as to deliver a more 'real-life' applicable approach when designing and/or auditing such systems in view of rendering them more robust. The authors of this paper based the analysis on research of previous papers in the field

2 DESIGN THINKING, VALUE SENSITIVE DESIGN (VSD), COMPUTATION ETHICS AND 'ETHICS-BY-DESIGN' OF AUTOMATED DECISION-MAKING SYSTEMS USING AI TECHNOLOGIES IN A HEALTH & CARE CONTEXT

Various analyses and discussions to-date, cover the entire lifecycle of an AI system: from its conception and design to its deployment and use. Thus, notions such as 'ethics-by-design' and 'auditing' of AI systems have been introduced to enable the uptake of AI systems and help optimise the outcomes when such systems are used to support users in a human-centric manner.

In 1980 writing, Marvin Minski had stated: "Nevertheless, the problem of meandering is certain to re-emerge once we learn how to make machines that examine themselves to formulate their own new problems. Questioning one's own "top-level" goals always reveals the paradox-oscillation of ultimate purpose. How could one decide that a goal is worthwhile -- unless one already knew what it is that is worthwhile?". In the same writing, Minsky discusses also common sense reasoning. (Minsky, 1980).

The discussion shall endeavor to provide an overview of the correlation and causality between various notions including those of 'ethics-by-design' and 'auditing' while keeping as a backbone of the analysis the following query: what is the purpose of use of AI systems in a health & care context and whether this purpose of use is worthwhile. The analysis shall also attempt to elaborate on whether and how common sense reasoning for AI system uses in a health & care context could be necessary, plausible and feasible.

In addition to the two aforementioned notions, the analysis shall attempt to introduce a correlation between the aforementioned query to the notions of Design Thinking (DT), Value Sensitive Design (VSD), Applied Ethics, Explainability, Trustworthiness, Computational Ethics and even a 'translation' and applicability of the Hippocrates oath to ensure that the ultimate purpose of use, i.e. saving human lives, is upheld when AI systems are used in a health & care context. Examples of uses will be embedded in a future extended paper to illustrate the uses and challenges derived from such uses.

AI has been described to be a set of sciences, theories and techniques (including mathematical logic, statistics, probabilities, computational neurobiology, computer science) that aims to imitate the cognitive abilities of a human being, leading computers to perform increasingly complex tasks, which could previously only be delegated to a human. Depending on the degree of reasoning and autonomous decision-making, AI has been defined as 'strong', i.e. of an intelligence equal to humans and capable of handling a wide range of tasks rather than one particular task or problem, or 'weak' AI alias known as 'narrow' AI, which is used to describe artificial intelligence systems that are specified to handle a singular or limited task. (IBM, 2020) (Council of Europe, 2021).

Throughout 2019 and 2020, the High Level Expert Group (HLEG) on AI that was appointed by the European Commission to provide advice on its artificial intelligence strategy, produced a number of deliverables. One of these deliverables, were

the 'Ethics guidelines for trustworthy AI'. In the Guidelines, the HLEG explored the conditions that would need to be fulfilled for AI systems to be considered trustworthy and introduced a set of seven key requirements that these systems should meet in order for them to be deemed as trustworthy (EC HLEG on AI, Ethics guidelines for trustworthy AI, 2019). The Guidelines were accompanied by another document, which provided a 'Definition of Artificial Intelligence' as understood and used in the context of the Guidelines.

In this document, the HLEG drew a distinction between AI systems and AI as a scientific discipline. They referred to AI systems as any AI-based component, software and/or hardware, which are usually embedded as components of larger systems, rather than stand-alone systems. This reference was a simple abstract description of an AI system, through three main capabilities: perception, reasoning/decision making, and actuation, as analysed in the 'Definition of AI' document. Considering the three aforementioned capabilities of AI systems, the HLEG referred to AI as a scientific discipline, as the AI techniques and sub-disciplines that are currently used to build AI systems. They broadly categorised the techniques in two main groups based on the AI systems' capability of reasoning and learning.

Their analysis concluded with the below definition of AI systems:

"Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions (EC HLEG on AI, Ethics guidelines for trustworthy AI, 2019).

For the purposes of the discussion in this paper and any further extended version, the aforementioned definition provided by the HLEG of AI as systems designed by humans, is used.

With the aforementioned definition in mind, a transdisciplinary approach and collaboration is necessary combining knowledge and experience from Science, Technology, Engineering & Mathematics (STEM) and Social Science and Humanities (SSH) disciplines. Taking also into account the principle of the purpose of use of AI systems within a health & care context and the worthwhileness of such purpose of use, it is useful to discuss the other notions of Design Thinking (DT), Value Sensitive Design (VSD), Computational Ethics and 'Ethics-by-Design', their interference and correlation to a worthwhile purpose of use.

Design Thinking (DT) is based, among others, on the principle that empathy is embedded in the design phase with the users for whom the innovation is developed to understand their pains and problems fully. The latter, in turn, is converted to the **Human-Centred Design (HCD)** that focuses on understanding the perception, the needs, and expectations of the person who are looking for a solution to a specific problem and whether the proposed solution has been designed in a way to and will effectively and efficiently resolve the problem for which it was designed. **HCD** can be further enriched by **Value Sensitive Design (VSD)** principles, which is a method that embeds values into a technical design.

As proposed in an analysis published at the beginning of 2021 by Steven Umbrello & Ibo van de Poel (2021), VSD could be integrated into AI systems design to address the challenges posed by the need for transparency, explicability, and accountability of AI systems as well as those posed by Machine Learning (ML) which may lead to AI systems adapting in ways that "disembody" the values embedded in them.

As an example, a study discussed the moral precepts and how could these VSD principles be operationalized in the design of the Quality of Life (QoL), QoL-ME, which is an eHealth and mHealth application that is expected to address important human values in the tool's design, using VSD principles for intregrating important human values during the development of the tool. (Maathuis, Niezen, Buitenweg, Bongers, & Nieuwenhuizen, 2020).

Another notion that the analysis shall factor in, is that of Computational Ethics. In a research paper, a framework of **Computational Ethics** is proposed, which specifies how the ethical challenges of AI can be better addressed by

incorporating the study of how humans make moral decisions. (Edmond Awad et. al, 2022). An interesting aspect reflected in the same paper is the one that asks how human ethics can inform machine ethics, and vice versa.

Abiding to the **Hippocrates Oath** that physicians are called to take, they swear that they will uphold this according to the best of their ability and judgment and that they shall abstain from all intentional wrong-doing and harm.

The **Turing-Holberton Oath** (The Holberton-Turing Oath, 2022) was initiated by Aurélie Jean[1] and Grégory Renard[2] and takes its name from Frances Elizabeth "Betty" Holberton[3] and Alan Mathison Turing[4]. The oath "*is made to gather all AI experts in the world around shared values and moral to drive them to use their skills by insuring their integrity and by avoiding any threat to any life being.*"

Similar to the Hippocrates Oath, the Turing-Holberton Oath states, among others, that the members of the data science and artificial intelligence profession should exercise their profession with the utmost respect of human life; will not permit bias of any kind (cultural, ethic, sexual orientation, race, gender, etc.); recall that human concerns outweigh technological ones; that it is human beings and their lives (family, social, economic, personal freedom) that can be and are affected by these technologies; that the data and AI scientists shall practice their profession with conscience and dignity; promise to create AI, first, to collaborate with people for the greater good, rather than usurp the human role and supplant them.

Introducing another angle in the discussion, the **Trolley Problem and the application of Game Theory** in medical ethics can be added in the exploration of ethics-by-design of AI systems used in health & care contexts.

In the 1967 essay "The Problem of Abortion and the Doctrine of the Double Effect" (1967), English philosopher Philippa Foot, defined the doctrine of double effect in terms of the distinction between what a person strictly (directly, explicitly) intends as the end and the means of a contemplated action and what a person "obliquely" (indirectly) intends as a foreseen consequence of the action but not as an end or a means. The term "The Trolley Problem" was first introduced as such by the American philosopher Judith Jarvis Thomson in her 1976 essay "Killing, Letting Die, and the Trolley Problem". Foot had pointed out that "*the doctrine of double effect is vulnerable to counterexamples if it is formulated too broadly as the principle that actions that have foreseeable bad consequences are morally permissible as long as those consequences are not directly intended—i.e., as long as they are intended only obliquely*" (Duignan, 2022).

In a 2019 paper on the subject of medical ethics and the trolley problem, the notion is put forward that medical ethics relies on the balancing of four main principles: autonomy, beneficence, non-maleficence and justice. The paper goes on to explore the question if and which of these principles should prevail over the other basing the analysis on the concept of the Trolley Problem. Therein it is also stated **"*The moral value of an action is not in its intrinsic nature, but rather in its consequences*"** (G.Andrade, 2019).

Another paper (Riggs JE, 2004) discusses the applicability of Game Theory and logic traps in medical ethics. **Game Theory** "*is the study of the ways in which interacting choices of economic agents produce outcomes with respect to the preferences (or utilities) of those agents, where the outcomes in question might have been intended by none of the agents.*" (Ross Don, 2021). The author uses to paradigm of the **'Prisoner's Dilemma'** from Game Theory to discuss the possible logic traps that can occur during decision-making by medical practitioners and which can impact such decisions from an ethics point of view as well. The 'Prisoner's Dilemma' is a decision-making and game theory paradox illustrating that two rational individuals making decisions in their own self-interest cannot result in an optimal solution.

Drawing from the aforementioned it is important to take into consideration a wide spectrum of data in order to reach an optimal output/decision especially under ethically sensitive circumstances. It could be explored how a non zero-sum approach inspired by the Prisoner's Dilemma paradox, the notion of consequences rather than intentions, the principles of autonomy, beneficence, non-maleficence and justice can be integrated in the design and deployment of AI systems used in

[1] Chief AI Officer and co-founder of DpeeX, an AI-based precision medicine deep tech startup | https://www.linkedin.com/in/aureliejeanphd/
[2] Head of Applied AI at AAICO | https://www.linkedin.com/in/gregoryrenard/
[3] https://en.wikipedia.org/wiki/Betty_Holberton
[4] https://en.wikipedia.org/wiki/Alan_Turing

health & care contexts. It is useful also to research whether the notion of a 'Do Good' instead of not solely 'Do No Harm' for AI systems in a health & care context is more appropriate based on the reasoning that these systems are designed, deployed and used in order to augment human agent decision making, reduce error rates (diagnosis, treatment, monitoring, etc) as the ultimate purpose of use is worthwhile - i.e. maintaining, improving human health and even more, saving human lives - and the consequences of wrong choices or data input can be severe, i.e. the loss of human lives.

Furthermore the principles of **'Ethics-by-Design' (EbD)** can be examined and how these affect the overall development of a human-centric framework. For example, In a 2021 paper published by the European Commission (DAINOW & BREY, 2021) the following definition of EbD is provided: "*Ethics-by-Design is intended to prevent ethical issues from arising in the first place by addressing them during the development stage, rather than trying to fix them later in the process. This is achieved by proactively using the principles as system requirements. What is more, since many requirements cannot be achieved unless the system is constructed in particular ways, ethical requirements sometimes apply to development processes, rather than the AI system itself."* Furthermore, Aral Balkan and Laura Kalbag have developed an 'Ethical Design Manifesto' that encapsulates the aspects of human rights, human effort and human experience presented in the form of a pyramid in a hierarchical manner with human rights as the basis of the pyramid and human experience at the top of the pyramid (Kalbag & Balkan, 2022). The analysis in this paper shall try to reflect the principles of these and other examples of 'Ethics-by-Design' methodology and principles and how these can be embedded in an 'Ethics Impact Assessment' for the use of AI technologies in the health & care context.

One question that rises is how could AI systems help improve or rather augment medical practitioners'/personnel's decision making capacity as to which individual's life or health is requires more urgent medical intervention over another's; which is the best therapy to choose for a specific condition; which patient necessitates an e.g. heart transplant more urgently over another in almost 'ex-aequo' severe cases; which is the precise nature of the pathology reducing diagnosis error? How can algorithms 'transform' the data input into ethically and medically 'non-zero sum' outputs? These and other questions can be served by AI and assist relevant actors with their tasks in health & care contexts.

3 EXPLAINABILITY, TRUSTWORTHINESS, AND CORRELATION TO LEGAL ASPECTS VIA AN 'ETHICS IMPACT ASSESSMENT' (EIA) FOR AUTOMATED DECISION-MAKING SYSTEMS USING AI TECHNOLOGIES IN A HEALTH & CARE CONTEXT

The notion of explainability and trustworthiness of AI systems has been widely discussed and debated. Depending on the context and the consequences, e.g. what happens when things go wrong, AI systems need to be explainable and trustworthy.

As per an NIST analysis (Phillips et al., 2021) explainable AI systems need to demonstrate the following features:

- "Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.
- Meaningful: A system provides explanations that are understandable to the intended consumer(s).
- Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.
- Knowledge Limits: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output"

In terms of legal consequences, an example can be drawn from the proposal of a Regulation for a EU Artificial Intelligence Act (Proposal for a Regulation laying down harmonised rules on artificial intelligence, 2022) wherein it is stated: "Furthermore, the exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial as well as the right of defence and the presumption of innocence, could be hampered, in particular, where such AI systems are not sufficiently transparent, explainable and documented."

For an AI system to be considered trustworthy, the following principles need to be embedded:

- Explainability: understanding the decision making tree and what data where used and why and what data were not used and why
- Fairness: how is the system conceived and designed to mitigate bias and ensure equitable access and outcomes
- Robustness: how can the system ensure the capacity to mitigate adversarial attacks safeguarding e.g. the users' privacy
- Transparency: how can the system and the whole process from design to deployment and use facilitate access to information to a wider audience
- Privacy: how can AI systems ensure that the data used through the entire lifecycle from training to production and governance are protected against an privacy hazards.

Various analyses have demonstrated that the design, development, deployment and use of trustworthy AI sysetms require a multi- and transdisciplinary approach comprising stakeholders from the technology, public, regulatory, users', etc. domains. For health & care uses, standardardised processes covering the entire lifecycle of AI systems from conception to use with underlying ethics and equity principles backed by appropriate regulatory frameworks, are of primordial importance for trustworthy AI systems (Crigger, 2022). An AI system 'Doing Good' and not only 'Doing No Harm'could be an ideal objective. The question is how could such a 'Doing Good' only AI system be achieved and how would this correlate with aspects such as e.g. liability in case of error? In a 2019 article on AI uses in health & care, it was stated (Kocher & Emanuel, 2019): "Of course, improving diagnostic and therapeutic outcomes are laudable goals. But AI is only as good as the humans programming it and the system in which it operates. If we are not careful, AI could not make health care better, but instead unintentionally exacerbate many of the worst aspects of our current health care system." This statement reflects the close links between the purpose of use, its worthwhileness, the 'do good' aspect and the interference with the human agent, which could create complex legal liability issues.

4 CONCLUSION

Drawing from the theory and accompanying examples, the future extended paper shall explore how could all these aforementioned principles be integrated in the conceptualisation and development of "Ethics Impact Assessment" (EIA) framework for AI systems used in health & care contexts in order to enhance the human-centric features of such systems with the goal of improving the quality of life of both healthy individuals and those requiring medical assistance and even more of saving human lives.

Accountability of systems and systems designers should be a backbone to the concept, design, development, deployment and use of AI systems in a health & care context. The consequences of 'when something goes wrong' in such contexts can be severe. Systems should be designed in an accountable manner. Systems' designers should also be held accountable in case of errors following appropriate due diligence. Liability provisions need to be embedded when and where best applicable so as to ensure that the ultimate purpose of use, which is to improve or even save human lives, can be served to its utmost worthwhileness.

An 'Ethics Impact Assessment' can serve as an initial compass for the design of a framework that can embed provisions that frame in a tailored and fair manner the accountability and, respectively, the share of liability of the intervening stakeholders in the lifecycle process of AI uses in health & care.

ACKNOWLEDGMENTS

This research is funded by the European Union and Greece (Partnership Agreement for the Development Framework 2014–2020) under the Regional Operational Programme Ionian Islands 2014–2020, project title: NEUROSYSTEM: Decision

Support System for the analysis of multilevel data of non-genetic neurodegenerative diseases", project number: MIS 5016116.

REFERENCES

[1] Council of Europe, C. o. (2021). *History of Artificial Intelligence*. Retrieved from What's AI: https://www.coe.int/en/web/artificial-intelligence/history-of-ai

[2] DAINOW, B., & BREY, P. (2021). *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence.* Brussels: European Commission.

[3] Duignan, B. (2022, April 20). *Philosophy & Religion*. Retrieved from Britannica: https://www.britannica.com/topic/trolley-problem

[4] EC HLEG on AI. (2019, April 2019). *Ethics guidelines for trustworthy AI*. Retrieved from Shaping Europe's digital future: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[5] EC HLEG on AI. (2019, April 8). *Ethics guidelines for trustworthy AI.* Retrieved from Shaping Europe's digital future: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651

[6] Edmond Awad et. al. (2022). Computational Ethics. *Trends in Cognitive Sciences*, 388-405.

[7] G.Andrade. (2019). Medical ethics and the trolley Problem. *J Med Ethics Hist Med.*, 12:3.

[8] IBM. (2020, August 31). *Strong AI*. Retrieved from IBM Cloud Learn Hub : https://www.ibm.com/cloud/learn/strong-ai#toc-strong-ai--YaLcx8oG

[9] Kalbag, L., & Balkan, A. (2022, April 17). *Ethical Design Manifesto*. Retrieved from Ind.ie: https://ind.ie/about/manifesto/

[10] Maathuis, I., Niezen, M., Buitenweg, D., Bongers, I. L., & Nieuwenhuizen, C. v. (2020). Exploring Human Values in the Design of a Web-Based QoL-Instrument for People with Mental Health Problems: A Value Sensitive Design Approach. *Science and Engineering Ethics*, 871-898.

[11] Riggs JE. (2004). Medical ethics, logic traps, and game theory: an illustrative tale of brain death. *Journal of Medical Ethics* , 30:359-361.

[12] Ross Don. (2021, Fall). *Game Theory*. Retrieved from The Stanford Encyclopedia of Philosophy (Fall 2021 Edition): https://plato.stanford.edu/archives/fall2021/entries/game-theory/

[13] *The Holberton-Turing Oath*. (2022, April 18). Retrieved from https://www.holbertonturingoath.org/