Medical Artificial Intelligence and Related Ethics Issues

Inaugural dissertation

to

be awarded the degree of Dr. sc. med.
presented at
the Faculty of Medicine
of the University of Basel

by

Giorgia Lorenzini

From Locarno, Ticino

Basel, 2023

Approved by the Faculty of Medicine

On application of

Prof Bernice Elger, first supervisor

Dr David Shaw, second supervisor

Prof Sigrid Sterckx, external expert

(Members of PhD committee)

Basel,

(Date of the acceptance of the Faculty)

……..

Dean

Prof Dr Primo Leo Schär

Table of Contents

Acknowledgments

I would like to thank all the people who helped and supported me during my PhD. First of all, my supervisors: Prof Bernice Elger and Dr David Shaw. Thank you for being always so full of ideas and inspiration, for your understanding and your immense availability, and for sharing your valuable expertise and experience with me. I would like to further thank Prof Bernice Elger for designing the EXPLaiN project and giving me the chance to work on such an important and interesting topic. I also owe many thanks to the Swiss National Science Foundation for funding the EXPLaiN project within the National research programme "Digital Transformations" (NRP77).

I was not alone in this project: I would like to thank Laura Arbelaez Ossa who embarked on this adventure with me. She has been an amazing colleague and friend who supported me both emotionally and academically. I am grateful for having worked with such an inspiring person and I wish her all the best for her future.

I would also like to express my gratitude to the other IBMB colleagues. Both the ones who were already here (or who left in the meantime) and the ones who joined later. They have always given me a word of advice and supported me even in the most difficult times. Particularly, I would like to mention Dr Tenzin Wangmo and Dr Michael Rost for sharing with me their impeccable methodology knowledge. Their help has been crucial when conducting empirical research. I am also grateful to (very soon to be Dr) Nadine Felber and Angelina Tian, who very kindly always listened to my problems and ideas. I am thankful for your friendship throughout my PhD. One last colleague I would like to mention is Anne-Christine Loschnigg, who has always been very patient with me and has been essential in helping me sort out all the administrative stuff. To them and all my other colleagues, I am deeply appreciative.

Of course, I want to thank my family, who always made sure I could follow my dreams. Thank you to mom, Ermita, and dad, Damiano, for working hard to provide me with all the opportunities I could ever imagine. I will always hold dear to your dedication, resilience, and love. I am grateful also to my grandmother Ester who selflessly welcomed me into her house for the main part of my studies. And when she was far away, she always made sure that I was happy and well. Lastly, but not less important, thank you to my sister Gaia who never stopped believing in me and encouraged me to pursue my aspirations, however crazy they might have seemed. I love you all so much and I could not have asked for better support.

A special thank you also to Robin, who was there during my Bachelor's and Master's defence and who has been my number one supporter during my PhD. Thank you for your unconditional love and understanding, for which I will forever be grateful.

Finally, I am grateful to my friends for always reminding me that I can achieve everything I set my mind to. Of course, they also reminded me that sometimes I need a break, and always made these breaks fun. Most of all, I am thankful to my best friend Flavia, with whom I shared many experiences and apartments. Thank you for bearing with my overthinking without ever judging me. I am also glad for Chiara's friendship, she helped me stay sane through it all: you are going to be a great psychologist! I am grateful for both the new friends I met in Basel and my longstanding friends who supported me from back home: your friendship has been priceless for me, thank you.

Summary

Chapter 1 introduces the topics and concepts discussed in the following Chapters of this thesis. In particular, it presents medical artificial intelligence (MAI) and it argues that it bears the potential to affect the doctor-patient relationship, both doctors' and patients' autonomy, patients' safety, cybersecurity, and shared decision-making (SDM). It concludes by highlighting the importance of a reflection on how we talk about MAI since our narratives have a performative power and can therefore influence its uptake and development.

Chapter 2 gives an overview of the methodologies used for the present research. It also aims to explain how theoretical and empirical research can be combined in bioethics while showing the benefits of this approach. Eventually, this thesis is based on what is known as empirical bioethics, although theoretical work is prevalent.

Chapter 3 analyses the role of AI-based clinical decision support systems (CDSS) for shared decision-making (SDM) to better comprehend its promise and associated ethical issues. Artificial intelligence (AI) based CDSS are becoming ever more widespread in healthcare and could play an important role in diagnostic and treatment processes. For this reason, AI-based CDSS has an impact on the doctor-patient relationship, shaping their decisions with its suggestions. We may be on the verge of a paradigm shift, where the doctor-patient relationship is no longer a dual relationship, but a triad. Moreover, Chapter 3 investigates how certain AI implementations may instead foster the inappropriate paradigm of paternalism. Understanding how AI relates to doctors and influences doctor-patient communication is essential to promoting more ethical medical practice. Both doctors' and patients' autonomy need to be considered in the light of AI.

Informed consent is at the core of the clinical relationship. With the introduction of machine learning (ML) in healthcare, the role of informed consent is challenged. Chapter 4 addresses the issue of whether patients must be informed about medical ML applications and asked for consent. It aims to expose the discrepancy between ethical and practical considerations while arguing that this polarization is a false dichotomy: in reality, ethics is applied to specific contexts and situations. Bridging this gap and considering the whole picture is essential for advancing the debate. In light of the possible future developments of the situation and the technologies, as well as the benefits that informed consent for ML can bring to shared decision-making, Chapter 4 concludes that it is necessary to prepare the ground for a future requirement of informed consent for medical ML.

Healthcare cybersecurity is increasingly targeted by malicious hackers. This sector has many vulnerabilities and health data is very sensitive and valuable. Consequently, any damage caused by malicious intrusions is particularly alarming. The consequences of these attacks can be enormous and endanger patient care. Amongst the already-implemented cybersecurity measures and the ones that need

to be further improved, Chapter 5 aims to demonstrate how penetration tests can greatly benefit healthcare cybersecurity. It is already proven that this approach has enforced cybersecurity in other sectors. However, it is not popular in healthcare since many prejudices still surround the hacking practice and there is a lack of education on hackers' categories and their ethics. Chapter 5 analyses hacker ethics to comprehend who ethical hackers are. Currently, hacker ethics has the status of personal ethics; however, to employ penetration testers in healthcare, it is recommended to draft an official code of ethics, comprising principles, standards, expectations, and best practices. Additionally, it is important to distinguish between malicious hackers and ethical hackers. Amongst the latter, penetration testers are only a sub-category. Acknowledging the subtle differences between ethical hackers and penetration testers allows to better understand why and how the latter can offer their services to healthcare facilities.

The discourse surrounding medical artificial intelligence (AI) often focuses on narratives that either hype the technology's potential or predict dystopian futures. AI narratives have a significant influence on the direction of research, funding, and public opinion and thus shape the future of medicine. Chapter 6 aims to offer critical reflections on AI narratives, with a specific focus on medical AI. This Chapter raises awareness as to how people working with medical AI talk about AI and discharge their 'narrative responsibility'. Qualitative semi-structured interviews were conducted with participants from different disciplines who were exposed to medical AI. The research represents a secondary analysis of data using a thematic narrative approach. Stories about the AI-doctor interaction depicted either a competitive or collaborative relationship. Some participants argued that AI might replace doctors as it performs better than physicians. However, others believed that doctors should not be replaced and that AI should rather assist and support physicians. The idea of excessive technological deferral and automation bias was discussed, highlighting the risk of 'losing' decisional power. The possibility that AI could relieve doctors from burnout and allow them to spend more time with patients was also considered. Finally, a few participants reported an extremely optimistic account of medical AI while the majority criticized this type of story. The latter lamented the existence of a 'magical theory' of medical AI, identified with techno-solutionist positions. The majority of the participants reported a nuanced view of technology, recognizing both its benefits and challenges, and avoiding polarized narratives. However, some participants did contribute to the hype surrounding medical AI, comparing it to human capabilities and depicting it as superior. Overall, the majority agreed that medical AI should assist rather than replace clinicians. Chapter 6 concludes that a balanced narrative (that focuses on the technology's present capabilities and limitations) is necessary to fully realize the potential of medical AI while avoiding unrealistic expectations and hype.

Finally, Chapter 7 provides a summary of the recommendations and conclusions presented in the previous chapters. It indicates further research directions and limitations while concisely recommending the next steps for ethically implementing MAI. The final considerations centre on patients' safety and rights, which should always be prioritised.

List of Abbreviations

AI	Artificial intelligence
BE	Bioethics
CDSS	Clinical decision support systems
CS	Computer science
DL	Deep learning
EC	Economy
EHR	Electronic health records
EKNZ	Ethics Committee of the Northwest and Central Switzerland
EXPLaiN	Ethical and legal issues of mobile health data – Improving the understanding and explainability of digital transformation and data technologies using artificial intelligence
LA	Laura Arbelaez Ossa
LW	Law
MAI	Medical artificial intelligence
ME	Medicine
ML	Machine learning
NRP77	National research project 77
GDPR	European General Data Protection Regulation
GL	Giorgia Lorenzini
Pen-tests	Penetration tests
Pen-testers	Penetration testers
Pen-testing	Penetration testing
PH	Public health
PL	Philosophy

PS	Psychology
SDM	Shared decision making
SNF	Swiss National Science Foundation
US	United States of America
UVM	University of Vermont

Chapter 1

Introduction

1.1.Introduction

Technologies based on artificial intelligence (AI) techniques are ever more popular. Debates about AI possibilities and challenges are never-ending. This is partly because it is difficult to reach a conclusion at the moment due to the novelty of their application, but also because the pace at which innovations happen is notoriously fast(1). For example, two AI tools that have drawn much attention lately are ChatGPT and OpenArt. Both are considered capable of closely imitating unique human abilities: converse and create art. OpenArt is an image-generating platform that aims to allow everyone to create artistic work. ChatGPT is a natural language processing tool that is said to have human-like conversations. It is often used to draft the most varied texts: from emails to school assignments and job applications. It alarmed particularly educational institutions, who feared that students might delegate their research and writing processes to this tool, thus negatively impacting their learning(2). Unlike other forms of 'cheating', it appeared impossible to discern students' work from ChatGPT-generated text. New tools were developed to tackle this issue and guarantee schools and universities the possibility to catch cheaters(3). At the same time, ChatGPT's limitations became apparent: it was not an all-powerful tool(4). It often reported inaccurate or wrong information, sometimes even inventing it. Human supervision and revision were still pretty much needed.

OpenArt AI was similarly capable of questioning humans' role in a field that is considered uniquely and profoundly human: art and creative expression. An artist submitted an AI-generated image to a major photography competition and won, unleashing a panoply of debates on the future of art and artists(5). Indeed, lately many people working in the creative industry expressed concern about AI tools that would both replace and steal their work(6). This called for attention to the deployment of this kind of tool while requiring further reflection on the role and meaning of art for humanity.

Although far from the healthcare context, these two examples illustrate concerns that are similarly present in the discussions about medical artificial intelligence (MAI). The job displacement concern does not spare any profession. Moreover, improper use or unreliable AI based-tools are possible and worrisome in healthcare as in other fields. Before diving into the ethics of introducing MAI in clinical practice it is fundamental to better understand what it is, how it works, and what it can do.

The goal of this thesis is to identify, analyse, and address some of the ethical issues surrounding the implementation and usage of MAI in healthcare. Particularly, the focus is on how MAI could impact and alter doctors' daily practice and professional autonomy, patients' safety and autonomy, and the doctor-patient relationship. For this reason, it will be considered how MAI relates to the shared-decision making (SDM) paradigm, whether it calls for an informed consent requirement, and how these new interconnected technologies could constitute a danger to patients' privacy and safety while promising great care improvements. Finally, it is also important to understand which narratives are prominent about MAI as these narratives shape its development and uptake, as well as the future of medicine.

Understanding the influence of MAI on doctors' clinical practice and professional autonomy is essential to imagine how the future doctor-MAI collaboration will look like. At the same time, on the basis of this collaboration, it is possible to conceive MAI's impact on patients and the new clinical interaction modes.

1.2.Medical Artificial Intelligence

AI in medicine is poised to play an increasingly prominent role as the computing power is advancing, algorithms are becoming more precise, and health data volume is ever-growing (7). With the advent of machine learning (ML) and deep learning techniques (DL), MAI research and development evolved dramatically(8). This began in the year 2000, particularly after the IBM question-answering computer system won first place on the television game show *Jeopardy!* and it gained a lot of popularity, fuelling hopes that AI could provide good answers also to doctors' questions. By combining new AI techniques with electronic health records (EHR), one could process large amounts of health data that would otherwise be very difficult to interpret and put to good use(9). The added benefit of ML is that it can learn from the data it is fed and become more accurate, improving its performance(10). DL is a subset of ML techniques that require less human intervention and supervision, relying on artificial neural network structures that mimic the functioning of the human brain(11). The advances in AI techniques allowed for unstructured and larger amounts of data to be processed faster, without demanding a lot of human labour. Therefore, the possibilities to implement AI-based tools in healthcare increased and many hopes were placed on the further advancement of MAI.

MAI applications can be very diverse: they range from administrative tasks (e.g. record keeping, data entry, and appointment scheduling) to supporting doctors' decision-making and diagnosis process. The latter are usually known as clinical decision support systems (CDSS) and can suggest diagnoses, make predictions, and recommend treatments(12). CDSS promise faster and more accurate decisions and diagnosis, aiming to improve the overall quality of care while reducing errors(13). The hope is to devote more time to the doctor-patient relationship, empower patients and enable them to manage their own care, reduce costs, and promote a more preventive and personalized healthcare model; apparently, all aspects of healthcare could be ameliorated with AI(14).

However, these promises are not always met with enthusiasm: some criticize the unrealistic expectations that surround novel MAI applications(14,15). The idea of MAI resolving the longstanding problems of healthcare is undoubtedly appealing but only theoretical: the actual efficacy and impact of these technologies remain to be assessed(16). The usage of MAI tools remains nascent and, as of now, there has not been a significant commercialization and clinical deployment(15).

The scarcity of studies proving MAI clinical efficacy is not the only criticism towards the introduction of these tools. Together with the promised benefits, deploying MAI in clinical contexts also entails some risks. Some of the most widely discussed are: inequality to access quality care as MAI

might not be deployed identically across different facilities, regions, and countries; transparency, as the workings of modern AI systems are often inaccessible to humans; biases and discrimination of underrepresented or vulnerable populations; and introducing automation bias, where doctors may trust MAI recommendations not on the basis of prove clinical efficacy, rather, on their perceived objectivity, accuracy, or complexity(15). Amongst the challenges introduced by MAI, there is the question of how it is going to transform doctors' daily practice and, in turn, their relationship with patients(17). Of particular importance is to understand how MAI could alter the shared decision-making (SDM) process between doctors and patients, what are implications it has for both doctors' and patients' autonomy, and which measures should be implemented to ensure patients' safety, health, and empowerment.

1.3.Artificial intelligence, doctors and patients

The impact of MAI on healthcare and the doctor-patient relationship is still uncertain and will probably vary with different applications and contexts(14,15). However, there seems to be a certain agreement that MAI will somehow significantly influence this relationship and modify the clinical interaction modes(17–19). Particularly discussed MAI applications are CDSS: they bear the greatest potential to modify the doctor-patient relationship since they enter the doctor decision-making and clinical evaluation processes thus possibly guiding and influencing doctors. Consequently, CDSS can affect the conversations that doctors have with patients, which information they share with patients, and to which degree the parties reach a shared decision. To try and answer these questions it is important to first define the doctor-patient relationship. After agreeing on the correct understanding of this relationship, a second relationship should be considered: the one between CDSS and doctors. Finally, it is possible to explore how to implement CDSS in clinical care without compromising the doctor-patient relationship and SDM.

The doctor-patient relationship is one of the most ethically significant aspects of healthcare(17). It is characterized by a fundamental asymmetry: a vulnerable patient seeks a doctor's comfort, support, and help to restore their health. The peculiarity of this relationship is that despite the asymmetry doctors and patients are partners in making decisions: both are, in their own way, experts(20). Doctors are experts in medical practice and patients in their values and preferences(21). The essence of this relationship is sharing information and decision-making, and giving care(17). The core elements are effective communication and respect for voluntary choices(22). It is therefore crucial to understand how CDSS can alter doctor-patient communication and autonomy, and whether, for example, they foster a more empathetic and compassionate relationship or stand in the way of true involvement(14,16).

There are two main concerns when considering the use of CDSS by doctors. The first one is commonly present in every field where AI is being introduced: job displacement. There are several claims about how CDSS is becoming more effective than doctors: it commits fewer errors, it is more accurate, it can process larger amounts of data, it does not get overworked or fatigued, it saves costs,

and it can identify patterns that the human eye does not catch(16). Alongside the idea that CDSS will eventually outperform doctors, hence rendering them obsolete, there are who maintain that CDSS will only augment, support, and empower doctors(7). According to the second attitude, CDSS can add value to medicine by assisting doctors in their clinical decision-making, guiding them through their reasoning, providing further information, and raising awareness of their cognitive and affective biases(18). Substituting doctors with AI tools is sometimes seen as an opportunity to diminish doctors' power and empower patients instead, thus freeing them from the burden of paternalism. This assumption is contradicted by those who believe this only introduces a new form of paternalism, where patients are dependent on technology instead of doctors(23). On the contrary, if doctors will only be assisted by CDSS, retaining the final decision and responsibility, they could exercise their professional autonomy. Since doctors' autonomy is a prerequisite for patients' autonomy and SDM, assistive CDSS might have a positive role to play in the doctor-patient relationship(14).

The second concern regards the understanding of CDSS by doctors. As many CDSS are based on ML and DL techniques, they usually possess what is described as a "black box": the reasons for their decisions cannot be accessed. The lack of causal insights into CDSS outputs can be worrisome since it could leave doctors clueless about why a tool is suggesting a certain diagnosis or treatment. This might entail that doctors will not discuss CDSS recommendations with patients, thus potentially preventing patients from making informed decisions(14). Black boxes have therefore a double implication in clinical practice: on the one hand, they might undermine SDM, on the other hand, they could prevent doctors from critically evaluating CDSS and deciding if, how, and when to integrate them into their clinical judgment(24). At the same time, it might not always be ethically significant to overcome this opacity. The accuracy of the system and the ability of doctors to include patients' values and preferences in their clinical judgment might be more important than transparent CDSS(14). Developing the necessary competencies for doctors to justify and explain procedures to patients in an understandable manner, even when suggested by black box CDSS, and discussing with them to reach a shared decision would be more important for patients' autonomy and empowerment than accessing the causal insights and workings of CDSS(23). These two concerns of the doctor-CDSS relationship are further explored in Chapter 3 of the thesis.

Transparency with patients is therefore of primary importance. This leads to questioning how important it would be to disclose and discuss MAI usage with patients. The doctor-patient relationship involves both interventions (for restoring health) and sharing of information for the sake of patients' knowledge, empowerment, and self-care(15). Not including MAI in these discussions may deprive patients of the ability to evaluate how it is affecting doctors' judgment(25). Especially in the case of underrepresented or historically marginalized populations, disclosing MAI could preserve the patients' autonomy and ensure their safety. At the moment there are no requirements to extend informed consent to MAI(25). However, since informed consent enables patients to manage their health and make their

voices heard, it should be considered whether it is the ethically appropriate thing to do. MAI influences doctors' considerations and constitutes a basis on which they make diagnoses and propose treatments(24,25). Usually, doctors are not expected to disclose their source of information (e.g. colleagues, journal articles). MAI could be an exception: it could influence doctors more than they are aware of while being black boxes, hence inscrutable and more difficult to assess and evaluate for both doctors and patients(26,27). Therefore, MAI might be different from any other source of information and tool used during the clinical assessment. However, there are a lot of obstacles to the implementation of informed consent for MAI and it might not always be beneficial for patients. For example, it could steer the conversation from care or it could further burden healthcare professionals with paperwork(28). Finally, although MAI bears the risk of undermining both doctors' and patients' autonomy, it could also be the case that instead it increases the opportunities for consent discussions(29). One way of doing this could be, for example, by automating routine tasks and hence freeing more time for doctors to discuss with patients(30). The idea of MAI gifting time to doctors is criticized: it could also be that doctors will visit more patients per day or would spend more time staring at screens(31,32). Informed consent requirements for MAI are therefore a very complex topic that combines ethical and practical considerations. Despite this complexity, it is an issue that requires to be addressed to ensure that patients' self-determination is not undermined by MAI applications.

Discussing MAI with patients and preserving doctors' professional autonomy might be the right path to promote patients' empowerment and safety with new technologies. But it might not be enough to avoid falling into a paternalistic trend, where doctors and patients are simply following MAI's recommendations. Therefore, consideration should be given to how to include patients' values and preferences in the final decision-making, also when MAI is informing and influencing it(33). Eventually, patients do not seek doctors merely for their medical knowledge, but also for someone who can understand them as socially embodied people with their uniqueness and values(15). The role of MAI in this ethically significant relationship should be attentively assessed. The intimate and dual relationship between doctors and patients might see a third actor enter the game, and it should be assured that balances will not change in a way that could preclude SDM. The doctor-patient relationship is further explored in Chapter 3 of the thesis, while Chapter 4 focuses on the eventuality of an informed consent requirement for MAI.

1.4.Cybersecurity issues are patients' safety issues

The doctor-patient relationship is based on a multitude of conditions. One of these is patients' safety: if patients do not feel like seeking medical assistance is going to benefit their overall well-being, they might renounce seeing a doctor. Patients' safety is a broad concept and captures many diverse aspects of healthcare which might be impacted by MAI. Informed consent and shared decision-making are two aspects that have already been mentioned. But while these are usually addressed when reflecting on the doctor-patient relationship, other relevant aspects are under-discussed, like cybersecurity.

Cybersecurity issues are not only IT problems: they constitute a major risk to patients' safety and care(34). When cyberattacks occur in healthcare facilities, entire networks are shut down and vital health data can no longer be accessed. For example, as a consequence of cyberattacks, surgeries and various therapies are often cancelled. Health data is particularly valuable for malicious hackers as it allows for identity theft, medical fraud, access to prescription drugs, and extortion(35). Therefore, health data is one of the most valuable types of data: it is not crucial only for patients' care, but it has a real value also on the dark web, where it can be sold for financial gain(34). Moreover, health data contains highly personal and sensitive information, that could potentially expose patients to stigma, discrimination, and embarrassment(36). Health data breaches can negatively impact public trust and undermine patients' willingness to seek out medical assistance and disclose sensitive, but needed, information(34). It is therefore imperative to strengthen healthcare cybersecurity to ensure patients' privacy, safety, and trust. Healthcare facilities cannot afford to have their activities disrupted by cyberattacks as this could lead to tangible harm to patients' health and even lives(15).

Robust cybersecurity is essential nowadays for patients' safety. The introduction of MAI systems in healthcare is one important technology issue and it can increasingly augment the longstanding cybersecurity vulnerabilities while introducing new ones(37,38). This is because MAI is often cloud-based, requires a vast amount of data, is geographically distant, integrates various networks and technologies, and relies on the electronic exchange of health data(38–40). The current cybersecurity model is inadequate to correctly tackle these new challenges(37). This can be particularly worrisome when faced with the continuous rise of cyberattacks on healthcare facilities(35,41). Coupled with the growing cybersecurity threat in healthcare, exacerbated by MAI but not solely caused by it, there is the longstanding systematic unpreparedness of this field in dealing with malicious intrusions. Indeed, no further cybersecurity measure or investment followed the implementation and adoption of new technologies in healthcare(35). Unfortunately, ignoring cybersecurity problems only prolongs time at risk(41).

AI does not only increase the cybersecurity vulnerabilities for healthcare but, at the same time, it can also contribute to protecting it. AI has, therefore, a dual role to play in healthcare cybersecurity: it further exposes the system to cyber threats and can be a tool for malicious hackers to try and access sensitive data, while also providing a means to analyse massive amounts of data and cybersecurity alters for identifying real dangers from false positives promptly(42). The limitation of AI-based strategies for cybersecurity is that, although they improve the degree of protection offered, they remain defensive approaches (e.g. preventing attacks with cyber-hygiene measures, detecting malicious activity, and responding to cyber incidents when they occur).

It is worth exploring cybersecurity strategies that have yet not been widely applied in healthcare and are not merely defensive. Instead, offensive strategies prioritise testing the system for identifying and patching vulnerabilities before they are exploited by malicious hackers. This is often referred to as

penetration testing, where professional ethical hackers are employed to simulate cyberattacks, hence testing the resilience and the weak points of the system. This approach was already proven to reinforce cybersecurity in other fields(43); it is worth exploring how it could be introduced in healthcare, where the stakes (human lives) are extremely high. This will be done in Chapter 5 of the thesis.

1.5.What our way of talking about AI tells about AI

The way in which we talk about AI, and MAI in particular, is not normatively neutral: it uncovers the preconceptions, values, and expectations we hold and how we understand and imagine MAI. Attention should be paid to the narratives of MAI that are present and perpetuated, in scientific discussions as well as in the media and pop culture. This is because narratives can influence the way we perceive and adopt MAI, its further developments and directions of research, and the availability of funding(44). The stories we tell about MAI, how we frame and conceptualise it, have a real-world impact how and where it will be implemented. At the same time, humans cannot refrain from creating and perpetuating narratives: it is in the nature of humans to make sense of things by narrating them. This meaning-making effort is therefore connatural to human existence and can, at the same time, influence its future. Narratives can be conceived as stories we tell about our lives and experiences that, in turn, impact our lives and experiences. Their being necessary creates a narrative responsibility for humans: we either assign meaning and values to things, situations, and so on, or we delegate this duty to other actors and passively accept their interpretation of reality. The narrative responsibility urges us to take on this duty ourselves to allow our worldview and values to be part of a collective effort for shaping humanity's future; or, in this case, the future of healthcare.

There are already many narratives about AI and they are widely studied. For MAI, instead, there are fewer studies. It is important to understand which narratives inhabit the MAI discourse and which future they envision for medicine. It seems that, particularly in healthcare, there is a lot of faith in the possibilities of MAI(45). MAI is expected to play an increasingly prominent role in medicine and tackle both old and new issues(7). These beliefs are usually influenced by hype narratives, which tend to focus on the benefits that these systems could provide, often exaggerating them and avoiding addressing the challenges they entail. Hype is very common in AI narratives: voices that think of AI as super-intelligent and imagine that it will soon take over the world are more widespread than the ones who see AI as limited systems that can perform one or some tasks usually associated with intelligent beings(46). In healthcare, one common hype narrative is the idea of a 'technological fix': every problem in healthcare could potentially be addressed with an appropriately designed MAI(45). Hype narratives misrepresent MAI and can be harmful as they envision futures outside of MAI's current capabilities while ignoring pressing issues and obstacles. It is of primary importance to identify misaligned narratives and to promote, instead, more truthful narratives, that are attentive to the reality of the technology, its limitations and advantages, and how it might better be implemented in healthcare. This requires also envisioning how the future doctor-patient relationship will be when MAI is involved in the clinical

evaluation and decision-making processes.

1.6.References

1. MIT Technology Review Insights. MIT Technology Review. 2021 [cited 2023 Aug 24]. Embracing the rapid pace of AI. Available from: https://www.technologyreview.com/2021/05/19/1025016/embracing-the-rapid-pace-of-ai/

2. O'Brien M. AP NEWS. 2023 [cited 2023 Jun 23]. EXPLAINER: What is ChatGPT and why are schools blocking it? Available from: https://apnews.com/article/what-is-chat-gpt-ac4967a4fb41fda31c4d27f015e32660

3. Khatsenkova S. euronews. 2023 [cited 2023 Jun 23]. ChatGPT: Is it possible to spot AI-generated text? Available from: https://www.euronews.com/next/2023/01/19/chatgpt-is-it-possible-to-detect-ai-generated-text

4. Marr B. Forbes. [cited 2023 Jun 23]. The Top 10 Limitations Of ChatGPT. Available from: https://www.forbes.com/sites/bernardmarr/2023/03/03/the-top-10-limitations-of-chatgpt/

5. Greenberger A. Artist Wins Photography Contest After Submitting AI-Generated Image, Then Forfeits Prize [Internet]. ARTnews.com. 2023 [cited 2023 Jun 23]. Available from: https://www.artnews.com/art-news/news/ai-generated-image-world-photography-organization-contest-artist-declines-award-1234664549/

6. Beyer EJ. The AI-Generated Art Debate Is Here. And It's Very Messy. [Internet]. nft now. 2022 [cited 2023 Jun 23]. Available from: https://nftnow.com/features/the-ai-generated-art-debate-is-here-and-its-very-messy/

7. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019 Oct 4;7:e7702.

8. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc. 2020 Oct;92(4):807–12.

9. Surya L. An Exploratory Study of AI and Big Data, and Its Future in the United States. Int J Creat Res Thoughts. 2015 May 2;3(2):991–5.

10. Burns E. SearchEnterpriseAI. 2021 [cited 2021 Jul 1]. In-Depth Guide to Machine Learning in the Enterprise. Available from: https://searchenterpriseai.techtarget.com/In-depth-guide-to-machine-learning-in-the-enterprise

11. Federchuk M. CENGN. 2022 [cited 2023 Jun 7]. The Difference Between AI, ML and DL. Available from: https://www.cengn.ca/information-centre/innovation/difference-between-ai-ml-

and-dl/

12. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digit Med. 2020 Feb 6;3(1):1–10.

13. Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. 'Brilliant AI Doctor' in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment. Proc 2021 CHI Conf Hum Factors Comput Syst. 2021 May 6;1–18.

14. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak. 2023 Dec;23(1):1–14.

15. Mittelstadt B. The Impact of Artificial Intelligence on the Doctor-Patient Relationship [Internet]. Council of Europe; 2021 Dec [cited 2023 Apr 14]. Report No.: F-67075. Available from: https://rm.coe.int/inf-2022-5-report-impact-of-ai-on-doctor-patient-relations-e/1680a68859

16. Van Cauwenberge D, Van Biesen W, Decruyenaere J, Leune T, Sterckx S. "Many roads lead to Rome and the Artificial Intelligence only shows me one road": an interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. BMC Med Ethics. 2022 Dec;23(1):1–14.

17. Dunn M. At the moral margins of the doctor–patient relationship. J Med Ethics. 2019 Mar 1;45(3):149–50.

18. Niel O, Bastard P. Artificial Intelligence in Nephrology: Core Concepts, Clinical Applications, and Perspectives. Am J Kidney Dis. 2019 Dec 1;74(6):803–10.

19. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics. 2021 Dec 1;47(12):e3–e3.

20. Aminololama-Shakeri S, López JE. The Doctor-Patient Relationship With Artificial Intelligence. Am J Roentgenol. 2019 Feb;212(2):308–10.

21. Godolphin W. Shared decision-making. Healthc Q Tor Ont. 2009;12 Spec No Patient:e186-190.

22. Chandra S, Mohammadnezhad M, Ward P. Trust and Communication in a Doctor-Patient Relationship: A Literature Review. J Healthc Commun [Internet]. 2018 Jul 19 [cited 2021 Apr 21];3(3). Available from: https://healthcare-communications.imedpub.com/abstract/trust-and-communication-in-a-doctorrnpatient-relationship-a-literature-review-23072.html

23. Segers S, Mertes H. The curious case of “trust” in the light of changing doctor–patient relationships. Bioethics. 2022;36(8):849–57.

24. Astromskė K, Peičius E, Astromskis P. Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. AI Soc. 2021 Jun 1;36(2):509–20.

25. Cohen IG. Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? Georgetown Law J. 2019 2020;108(6):1425–70.

26. Cohen IG, Graver H. A Doctor’s Touch: What Big Data in Health Care Can Teach Us About Predictive Policing [Internet]. Rochester, NY: Social Science Research Network; 2019 Aug [cited 2021 Jul 2]. Report No.: ID 3432095. Available from: https://papers.ssrn.com/abstract=3432095

27. Taddeo M, Floridi L. How AI can be a force for good. Science. 2018 Aug 24;361(6404):751–2.

28. Kent C. Medical Device Network. 2020 [cited 2021 Jul 8]. Artificial consent: should doctors be telling patients more about AI? - Verdict Medical Devices. Available from: https://www.medicaldevice-network.com/features/artificial-consent-should-doctors-be-telling-patients-more-about-ai/

29. Michalski A, Stopa M, Miśkowiak B. Use of Multimedia Technology in the Doctor-Patient Relationship for Obtaining Patient Informed Consent. Med Sci Monit Int Med J Exp Clin Res. 2016 Oct 26;22:3994–9.

30. Topol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. First Edition. New York: Basic Books; 2019. 378 p.

31. Sparrow R, Hatherley J. High Hopes for “Deep Medicine”? AI, Economics, and the Future of Care. Hastings Cent Rep. 2020;50(1):14–7.

32. Nittas V, Daniore P, Landers C, Gille F, Amann J, Hubbs S, et al. Beyond high hopes: A scoping review of the 2019–2021 scientific discourse on machine learning in medical imaging. PLOS Digit Health. 2023 Jan 31;2(1):e0000189.

33. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics. 2019;45:156–60.

34. Jarrett MP. Cybersecurity—A Serious Patient Care Concern. JAMA. 2017 Oct 10;318(14):1319–20.

35. Kruse CS, Frederick B, Jacobson T, Monticone DK. Cybersecurity in healthcare: A systematic review of modern threats and trends. Technol Health Care. 2017 Jan 1;25(1):1–10.

36. Hellsten H. Cyber risk management in the Finnish healthcare sector [Internet]. [Tampere, Finland]: Tampere University; 2018 [cited 2022 May 5]. Available from: https://www.semanticscholar.org/paper/Cyber-risk-management-in-the-Finnish-healthcare-Hellsten/567b836f752c4afabe95750f222c83b9df9ab357

37. Lohn AJ. Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity [Internet]. Center for Security and Emerging Technology; 2020 Dec [cited 2023 Jun 21]. Available from: https://cset.georgetown.edu/publication/hacking-ai/

38. Luna R, Rhine E, Myhra M, Sullivan R, Kruse CS. Cyber threats to health information systems: A systematic review. Technol Health Care Off J Eur Soc Eng Med. 2016;24(1):1–9.

39. Segal E. IEEE Computer Society. 2020 [cited 2023 Jun 14]. The Impact of AI on Cybersecurity. Available from: https://www.computer.org/publications/tech-news/trends/the-impact-of-ai-on-cybersecurity/

40. Das S, Siroky GP, Lee S, Mehta D, Suri R. Cybersecurity: The need for data and patient safety with cardiac implantable electronic devices. Heart Rhythm. 2021 Mar 1;18(3):473–81.

41. Martin G, Kinross J, Hankin C. Effective cybersecurity is fundamental to patient safety. BMJ. 2017 May 17;357:j2375.

42. Ravichandran H. Forbes. [cited 2023 Jun 23]. How AI Is Disrupting And Transforming The Cybersecurity Landscape. Available from: https://www.forbes.com/sites/forbestechcouncil/2023/03/15/how-ai-is-disrupting-and-transforming-the-cybersecurity-landscape/

43. Caldwell T. Ethical hackers: putting on the white hat. Netw Secur. 2011 Jul 1;2011(7):10–3.

44. Cave S, Craig C, Dihal K, Dillon S, Montgomery J, Singler B, et al. Portrayals and perceptions of AI and why they matter. R Soc. 2018 Nov;

45. Gardner J, Warren N. Learning from deep brain stimulation: the fallacy of techno-solutionism and the need for 'regimes of care'. Med Health Care Philos. 2019 Sep 1;22(3):363–74.

46. Kayser-Bril N. Using ChatGPT for (re)search is a philosophical break [Internet]. Algorithm Watch. 2023 [cited 2023 Jun 21]. Available from: https://r.algorithmwatch.org/nl3/q_7P5fwseQoWIpcLRde4Kw?m=AWcAABxNlGkAAAARy_AAAChn8AUAAAAA98EAAB1gABB0KQBkgrJtzIb6hsKZQoqc6Mnb_drzoQAQJ0I&b=eac9a7ac&e=13493a7f&x=KNeUbuOyH-FBJoqIxZ5AXZ_W8G65_G9sz6N3PfbjexM

Chapter 2

Methodology

2.1. Methodology

This thesis is part of a larger project (EXPLaiN – Ethical and Legal issues of Mobile Health-Data: Improving understanding and eXPlainability of digitaL transformAtion and data technologies using artificial IntellgeNce) funded by the Swiss National Science Foundation (project number 407740_187263) as part of a national research programme (NRP77). EXPLaiN investigated concerns regarding bias, data protection, lack of transparency and explainability. EXPLaiN also aimed to clarify the legal and ethical issues that need to be resolved for the collection, usage and analysis of health data with artificial intelligence methods. The first part of the study consisted of 41 semi-structured interviews with international experts on medical artificial intelligence (MAI) from the fields of medicine, law, ethics, public health, and computer science. The interviews focused on the barriers and facilitators for the implementation of AI in clinical settings, particularly regarding clinical decision support systems and wearable devices. The aim was to examine current attitudes, knowledge, and barriers to using AI models to analyse health data and to support doctors and patients in their decision-making. The second part of the study consisted of 22 interviews with patients from Swiss hospitals about medical AI, with the same goal of exploring current attitudes, knowledge, and barriers to its implementation. The third part of EXPLaiN research project aimed to develop an explainable machine-learning algorithm with health data collected by the anesthesiology department of the University Hospital of Basel in collaboration with ETH.

The present thesis focused on the ethical aspects of MAI for the doctor-patient relationship and the doctors' daily practice. It emphasised particularly the role of both doctors' and patients' autonomy while drawing attention to the way we talk about MAI. Indeed, narratives of MAI can play an important role in how it is developed, perceived, and introduced into healthcare.

2.2. Research Objectives

In order to achieve the goal of an ethical adoption of MAI, this thesis investigated the following questions. Chapter 3 addressed the question of how MAI could impact the doctor-patient relationship within a normative framework. Considering the huge impact that MAI is expected to have on clinical practice, it seemed reasonable to imagine that it could affect the relationship between doctors and their patients(1). Starting with a description of SDM, identified as the ethically appropriate paradigm for the relationship, we analysed how MAI could change doctor-patient communication. To this end, the principle of respect for autonomy was recognized as central and further investigated both for doctors and patients. With the routine implementation of MAI, the SDM dual relationship might be reimagined as a triad, involving the patient, the doctors, and the AI.

Chapter 4 faced the question of an informed consent requirement for MAI. Continuing the discussion of the previous chapter on doctor-patient communication and relying on a normative approach, it evaluated the benefits and drawbacks of introducing such a requirement. Introducing MAI

in clinical practice challenges the boundaries of informed consent and raises a panoply of questions about its role, scope, and appropriateness. The analysis explored the theoretical arguments and the practical implications of informed consent. The goal was to bridge these two sides of the discourse about informed consent for MAI to identify the way forward. Eventually, some sort of informed consent should be implemented.

In Chapter 5 there is the question of how to bolster healthcare cybersecurity. Since the widespread use of interconnected technologies, such as MAI, further augment the system's vulnerabilities, it is crucial to safeguard sensitive health data and patients' well-being(2,3). To reinforce the cybersecurity of healthcare facilities a different approach is examined: employing ethical hackers (specifically penetration testers) for ensuring the system's resilience. A normative analysis considered hacker ethics and its compatibility with the healthcare situation and needs, concluding that pen-testers could in fact contribute to healthcare cybersecurity.

Finally, Chapter 6 posed the question of which narratives are there about MAI and what is their role. Identifying and evaluating MAI narratives is important to understand the direction in which MAI is going to be developed and implemented in clinical practice. In order to do so, a thematic narrative approach was adopted. This allowed us to identify narratives in the relatively short fragments that were selected for the secondary analysis. The data is part of the experts' interviews and consists of a subset of 30 participants. Together with the qualitative analysis, there was a normative intent to bring attention to the narrative responsibilities that we carry. Here, the normative conclusions were based on the empirical results and mainly revolve around the importance of establishing a more nuanced and truthful narrative for MAI.

2.3. Normative Analysis

This thesis relied on normative analysis, namely the investigation of how one ought to act or the world ought to be. It entails thinking systematically about the concepts of good or bad, right or wrong, just or unjust, required or prohibited(8). This systematic reflection makes use of ethical principles, values, and virtues. In the context of bioethics, of particular importance are the four principles enunciated by Beauchamp and Childress: respect for autonomy, beneficence, non-maleficence, and justice(9). The first principle had more relevance for this thesis as it is central to both Chapters 3 and 4. Here respect for autonomy was considered in the light of the doctor-patient relationship, the doctors' need for professional autonomy, and the patient's right to self-determination. The doctor-patient relationship was conceived as ethically relevant and therefore worthy of reflection(10). On this basis, we attempted to develop a moral framework that could describe and prescribe the ethically appropriate paradigm for a good doctor-patient relationship, also when MAI is employed. Eventually, this is the overarching goal of normative analyses in bioethics: advancing prescriptive or evaluative statements about how an issue should be understood and which actions (or policies) should be undertaken(8).

A second focus of the normative analysis conducted here regarded hackers' codes of conduct. These codes of conduct were subjected to a normative analysis to determine their conception of ethics and whether or how it could be integrated with the needs of healthcare cybersecurity. Here the starting point already had a normative valence as codes of conduct prescribe the appropriate actions to undertake and define what is morally good. The question was to identify, describe, and categorize the type of ethics implicit in these codes. Therefore, the goal was to explicate the implicit principles and values and further evaluate them in the sensible context of healthcare.

In synergy with a qualitative analysis, narratives of MAI were also analysed normatively. Drawing from empirical evidence, where narratives were described and their impact on MAI assessed, the normative analysis aimed to identify which narratives could be considered more ethically appropriate. This normative analysis appealed to the idea of narrative responsibility. Narrative responsibility constituted the basis on which it was possible to justify normative claims. Besides the idea of narrative responsibility, central to this normative analysis was also hermeneutic theory, according to which meaning units are not isolated but always in relation to the whole (e.g. the social context)(11,12). Therefore, narratives were not analysed as isolated stories, but hermeneutically as means to create and perpetuate meaning, actions, and ethical values.

The general structure of the normative analyses conducted in this thesis is as follows: the first step required an overview and description of the problem. This was attained either with literature research or empirical evidence, which informed the subsequent systematic reflection. The description was essential for understanding the state of things, identifying and contextualising the ethical issues, acknowledging the conclusions that were already been made, and recognising the relevant ethical principles that needed to be further addressed and examined. The systematic reflection considered the different values and principles at stake and how they could impact MAI deployment. This called for the identification of best practices that could foster an ethical adoption of MAI for doctors, patients, and the healthcare system. Finally, conclusions (in the mentioned form of "best practices") had a prescriptive, and not only descriptive, value.

2.4. Empirical Bioethics

The present research (particularly Chapter 6) brings together normative and empirical analyses under the framework of empirical bioethics. Over the past two decades, the interest in empirical bioethics increased (this is often referred to as the "empirical turn")(4). Empirical bioethics can be described as a type of research that investigates bioethical problems by drawing on the strengths of both normative and empirical analyses(5). Its methodology is still evolving as it needs clarification on how to better combine the normative and the empirical approaches(4,5). Therefore, the term "empirical bioethics" does not refer to one established and definitive methodology, but rather, to a varied and broad range of different approaches.

The arguments for using empirical bioethics methodologies rely on the complementarity of the normative and empirical analyses, which would both profit one from the other. Empirical research can better contextualize ethical concepts and problems in their broad social context, can provide evidence for normative claims, can capture the relevant concrete aspects of medical practice, can assess the consequences of the chosen normative solution, and it can decrease the odds of biased conclusions(4,6). However, empirical research alone cannot solve moral dilemmas: it simply describes the world as it is and only a normative approach can prescribe how this world ought to be(4). The goal of introducing empirical research in bioethics is not to replace normative analysis, instead, it should enrich moral reasoning(6). The normative analysis is essential for avoiding stopping at the individual case and it stands against the idea that the majority creates the moral rules(4,6).

Depending on the approach, the empirical and the normative analysis can be present at different stages of the research but they are always complementary. For example, problems can be identified either way. If the problem was identified empirically, the normative analysis of the (sometimes hidden) values entailed becomes crucial. Instead, beginning from the normative definition of a problem requires the empirical exploration of how the problem presents itself in everyday reality(7).

In this thesis, normative analysis is prevalent; only in Chapter 6 it is accompanied by an empirical approach (although not a classic one as it was combined with narrative analysis). There, the problem was identified through empirical analysis and emerged from the data analysis: it became apparent that there were peculiar narratives that called for further research. Based on the empirical results, a normative discussion on narratives was conducted. Chapter 6 used empirical bioethics and supplemented an empirical analysis (more on it in the section "Thematic Narrative Analysis") with a normative analysis of MAI narratives based on data.

2.5. Empirical Analysis

Part of the research results presented in this thesis are based on an empirical analysis (data from part 1; patients' data, namely part 2, was not included in this thesis). It consisted of a qualitative exploration study. Participants were purposively sampled on the criterion of being knowledgeable in the field of MAI and having a background either in Medicine, Philosophy, Public Health, Law, Computer Science or related disciplines. Another inclusion criterion was that of holding a senior position; the goal was to identify participants who have a longstanding experience with MAI (either first-hand or theoretical). Participants were recruited internationally, however, we also aimed to investigate experts' opinions on MAI on a Swiss level. Hence, almost half of the participants (n=14) were Swiss experts. The rest of the participants were from Europe (n=20), America (n=6), and Africa (n=1), for a total of 41 experts interviewed.

First contact with participants was through email where they were invited to be interviewed by introducing the project, explaining the aims, and the implication of their participation (e.g. time

commitment, voice recording, the method of transcription, data pseudonymisation format). Ethics approval was obtained from the Ethics Committee of Northwest and Central Switzerland (EKNZ). We obtained oral informed consent (in this case written informed consent was not required as the research did not fall within the Human Research Ordinance) before each interview; all participants consented to be involved and declared their understanding of the aims of the research.

The semi-structured interviews were conducted online (the first two – the case studies – were exceptionally conducted in person between March and April 2021) from November 2021 to May 2022. We used Zoom for conducting and recording the interviews, which lasted between 40 and 80 minutes. Interview questions were based on an interview guide composed of 13 questions, each with several prompts or follow-ups. The questions were divided into 6 blocks: introductory questions (about the experience of the participant), general questions about using AI in medical practice, context-related questions about AI-patient relationships (vignette 1 involving a wearable device), context-related questions about doctor-patient relationships with AI (vignette 2 involving CDSS), context-related questions about private-public relationships (vignette 3), and closing questions.

All the interviews were pseudonymised and transcribed. Once data saturation was reached, the interviews were inductively coded(13). The first ten were coded by the team; during these sessions, the code system was developed. Topics emerged from the interviews and were subsequently gathered in the codes. Therefore, the code system grouped the main topics found in the data, for example, informed consent, data handling, narratives and value judgments, doctor-AI interaction, patient-doctor relationship with AI, the potential benefit of AI, private-public interaction, and regulations. Each code had several sub-codes for a total of 106 codes. The rest of the interviews were coded individually by L. Arbelaez Ossa and me. During this process, we held weekly meetings to discuss the code tree and make any necessary changes. Some codes have been merged, re-arranged, or created. We modified the initial tree code only to the necessary extent. During these meetings, we also started discussing themes and have thus preliminarily combined codes into sets. The approach chosen for analysing the data collected through these interviews was thematic reflexive analysis(14).

In Chapter 6, data was analysed by combining thematic reflexive analysis with narrative analysis, namely using a thematic narrative approach to identify and reflect on the stories that participants told about MAI(15,16). For this purpose, the chosen sub-set of data was re-coded, giving rise to a secondary analysis of the experts' interviews. The specific method used in Chapter 6 is described in detail in the method section of the aforementioned Chapter and it is therefore not presented here.

2.6. Individual Contributions

L. Arbelaez Ossa and I prepared the documents to obtain ethical approval from the competent Research Ethics Committee (EKNZ). L. Arbelaez Ossa prepared the submission for the qualitative part

of the study while I took care of the health data transfer part. I also managed the legal agreement between USB, Unibas, and ETH for the transfer of anaesthesiology data from USB to ETH.

Interview guides were drafted by L. Arbelaez Ossa and me under the direct supervision of Dr D. Shaw, Dr M. Rost, and Prof B. Elger. Recruitment and interviews were carried on by L. Arbelaez Ossa and I, who also coded the interviews. In the team coding sessions participated Dr T. Wangmo, Dr D. Shaw, Dr M. Rost, and Dr S. Milford. Dr E. De Clerq together with Dr D. Shaw supervised the narrative thematic analysis that I carried out in Chapter 6.

Chapters 3 and 4 were written with the contribution of Prof B. Elger, Dr M. Shaw, and L. Arbelaez Ossa. Prof B. Elger and Dr M. Shaw also contributed to Chapter 5. Chapter 6 saw the collaboration of the previously mentioned authors plus Dr E. De Clerq and Dr S. Milford.

2.7. References

1. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics. 2021 Dec 1;47(12):e3–e3.

2. Luna R, Rhine E, Myhra M, Sullivan R, Kruse CS. Cyber threats to health information systems: A systematic review. Technol Health Care Off J Eur Soc Eng Med. 2016;24(1):1–9.

3. Muthuppalaniappan M, Stevenson K. Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health. Int J Qual Health Care. 2021 Jan 1;33(1):mzaa117.

4. Leget C, Borry P, de Vries R. 'Nobody tosses a dwarf!' The relation between the empirical and the normative reexamined. Bioethics. 2009 May;23(4):226–35.

5. Davies R, Ives J, Dunn M. A systematic review of empirical bioethics methodologies. BMC Med Ethics. 2015 Dec;16(1):1–13.

6. De Vries M, Van Leeuwen E. Reflective Equilibrium and Empirical Data: Third Person Moral Experiences in Empirical Medical Ethics. Bioethics. 2010;24(9):490–8.

7. Birnbacher D. Ethics and Social Science: Which Kind of Co-operation? Ethical Theory Moral Pract. 1999 Dec 1;2(4):319–36.

8. Viens AM. The Fundamental Importance of the Normative Analysis of Health. Health Care Anal. 2019 Mar 1;27(1):1–3.

9. Beauchamp TL, Childress JF. Principles of biomedical ethics. New York: Oxford University Press; 1979.

10. Segers S, Mertes H. The curious case of “trust” in the light of changing doctor–patient relationships. Bioethics. 2022;36(8):849–57.

11. Gadamer HG. Truth and Method. A&C Black; 2013. 665 p.

12. Geanellos R. Exploring Ricoeur’s hermeneutic theory of interpretation as a method of analysing research texts. Nurs Inq. 2001 Dec 25;7(2):112–9.

13. Joffe H. Thematic Analysis. In: Qualitative Research Methods in Mental Health and Psychotherapy [Internet]. 1st ed. John Wiley & Sons, Ltd; 2011 [cited 2023 Jul 26]. p. 209–23. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9781119973249#page=212

14. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol. 2022;9:3–26.

15. McAllum K, Fox S, Simpson M, Unson C. A comparative tale of two methods: how thematic and narrative analyses author the data story differently. Commun Res Pract. 2019 Oct 2;5(4):358–75.

16. Braun V, Clarke V. Using thematic analysis in psychology. Springer. 2016 Jul 27;3(2):77–101.

Chapter 3

Artificial Intelligence and the Doctor-Patient Relationship:

Expanding the Paradigm of Shared Decision-Making

Citation: Lorenzini, G., Arbelaez Ossa, L., Shaw, D. M., & Elger, B. S. (2023). Artificial intelligence and the doctor–patient relationship expanding the paradigm of shared decision making. *Bioethics*, 37, 424–429. https://doi.org/10.1111/bioe.13158.

Artificial Intelligence and the Doctor-Patient Relationship: Expanding the Paradigm of Shared Decision Making

Abstract

Artificial intelligence (AI) based clinical decision support systems (CDSS) are becoming ever more widespread in healthcare and could play an important role in diagnostic and treatment processes. For this reason, AI-based CDSS has an impact on the doctor-patient relationship, shaping their decisions with its suggestions. We may be on the verge of a paradigm shift, where the doctor-patient relationship is no longer a dual relationship, but a triad. This paper analyses the role of AI-based CDSS for shared decision-making (SDM) to better comprehend its promise and associated ethical issues. Moreover, it investigates how certain AI implementations may instead foster the inappropriate paradigm of paternalism. Understanding how AI relates to doctors and influences doctor-patient communication is essential to promote more ethical medical practice. Both doctors' and patients' autonomy need to be considered in the light of AI.

Keywords

Artificial Intelligence, Autonomy, Doctor-Patient Relationship, Healthcare, Shared Decision-Making.

Introduction

The introduction of *artificial intelligence* (AI) technologies in healthcare promise enormous benefits, particularly concerning quality, costs, efficiency, and access. AI applications are broad and diverse: the present analysis considers the implications of AI used in hospitals as clinical decision support systems (CDSS). CDSS can suggest diagnoses, make predictions, and recommend treatments, thus assisting physicians, nurses, patients, and other caregivers' decision-making processes(1,2). CDSS are not a novelty: computer-based ones already existed in the 1970s(1). However, they were poorly integrated with patient care and it was not until recently, combined with AI and electronic health records (EHR), that they started to become increasingly desirable for clinical practice. Thanks to the embedding of AI and EHR with CDSS, they can provide valuable diagnostic suggestions based on patient data and test results, as well as support patient safety, clinical management, cost containment, and administrative functions(1). AI-based CDSS are gaining momentum as they promise faster and more accurate decisions and diagnoses(3). As the focus of this paper lies on CDSS applications of AI, the general term "AI" will be used as an abbreviation of AI-based CDSS.

Any tool aiming to enhance doctors' diagnostic abilities and quality of care may be life-saving for many people: diagnostic errors alone contribute to approximately 10% of patients' death in the United States(4). However, the adoption of these technologies is not unproblematic. AI generates new questions and challenges for the doctor-patient relationship as it bears the potential to transform clinical interaction modes(5): "although AI systems have the potential to empower humans in medical decision-making, they also run the risk of limiting autonomy and creating new obligations"(6). This paper focuses on the impact that AI can have on the doctor-patient relationship while trying to identify its benefits and burdens for the shared decision-making (SDM) paradigm. Although the consequences of AI for SDM are beginning to be discussed by academics (e.g. Eric Topol's *Deep Medicine*(7)), this paper aims to contribute to the discussion by providing a cautionary tale of the potential consequences of AI for both doctors and patients. After introducing the SDM paradigm, it will show how AI could relate to it, and evaluate the importance of both doctors' and patients' understanding, communication, and autonomy. There is a danger of a shift back towards paternalism if sufficient attention is not given to preserving the foundations of SDM, namely doctors and patients' understanding, communication, and autonomy.

The shared decision-making paradigm

Nowadays, much attention is paid to patients' autonomy, and it is generally thought that doctors should facilitate patients' participation in managing their health. Patients' right to direct their own care, namely to hold views, make choices, and act according to their personal values, should be acknowledged(8,9). Respect for patients' autonomy is considered to be one of the fundamental principles of contemporary medical practice: "respect for autonomy is not a mere *ideal* in health care; it is a professional *obligation*. Autonomous choice is a *right* – not a *duty* – of patients"(9). Therefore, it

can be argued that a paradigm promoting patients' autonomy is more ethical than one suppressing it. Indeed, it is broadly accepted as the ethical appropriate paradigm(10). The SDM paradigm can empower patients and get them more involved in their healthcare, hence allowing them to exercise their values and autonomy(11). By promoting SDM, patients' self-determination is promoted too. Ideally, this collaborative doctor-patient relationship should be an encounter between two experts: physicians are experts in medicine while patients are experts in their own values(12,13). Both sides need to respect the other's expertise, and have the duty to inform the other: the doctors should disclose the procedure and associated risks and benefices, possible alternatives, prognosis, and consequences of each clinical decision, and the patients should indicate their preferences and personal values. Doctors no longer "care for" as much as "care with" their patients(8). SDM is a process and, as such, it involves many factors that can either contribute or hinder making a shared decision. An element that can foster this process is informed consent since it calls for doctors to disclose and explain information to patients. Therefore, informed consent can be a part of SDM: without adequate information, patients have an inadequate basis for decision-making(9). Informed consent is the disclosure of appropriate information to competent patients who then can actively participate in decisions concerning their health(14). The information given to patients forms the basis for them to exercise their autonomy: without this information, patients would not be able to consciously contribute to the decision-making. Only when patients have received enough information SDM can take place.

When AI is included in the relationship, it can support SDM, if carefully implemented. This is because AI plays a role in the decision-making process. The aspects that need to be considered are how AI influences doctors and patients' communication and autonomy. In the event that CDSS is used without careful attention to these aspects, it could lead to a paternalistic doctor-patient relationship. While SDM has been identified as the ethically appropriate paradigm for healthcare, paternalism disregards patients' values, understanding, and autonomy.

According to the paternalistic paradigm, doctors attempt to steer patients' decisions to what they think is in the patients' best interest, e.g. by strongly recommending a course of action or by offering them a partial range of options. In the extreme form of paternalism, patients may have no decisional power, but that is rarely the case: usually, paternalistic doctors try to convince patients to choose whichever option the clinicians think is best. When decisions have to be made, paternalistic doctors may override patients' wishes(15). As a consequence, the doctors' decisions are not always in line with the patients' wishes, which often remain unheard. Indeed, in a paternalistic doctor-patient relationship, the doctors act on the patients' behalf, but not at the patients' behest(8). This does not mean that doctors intend to mistreat patients, rather, their non-observance of the patients' autonomy has the ultimate goal of doing what they consider to be the best for them. However, without opening a dialogue with patients, doctors cannot know if what they consider to be best is truly good from the patients' point of view. In this situation, it is difficult for patients to understand what is happening to their bodies, because they are

not given sufficient explanations about their diagnosis, prognosis, and treatment options, with the correlated benefits and risks. Therefore, a paternalistic doctor would not seek the patient's consent and might not disclose all relevant information. Today's doctors who tend towards paternalistic behaviour, although under the legal obligation to ask for patients' consent, would try to limit information and influence patients towards what they believe is best for them. Thus, paternalistic doctors do not promote their patients' autonomy.

This paper holds SDM as the opportune paradigm for the doctor-patient relationship. Instead, a paternalistic relationship is considered undesirable in principle. Accordingly, it conceptualizes different ways in which AI-CDSS could preserve SDM while trying to tackle those aspects that may promote paternalism.

AI and shared decision-making

The doctor-patient relationship is a consensual partnership in which patients seek and accept the assistance of a physician to manage their health. They collaborate to achieve the highest standard of care while respecting patients' autonomy, communicating and explaining options, and obtaining informed consent(16). Therefore, key elements of the relationship between doctors and patients are effective communication and respect for voluntary choices(17). These are the pre-conditions for SDM, and their compliance is independent from the presence of AI. However, with the introduction of AI into the equation, the enforcement of these key elements may be at risk: both the interaction between doctors and AI and the communication of this interaction to patients should be considered.

a) AI-doctor communication and autonomy

During the clinical evaluation process, doctors assess complex clinical evidence to reach a diagnosis. When AI is used to assist in diagnosis, its suggestions will become part of this evaluation process. AI's suggestions could guide doctors' decisions more than they are aware since their outputs affect, shape, and even stand in tension with doctors' judgments, thus raising questions on who is truly guiding the decision-making process(5,18). AI's extensive influence could limit doctors' autonomy(6,19). Accordingly, some doctors worry that AI could "decrease their control over decision making" and that it may be a "threat to professional autonomy"(3). Doctors' professional autonomy is a condition for freely exercising their clinical judgment in patient care: doctors need to have the necessary autonomy to take decisions on the care of their patients(20). This autonomy also corresponds to a responsibility on the part of doctors to provide their best care to patients(20). The close link between doctors' autonomy and responsibility is due to the nature of this autonomy: it is granted to doctors because doing so provides a benefit to society, and that benefit would be good care(21). Therefore, doctors' professional autonomy is important because it is a requisite for practising their judgment and

decision-making, while also holding them accountable for those, and it needs to be preserved as it provides a service to society. What is at stake here is doctors' freedom to decide both the conditions for practice – for example, how AI will be implemented – and to act according to their best clinical judgment "to promote patients' best interest, not their best interest"(22). Since SDM is a collaborative relationship, it is essential that both parties preserve their ability to make informed and autonomous decisions. Indeed, doctors are an active part of the SDM process and, as such, they shall be capable of autonomous clinical judgment and decision-making. This requires both doctors' and patients' autonomy to be respected. Two requirements for doctors' autonomy are competence, both clinically and with AI, and ability to make their own decisions based on clinical and contextual evidence.

Competent doctors collaborate with AI while assessing its recommendations and checking for errors(6,19,23). Of foremost importance is to identify which level of understanding is necessary for doctors to integrate AI's recommendations into daily practice while maintaining a critical eye. If doctors understand the implications and underlying assumptions of AI, they will be better positioned to evaluate its outputs, thus more confidently deciding whether or not to rely on them for their own decision-making. Hence, physicians may seek to understand the reasons underlying AI's recommendations to evaluate their validity and to be able to explain to patients their impact on the clinical evaluation process(24). One obstacle to doctors' understanding can be the opaqueness of some AI systems, namely the *black box* problem. Black box AI increases the complexity of the communication process as it does not offer explanations of its decisions and operations(23). Solving the black box problem can support the AI-doctor relationship, but requirements for AI to be explainable should be endorsed only if explanations consider the specific context, background knowledge, and interests of doctors, rather than being solely mathematical(25). In the explanatory process, several factors should be included, such as premises, implications, and the AI's output in relation to the real-life context(25,26). However, it must be noted that this should not be regarded as a *deus ex machina* solution. The offer of causal explanations of AI behaviour and the apparent transparency of knowing the causal relationships between the input and output does not necessarily translate to understanding the implications of using AI and its assumptions. In the same way, understanding AI's causal inferences may not be always required to evaluate its recommendation. It is above all fundamental for doctors to be aware of AI's usability and limitations in the context of implementation. Therefore, doctors would not need to know everything about AI and how it arrived at a certain recommendation, but rather, they would benefit from understanding the underlying assumptions of a decision. For example, is AI basing its analysis on similar data/situations or is it considering family history to increase or decrease a risk assessment?

While explainability may be useful, excessively concentrating on the black box issue can potentially overshadow other issues that can equally or more strongly impede doctors' capacity to work with AI. Supposing that there is fully transparent and highly performing AI, the problem remains as to how to train doctors to understand and evaluate its results to a degree they are competent enough to

remain autonomous and decide how, when and if to integrate AI in their clinical judgement. Explainability alone does not guarantee AI-doctor or doctor-patient communication; rather, motivation and time constraints may be equally important factors to be addressed.

Another consideration in terms of optimal AI-doctor collaboration and preserving physicians' professional autonomy, is the assistive nature of these tools: they are designed to inform, assist, and empower clinicians, not to replace them(2,19). It is unlikely that in the near future AI will replace humans as the final decision-makers in the healthcare context(27). Despite this, there is a strong narrative suggesting that AI is a competitor with doctors, standing against their expertise, and undermining their profession. For example, a venture capitalist from Silicon Valley once proclaimed that "machines will replace 80% of doctors" and "radiologists will be obsolete in five years"(28). Although they made these statements a long ago, this has not occurred; currently AI can only be a tool for clinicians and not a substitution(29). Describing AI as a rival and autonomous agent does not foster a good ground for introducing AI as a further collaborator in the clinical practice.

Ultimately, current AI-based tools lack the contextual and emotional intelligence needed to make decisions in uncertain, risky, and emotionally fraught circumstances: "some decisions are not simply a matter of survival-based logic"(30). The conclusion is not to avoid using AI until when they are "intelligent" enough to be autonomous deciders. On the one hand, this is something that we may never want to happen; on the other hand, underuse of AI could increase risk of harm for patients and be burdensome(31). If it is not desirable for doctors to avoid using AI, neither is it desirable to exclude doctors from the clinical decision-making process: clinical practice is more likely to implement human-in-the-loop setups, where doctors actively collaborate with AI systems, provide oversight, and decide what, when, how and why integrate its outputs in their clinical judgment(6). What should be advocated is then a collaborative partnership between AI and doctors: AI systems should collaborate with humans instead of competing against them(6). This collaboration would allow doctors to exercise their autonomy and to take care of those aspects that even a perfect algorithm cannot handle (such as empathy, risk communication, assessment of patients' values, hopes, fears, and expectations(30)) while also promising better performance(6). By joining forces, AI and doctors may provide better care than either AI or clinicians alone(32). This is crucial because doctors' professional autonomy is a pre-requisite for SDM with AI: only if doctors' autonomy is preserved they can promote patients' autonomy as this allows them to communicate transparently with patients and explain them difficult information(13). As a consequence, patients are better positioned to participate in the SDM process even if AI is used. Otherwise, a double paternalism[1] would be established: first between AI and doctors, with the latter

1 The concept of "double paternalism" is not new, nonetheless, in this context it is used in a new way. Traditionally, double paternalism refers to a combination of medical and social paternalism in the case involuntary hospitalisations, e.g. forced psychotherapy. However, here double paternalism means the establishment of a paternalistic relationship first between the AI and the doctors, and consequently between the doctors and the patients. These two usages of the "double paternalism" concept are therefore different.

doing as they are told, then between doctors and patients, with patients doing as they are told in turn and the doctors left as intermediaries. In this scenario, it would be more difficult for doctors to consider results, detect errors, and disagree with a paternalistic AI. AI has the potential to sustain doctors' autonomy, but only on the precondition of good communication, otherwise, they may become mere passive executors of AI's decisions. Indeed, some experts fear that doctors may become less the deciders, and more the messengers of AI's outputs(10,33). AI-doctor communication enables doctors to actively participate in the decision-making process because this comprehension makes them more aware of the motives of their (dis)agreement(34). Doctors' ability to make decisions autonomously is an essential part of the SDM process. Accordingly, good AI-doctor communication is essential both for doctors' autonomy and for SDM.

b) Doctor-patient communication and autonomy with AI

AI can have an enormous role in shaping doctors' decisions, so doctors may be required to inform their patients when AI is included in the clinical evaluation(35). Providing this type of information to patients may help them understand better the reasons for a diagnosis, the different alternatives, and the prognosis. As a consequence, patients would be better positioned to participate in the decision-making process. While explainability can contribute to doctors' understanding and evaluation of AI's recommendations, alone, it is not sufficient to safeguard patients' autonomy, as has been previously argued. Doctors not only need to assess AI's suggestions, but also need to be able and willing to communicate with patients. At the core of doctor-patient collaboration, there is the willingness of both parties to communicate. AI alone cannot ensure that this communication takes place. AI is not a threat to patients' autonomy only if doctors are predisposed to disclose and discuss it. The prerequisite for this is good communication between AI and doctors. Therefore, AI-doctor communication not only serves the function of preserving doctors' autonomy, but also enables doctors to include patients in this decision-making process, thus fostering patients' autonomy.

AI poses a risk of establishing a new form of paternalism, where the "computer knows best"(10). This is a possibility because AI's recommendations might not take into consideration patients' values; for example, the only value guiding its recommendation might be the goal of maximising lifespan. While it can be argued that it is a common shared value, it is also true that not all patients aim to prolong their lives: at the same stage of a terminal disease, one patient will choose palliation, while another will opt for further therapy(30). "In clinical settings, there can be no one-size-fits-all decision threshold"(27). Medical decisions are not based solely on clinical information, but are intertwined with preferences, values, risk tolerance, and many other personal factors that must be weighed in the decision-making process(10). AI can better support patients by considering their preferences and unique situations: while the doctors are important intermediaries and can inquire about patients' preferences and ensure these

are considered, patients would have one more guarantee that their values were being respected if these preferences were already included in the AI evaluation.

Currently, the values behind AI decisions are hidden behind the algorithm; moreover, companies and institutions, rather than patients, influence these values(30). The first required step is to identify and expose the values embedded into the AI. Therefore, doctors (and possibly patients) should be aware of AI core values, and ensure that, eventually, patients' values are safeguarded and prevail over competing views. Doctors should ensure that patients' specific preferences are taken into account, thus facilitating SDM between patients, doctors and AI. It could be imagined to incorporate patients' preferences and risk-taking attitudes in the algorithm so that they would be considered e.g. when proposing a treatment: "respect for patients' autonomy means that patients' values should drive the ranking process"(10). That would be the second step (together with a successful AI-doctor-patient communication) to attaining an optimal AI-doctor-patient partnership. Ensuring that AI respects patients' autonomy is fundamental for avoiding paternalism and enabling patients to participate in SDM.

Conclusion

It is certainly challenging to introduce AI in healthcare, but this should not be a sufficient reason to desist. While AI may profoundly alter the doctor-patient relationship, this change is not necessarily for the worst since it could further foster SDM, if carefully implemented. However, it should be borne in mind that this will also involve a paradigm shift: while SDM principles may not vary, the fundamental relationship that lies at their core will. Not only could the modes of interaction be altered, but the parties involved will be different as well. As AI is increasingly being implemented, the SDM dual relationship should be re-drawn as a triad, involving the patient, the doctor, and the AI. The introduction of AI shifts the medical relationship paradigm to a new form of SDM that is shared between AI, doctors, and patients.

The new triadic SDM relationship should ensure good AI-doctor-patient communication. This could be attained on the hand, by ensuring that doctors have the competence to understand and evaluate AI's outputs while bearing in mind its limitations. On the other hand, patients should be informed of AI's involvement to allow them to better participate in the shared decision-making process. Therefore, decision-making is truly shared when both doctors and patients are in a position to contribute, each with their unique expertise (would this be medical knowledge, contextual clues, empathy, or personal values and preferences), to the final decision, even when an AI is involved.

Including both doctors and patients in the AI decision-making should guarantee that patients' values and preferences are considered, thus preserving their autonomy, and that doctors' professional autonomy is safeguarded. Similar collaborative relationship allows for AI, doctors, and patients to join forces. Eventually, this collaboration could result in better care.

References

1. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digit Med. 2020 Feb 6;3(1):1–10.

2. Sloane EB, J. Silva R. Artificial intelligence in medical devices and clinical decision support systems. Clin Eng Handb. 2020;556–68.

3. Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. 'Brilliant AI Doctor' in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment. Proc 2021 CHI Conf Hum Factors Comput Syst. 2021 May 6;1–18.

4. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care [Internet]. Balogh EP, Miller BT, Ball JR, editors. Washington (DC): National Academies Press (US); 2015 [cited 2023 Sep 6]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK338596/

5. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics. 2021 Dec 1;47(12):e3–e3.

6. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan;28(1):31–8.

7. Topol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. First Edition. New York: Basic Books; 2019. 378 p.

8. Jauhar S. When Doctors Need to Lie. The New York Times [Internet]. 2014 Feb 22 [cited 2021 Apr 21]; Available from: https://www.nytimes.com/2014/02/23/opinion/sunday/when-doctors-need-to-lie.html

9. Beauchamp TL, Childress JF. Principles of biomedical ethics. New York: Oxford University Press; 1979.

10. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics. 2019;45:156–60.

11. Stiggelbout AM, Van der Weijden T, De Wit MPT, Frosch D, Légaré F, Montori VM, et al. Shared decision making: really putting patients at the centre of healthcare. BMJ. 2012 Jan 27;344:e256.

12. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychother Theory Res Pract. 1979;16(3):252–60.

13. Godolphin W. Shared decision-making. Healthc Q Tor Ont. 2009;12 Spec No Patient:e186-190.

14. Appelbaum PS. Clinical practice. Assessment of patients' competence to consent to treatment. N Engl J Med. 2007 Nov 1;357(18):1834–40.

15. McKinstry B. Paternalism and the doctor-patient relationship in general practice. Br J Gen Pract. 1992 Aug;42(361):340–2.

16. Ha JF, Longnecker N. Doctor-patient communication: a review. Ochsner J. 2010;10(1):38–43.

17. Chandra S, Mohammadnezhad M, Ward P. Trust and Communication in a Doctor-Patient Relationship: A Literature Review. J Healthc Commun [Internet]. 2018 Jul 19 [cited 2021 Apr 21];3(3). Available from: https://healthcare-communications.imedpub.com/abstract/trust-and-communication-in-a-doctorrnpatient-relationship-a-literature-review-23072.html

18. Taddeo M, Floridi L. How AI can be a force for good. Science. 2018 Aug 24;361(6404):751–2.

19. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. JAMA. 2018 Dec 4;320(21):2199.

20. Wilson CB. Physician Autonomy Essential to Patient Care [Internet]. 2013 [cited 2022 Oct 19]. Available from: https://www.wma.net/blog-post/physician-autonomy-essential-to-patient-care/

21. McAndrew S. Internal morality of medicine and physician autonomy. J Med Ethics. 2019 Mar 1;45(3):198–203.

22. Emanuel EJ, Pearson SD. Physician autonomy and health care reform. JAMA. 2012 Jan 25;307(4):367–8.

23. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46:205–2011.

24. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J Am Med Inform Assoc JAMIA. 2020 Apr 1;27(4):592–600.

25. Páez A. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). Minds Mach. 2019 Sep 1;29(3):441–59.

26. Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. Digit Health. 2022 Jan 1;8:20552076221074488.

27. Birch J, Creel KA, Jha AK, Plutynski A. Clinical decisions using AI must consider patient values. Nat Med. 2022 Jan 31;1–3.

28. Farr C. CNBC. 2017 [cited 2022 Apr 21]. Here's why one tech investor thinks some doctors will be 'obsolete' in five years. Available from: https://www.cnbc.com/2017/04/07/vinod-khosla-radiologists-obsolete-five-years.html

29. Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. Eur J Intern Med. 2018 Feb;48:e13–4.

30. Liu X, Keane PA, Denniston AK. Time to regenerate: the doctor in the age of artificial intelligence. J R Soc Med. 2018 Apr;111(4):113–6.

31. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. In: Ethics, Governance, and Policies in Artificial Intelligence. Springer; 2021. p. 19–39.

32. Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. Npj Digit Med. 2019 Nov 18;2(1):1–10.

33. Cohen IG, Graver H. A Doctor's Touch: What Big Data in Health Care Can Teach Us About Predictive Policing [Internet]. Rochester, NY: Social Science Research Network; 2019 Aug [cited 2021 Jul 2]. Report No.: ID 3432095. Available from: https://papers.ssrn.com/abstract=3432095

34. Quintarelli S, Corea F, Fossa F, Loreggia A, Sapienza S. AI: profili etici. Una prospettiva etica sull'Intelligenza Artificiale: princìpi, diritti e raccomandazioni. BioLaw J - Riv BioDiritto. 2019 Nov 27;18(3):183–204.

35. Lorenzini G, Shaw DM, Arbelaez Ossa L, Elger BS. Machine learning applications in healthcare and the role of informed consent: Ethical and practical considerations. Clin Ethics. 2022 Apr 24;1–6.

Chapter 4

Machine Learning Applications in Healthcare and the Role of Informed Consent: Ethical and Practical Considerations

Citation: Lorenzini G, Shaw DM, Arbelaez Ossa L, Elger BS. Machine learning applications in healthcare and the role of informed consent: Ethical and practical considerations. *Clinical Ethics*. 2022;0(0). doi:10.1177/14777509221094476.

Machine Learning Applications in Healthcare and the Role of Informed Consent: Ethical and Practical Considerations

Abstract

Informed consent is at the core of the clinical relationship. With the introduction of machine learning (ML) in healthcare, the role of informed consent is challenged. This paper addresses the issue of whether patients must be informed about medical ML applications and asked for consent. It aims to expose the discrepancy between ethical and practical considerations, while arguing that this polarization is a false dichotomy: in reality, ethics is applied to specific contexts and situations. Bridging this gap and considering the whole picture is essential for advancing the debate. In the light of the possible future developments of the situation and the technologies, as well as the benefits that informed consent for ML can bring to shared decision-making, the present analysis concludes that it is necessary to prepare the ground for a future requirement of informed consent for medical ML.

Keywords

Ethics, Healthcare, Informed Consent, Machine Learning, Shared Decision-Making, Transparency.

Introduction

Informed consent is almost a novelty in medical practice. While the earliest form of medicine can be dated up to 2500 years ago, with the inception of the Hippocratic Oath(1), the first mention of informed consent was in 1957(2). Another important milestone in the history of informed consent is the Declaration of Helsinki, originally adopted in 1964. Although the Declaration focuses on medical research rather than clinical practice, it has been pivotal for disseminating the concept of informed consent in healthcare. Several controversial human experiments and clinical practices have since brought the debate about informed consent to a wider public. In response, bioethics became an increasingly relevant discipline and various approaches to navigate challenging ethical dilemmas have been proposed. One of such approaches are the four bioethical principles enunciated by Beauchamp and Childress(3): respect for autonomy, beneficence, non-maleficence, and justice. These principles became crucial in the subsequent bioethical considerations. Although this paper does not intend to adopt Beauchamp and Childress' approach to address the current issue, it aims to highlight the special role of the principle of respect for autonomy when it comes to reflecting on informed consent. Autonomy is closely linked with informed consent, which aims to protect and promote patients' autonomy. While the four principles are intended to be equally important, particular emphasis is given to the principle of autonomy in modern healthcare.

It is necessary to elucidate further the concept of informed consent to better grasp its centrality in the contemporary bioethical debate. Informed consent is a process required for clinical practice and research with human subjects. The practitioner informs the patient, in a simple and understandable language, about the procedure and associated risks and benefits, possible alternatives, prognosis, and consequences of each clinical decision(4–6). It is difficult to define the details of what ought to be disclosed; generally, clinicians need to disclose everything that a reasonable patient would want and need to know to understand their clinical situation and consequentially make an informed decision(5,7–9). However, this rule is very broad – what matters varies depending on the case – and can still result in paternalism since it is the clinician who decides which information is indispensable(10). This highlights the asymmetry of power and information between doctors and their patients. It is therefore necessary to carefully assess and evaluate informed consent in order to preserve patients' autonomy, self-determination and inviolability(11). Simply signing a form does not guarantee that informed consent is valid(12,13); instead, it is necessary to ensure that information is understood and that patients can participate in the decision-making process(14). To ensure this, sensitive and effective doctor-patient communication is crucial.

Informed consent has two major functions in the clinical context. First, it waives ethical and legal norms prohibiting invasive interventions, ensuring that patients understand and agree to the mentioned intervention, thus seeking to avoid patients' abuse and manipulation. At the same time, informed consent fosters *shared decision-making* (SDM). SDM can be defined as a mutual agreement

between doctors and patients that takes into account both doctors' expertise and patients' preferences(6,15). Regarding the first function, patients consent to practices that would otherwise be prohibited and would result in the prosecution of the doctor: consent is not required for actions that do not violate ethical or legal norms, and are therefore already permissible(10). For example, clinicians are not expected to seek consent when asking patients how they are feeling, instead, they should do so when administering drugs or performing surgery because the ethical and legal norms regarding the inviolability of the human body and personal freedom may otherwise be infringed(10). Contemporarily, informed consent is used to reach decisions shared between the doctor and the patient. According to the SDM paradigm, both parties are experts: doctors are experts in medicine, while patients are experts in their values(16). The common ground for SDM is informed consent: the doctor discloses what is deemed as necessary information, and answers the patient's questions honestly. At the same time, patients disclose information regarding their preferences and their values. The goal is to allow and motivate patients to be more involved in their own healthcare(17). It is commonly argued that SDM promotes a more ethical medical practice while improving the quality of care; therefore, it can be considered as the ethically appropriate paradigm(18–20).

After this brief and non-exhaustive elucidation of informed consent, it is time to introduce the second main theme of this paper: *machine learning* (ML). ML is a type of artificial intelligence that is able to learn from data and become more accurate, improving its performance(21,22). In healthcare, ML can be used as a *clinical decision support system* (CDSS) for its predictive capabilities; it provides recommendations and diagnoses for a wide range of situations. However, ML applications in healthcare generate concern because of the problem of the black box: although developers understand the process by which ML generates new models, the models themselves are inscrutable to humans, hence they are black boxes(23). Black boxes generate concern since they can prevent causal insights into ML's outputs. In the worst-case scenario, doctors are confronted with ML's recommendations with which they do not agree and that they do not understand, and cannot understand unless the ML provides any sort of justification or reason for its output.

The introduction of ML in clinical practice challenges the boundaries of informed consent and raises a panoply of questions. First of all, are doctors required to disclose to patients that their decisions are supported by ML? Should they invest their time explaining to patients information that they, in the first place, may not completely understand? Or is ML simply another tool amongst the others that doctors are not required to disclose? What would be the purpose of disclosing ML usage? The present discussion aims to present the discrepancy between many of the arguments found in the literature: often a more theoretical or more practical stance is inadvertently taken and the conclusions largely rely on the chosen angle. By bringing to light this tension between practical and ethical considerations, the goal is to position the issue in the frame of possible future developments, while attempting to bridge the gap.

Ethical considerations

Transparency can be conceived as a condition that often endorses or enables ethical practices(24), such as informed consent. In this paper, transparency is intended as doctors honestly disclosing information to their patients. Accordingly, transparency is necessary, even if not sufficient, to achieve informed consent. This is valid in clinical practice as well: it is generally thought that doctors who are transparent with their patients have more ethical conduct than the ones who are not. Like almost everything in ethics, transparency needs to be balanced: there are some cases in which clinicians may find transparency a constraint and disclosure would bring little or even no good. Therefore, one should ask whether clinicians should be required to be transparent with patients about ML usage.

One preliminary question is the reason for disclosing this information to patients(25). The ultimate goal of the disclosure can justify the expenditure of resources for informing patients about medical ML. Since transparency is positively linked with informed consent, which in turn is linked with SDM, it is possible to conclude that disclosing ML usage to patients is ethically recommended when it promotes SDM. This is because SDM is considered to be the ethically appropriate paradigm for the doctor-patient relationship(18). Being transparent about ML employment can be beneficial for SDM to take place while ensuring trust and setting up a solid base for communication to take place. Disclosing all relevant information is essential for SDM to succeed, and ML outcomes can be relevant for the role they play in clinical assessment. Interviews with patients on SDM have revealed that "shared decision-making is hard because [doctors] have so much more knowledge. So it can't be totally shared unless we are totally informed"(19). To facilitate patients' involvement in the SDM process, it is therefore extremely important to provide them with complete information. Although not all patients intend to actively participate in the decision-making process, and some prefer to leave the final decision to doctors, nearly all patients wish to be informed, given options, and asked for their preferences(19,26). Ultimately, when faced with the question of whether to disclose or not medical ML usage, one should try to understand what is the reason for disclosing. In cases where disclosing can empower patients and foster better communication to reach a shared decision, it is ethically advisable to do so. If transparency about ML leads to patients being better informed and hence better positioned to participate in the decision-making process, ethical considerations favour the extension of informed consent requirements to ML.

Unfortunately, it is not always so simple, but before coming to the objections let us consider another ethical argument in favour of disclosing ML to patients. ML employed as CDSS influences clinicians' decisions to the point where the extent to which it influences their judgement is debatable(27,28). Indeed, ML suggestions may prompt practitioners more than they are aware(28). This influence generates concern, especially regarding human control and autonomy. When ML is used to support doctors' decisions, some of the decision-making power is ceded to the algorithm(29). It is therefore of the foremost importance to balance human decision-making with ML-led decisions to avoid

arbitrary decision-making. Accordingly, most experts argue for maintaining human agency, responsibility, control, and oversight over ML systems(27,29–31). Continuous supervision, overview, and control by human agents shall ensure that ML is not a stand-alone system: it is rarely[2] the case that ML decisions are not considered by a clinician before they are accepted or refused. Hence the final decision can be conceived as shared between ML and doctors. On the other hand, disclosing ML usages to patients and discussing these suggestions with them could augment the human factor in the decision-making process, thus guaranteeing a larger amount of human scrutiny. Moreover, this would enable patients to participate in the decision-making process, hence making the final decision shared between ML, doctors, and patients. Although this tactic alone cannot preserve human control and autonomy, it could be part of a larger strategy for balancing human and ML decision-making power.

The principle of respect for autonomy and informed consent are closely linked: one of the aims of informed consent is to help protect and promote patients' autonomy. For example, patients' autonomy is preserved when they are enabled to decide whether to undergo a certain medical treatment or not. Therefore, the opt-out option is an essential part of respecting patients' autonomy; at the same time, patients need to be informed in order to opt-out. This is true also with medical ML: being unaware of its use, patients are not given the choice of whether to opt-in or opt-out. Their personal health data is being fed to ML without seeking their consent and hence without respecting their preferences. This can have detrimental consequences on their autonomy. Introducing informed consent for ML usage in healthcare allows patients to decide on their health data, which is perceived as one of the most sensitive types of data(32).

These three arguments in favour of introducing informed consent for medical ML have in common a theoretical stance, meaning that they do not acknowledge the practical difficulties and obstacles involved in implementing this requirement. The present limitation might be caused by the complexity of the issue: an interdisciplinary appraisal is needed to appraise the legal implications, the ethical consequences, and the reality of clinical practice. However, interdisciplinary perspectives can be difficult to attain(33,34).

Nonetheless, not all theoretical arguments are in favour of extending the informed consent requirement to medical ML. For example, the debate over the ontology of ML – whether ML is a tool, like many others employed in everyday medical practice, or if is it something different – can offer a ground for denying the need for informed consent. If ML was simply another tool that clinicians use during their decision-making process, it could be argued that not informing patients does not conflict with ethical and legal requirements. Since it is generally not required for doctors to disclose how tools

[2] There are some exceptions, such as the *IDx-DR*, the first autonomous FDA-authorized AI diagnostic system for diabetic retinopathy and macular enema.

influence their decisions, many do not see any logic in imposing this type of obligation for ML. Therefore, failing to inform patients about ML uses would not violate the doctrine of informed consent(35).

However, the idea that ML is a mere tool does not come without controversies. A major argument regarding ML not being a traditional tool is the problem of the black box: the inherent opaqueness of its algorithms may differentiate it from other tools, whose mechanisms are usually transparent. In contrast with other tools, ML is inaccessible to doctors' scrutiny, and its causal insights remain hidden. For this reason, some experts fear that doctors may become less the deciders and more the implementers of ML outputs(18,27). Other doubts about ML classification as a tool are raised by its suggestions, which can influence clinicians' decisions more than they think (similarly to a nudge rather than a simple suggestion)(28). The extensiveness of ML's influence could differentiate it from other tools, which generally do not present this feature. Lastly, the idea that ML does not violate ethical and legal requirements can be questioned as well: ML entails many risks as it can make widely incorrect evaluations and its recommendations tainted by discriminatory bias(31,36,37).

Suggesting that ML does not violate any fundamental ethical or legal requirement may be naïve considering the novelty of its applications: the consequences of its widespread application in healthcare still have to be identified and assessed. The risks posed by ML must not be overlooked; instead, there is the need to conduct further research – for example, some CDSS are not well integrated into clinical practice, require doctors to invest their limited time in inputting data into the system, and present too many alerts that eventually doctors ignore(38); ML can lead to benefits as well as burdens for the healthcare sector. In the future, we might be more aware of the risks it entails and therefore the stance towards the necessity of informed consent for medical ML may differ accordingly. As of now, the heated debate on ML ontology remains open.

Practical considerations

The previous ethical analyses should be considered in light of the available resources. Transparency about ML applications may not always be the best option, and doctors who choose not to do so must not be immediately judged as unethical. There can be a variety of reasons why doctors do not disclose ML to patients. First of all, doctors already have a tight schedule and it is sometimes challenging to take the time to discuss ML usage and its suggestions with patients. If the process of SDM were to be made more complex, in some circumstances this could put at risk patients' health rather than benefiting it. There is already a deficit in providing the necessary information to patients(5). This is in part due to the already-existing barriers clinicians encounter when communicating with patients, which include time constraints(39). Many practitioners maintain that obtaining informed consent every time ML is used is very time-consuming and could take up time from the conversation about care(31). Doctor-patient communication must prioritize health-related discussions; when informing the patients

about ML derails the dialogue from this priority, disclosing could potentially be harmful(31). Nevertheless, time-constraint barriers could be overcome with the automation of parts of doctors' workload with ML, hence reducing the burden of routine tasks(40). However relevant, the time-constraint objection is not categorical since ML itself could invalidate it. As ML takes over clinical routine tasks, doctors can have more time to invest in communicating with patients. Eventually, the time-constraints preventing doctors from discussing ML usage with patients could be resolved by ML itself.

It must not be assumed that all clinicians have a meaningful understanding of ML, hence it would put them in a difficult position if they had to disclose and answer patients' questions about ML. In addition to doctors who do not understand how ML works, some patients are technologically illiterate, sometimes to the extent that any comprehension of ML functions is denied to them. The lack of understanding of ML can represent a difficulty in the practical implementation of informed consent. On the other hand, doctors can be trained and educated, allowing them to learn about ML to the extent necessary to safely use it and, at the same time, to be able to discuss it with patients. It is probably not necessary for doctors to have exhaustive knowledge of ML – for instance, knowledge of ML programming or its mathematical constructs. In reality, doctors need enough information to assess the technology and comprehend its impact on patients. With this information, doctors may be better positioned to address patients' issues and find the best-tailored solution for them. At the same time, patients can be taught as well. Moreover, it can be reasonably expected that in the future humans will be more and more acquainted with this kind of technology, hence gaining better insights and some general knowledge of its functioning. Increased familiarity with ML could be pivotal for the success of any eventual future requirement for informed consent.

Bridging the gap

On the one hand, many ethical arguments favour the implementation of informed consent for ML applications in healthcare. On the other hand, the majority of the practical considerations focus on the difficulties of imposing this requirement for everyday clinical practice. However, there is an exigency – as well as a possibility – to find common ground to further advance the debate. The first step is to bear in mind that this polarization between ethical and practical considerations is not a real distinction: In clinical practice, ethics is applied to specific contexts and situations, hence considering practical barriers and enablers. Although the debate on ML informed consent seems to be split into these two distinct viewpoints, in reality they overlap and cooperate.

A second reason for surmounting the polarization is that some theoretical arguments argue against the implementation of informed consent for ML. At the same time, there are practical considerations in favour of an informed consent requirement for ML. As previously discussed, disclosing ML to patients may divert time from the conversation about care, which could result in lower

quality of care and diminished SDM for patients. A doctor communicating about ML could mean that, although the patient has the possibility to opt in or out, their doctor may not have time to discuss other important information, such as alternative treatment. As a consequence, the patients would not be able to participate fully in the treatment decision, resulting in non-ideal clinical practice. At the same time, transparency about ML could foster patients' trust and ameliorate the doctor-patient relationship. This may correlate with better SDM and consequently with improved quality of care(9,16).

The previous examples blur the boundaries of the distinction between a theoretical and a practical stance. Ethical arguments are based on practical circumstances and defined contexts, while practical reflections are informed by ethical principles. A stark division is inconceivable and unrealistic. Of course, differentiations can be – and sometimes must be –made for the sake of the discussion: it can be easier to reason with categories. However, it is crucial to remember that in reality ethics and practice are strictly intertwined, and such a polarization between ethical and practical considerations is not possible.

Despite the lack of a legal requirement to disclose ML usage to patients and to extend the informed consent process to its applications, in the present analysis we argue that it is ethically advisable to implement informed consent for medical ML in a way that further promotes and fosters SDM. Confronted with the ethical dilemma of whether to disclose ML to patients, being transparent seems the better option, given that the SDM paradigm is widespread and well-respected. At the same time, it would not be fair to simply deem doctors that choose not to disclose ML as unethical. As there are currently many obstacles and uncertainties to the implementation of ML informed consent, not informing patients about it can be excused. However, this does not mean that one should not aim to change the situation and orient towards a better alternative – namely, informed consent for ML applications in healthcare. As has been noted previously, the objections to ML informed consent are not categorical, but rather contingent, meaning that solutions can be found. It is possible to find ways to overcome the impracticability of a requirement for ML informed consent, sometimes with the help of ML itself. The moral duty would not be that of imposing a requirement for informed consent, rather, to prepare the ground for this requirement to flourish. The current situation should be better adapted to the new context, where technologies such as ML are integrated into clinical practice.

Conclusion

The discussion about informed consent for medical ML is very recent and it is linked to all the uncertainties surrounding ML uses in clinical contexts. Some arguments maintain that informing patients about ML usage in healthcare is impractical and there would be little benefit in disclosing at present. This idea is in line with the practical considerations analysed in this paper, which expose the unfeasibility of informed consent for ML. However, these objections are contingent; hence they show the problems that should be addressed, rather than constituting structural impossibilities. Concluding

that informed consent for medical ML is not necessary suggests that the whole picture has not been considered. The ethical value of disclosing ML to patients has been acknowledged: disclosing is ethically advisable when it empowers the patients and enables them to participate in the management of their care. This is to say, informed consent for medical ML is ethically desirable when it supports SDM.

The discrepancy between ethical and practical considerations highlights an already existing tension between ethics and practice. Recognizing and comprehending the difficulties of introducing informed consent is indispensable to overcome them. Indeed, these barriers are mostly non-structural, they do not constitute a reason for abandoning the idea of informed consent for medical ML. This is especially true when the ethical analysis illustrates the significance of doing so. It should be recognized that informed consent for ML is a requirement that may need some time before being practically feasible, and hence implemented. It might be that in future its necessity will be more evident, as the practical counter-arguments are solved and ML risks are better comprehended. Therefore, the present conclusion invites further exploration of this issue, in order to bridge and bring together ethical and practical considerations. It is now the time to prepare the healthcare sector for a future implementation of informed consent for ML applications, without compromising patient care or SDM.

References

1. Jouanna J. Hippocrates [Internet]. Baltimore, Md. : John Hopkins University Press; 1999 [cited 2021 Aug 2]. 546 p. Available from: http://archive.org/details/hippocrates0000joua

2. Hick C, Corbellini G. Consenso informato in 'Enciclopedia della Scienza e della Tecnica' [Internet]. 2007 [cited 2021 Jul 1]. Available from: https://www.treccani.it/enciclopedia/consenso-informato_(Enciclopedia-della-Scienza-e-della-Tecnica)

3. Beauchamp TL, Childress JF. Principles of biomedical ethics. New York: Oxford University Press; 1979.

4. Schiff D, Borenstein J. How Should Clinicians Communicate With Patients About the Roles of Artificially Intelligent Team Members? AMA J Ethics. 2019 Feb 1;21(2):E138-145.

5. Shah P, Thornton I, Turrin D, Hipskind JE. Informed Consent. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 [cited 2021 Aug 3]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK430827/

6. Slim K, Bazin JE. From informed consent to shared decision-making in surgery. J Visc Surg. 2019 Jun 1;156(3):181–4.

7. Palmboom GG, Willems DL, Janssen NBAT, de Haes JCJM. Doctor's views on disclosing or withholding information on low risks of complication. J Med Ethics. 2007 Feb;33(2):67–70.

8. Murray B. Informed Consent: What Must a Physician Disclose to a Patient? AMA J Ethics. 2012 Jul 1;14(7):563–6.

9. Coulter A, Collins A. Making shared decision-making a reality: no decision about me, without me. London: King's Fund; 2011. 45 p.

10. Manson NC, ONeill O. Rethinking Informed Consent in Bioethics [Internet]. Cambridge: Cambridge University Press; 2007 [cited 2021 Jul 1]. Available from: http://ebooks.cambridge.org/ref/id/CBO9780511814600

11. Hall DE, Prochazka AV, Fink AS. Informed consent for clinical treatment. CMAJ Can Med Assoc J. 2012 Mar 20;184(5):533–40.

12. Morton R. The Doctors Company. 2020 [cited 2022 Jan 11]. Informed Consent: Substance and Signature. Available from: https://www.thedoctors.com/articles/informed-consent-substance-and-signature/

13. Weiner S. AAMC. 2019 [cited 2022 Jan 11]. What "informed consent" really means. Available from: https://www.aamc.org/news-insights/what-informed-consent-really-means

14. Millum J, Bromwich D. Informed Consent: What Must Be Disclosed and What Must Be Understood? Am J Bioeth. 2021 May 4;21(5):46–58.

15. Frosch DL, Kaplan RM. Shared decision making in clinical medicine: past research and future directions. Am J Prev Med. 1999 Nov 1;17(4):285–94.

16. Godolphin W. Shared decision-making. Healthc Q Tor Ont. 2009;12 Spec No Patient:e186-190.

17. Stiggelbout AM, Van der Weijden T, De Wit MPT, Frosch D, Légaré F, Montori VM, et al. Shared decision making: really putting patients at the centre of healthcare. BMJ. 2012 Jan 27;344:e256.

18. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics. 2019;45:156–60.

19. Say RE, Thomson R. The importance of patient preferences in treatment decisions—challenges for doctors. The BMJ. 2003 Sep 6;327(7414):542–5.

20. Staren D. HealthValueHub [Internet]. 2019 [cited 2021 Sep 16]. Available from: https://www.healthcarevaluehub.org/advocate-resources/publications/consumer-benefits-patient-shared-decision-making

21. Burns E. SearchEnterpriseAI. 2021 [cited 2021 Jul 1]. In-Depth Guide to Machine Learning in the Enterprise. Available from: https://searchenterpriseai.techtarget.com/In-depth-guide-to-machine-learning-in-the-enterprise

22. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In 2020. p. 295–336.

23. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Cent Rep. 2019 Jan;49(1):15–21.

24. Turilli M, Floridi L. The ethics of information transparency. Ethics Inf Technol. 2009 Jun 1;11(2):105–12.

25. Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harv J Law Technol [Internet]. 2018 Mar 21 [cited 2021 Jul 12];31(2). Available from: http://arxiv.org/abs/1711.00399

26. Levinson W, Kao A, Kuby A, Thisted RA. Not All Patients Want to Participate in Decision Making. J Gen Intern Med. 2005 Jun;20(6):531–5.

27. Cohen IG, Graver H. A Doctor's Touch: What Big Data in Health Care Can Teach Us About Predictive Policing [Internet]. Rochester, NY: Social Science Research Network; 2019 Aug [cited 2021 Jul 2]. Report No.: ID 3432095. Available from: https://papers.ssrn.com/abstract=3432095

28. Taddeo M, Floridi L. How AI can be a force for good. Science. 2018 Aug 24;361(6404):751–2.

29. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach. 2018;28(4):689–707.

30. AI HLEG. Ethics guidelines for trustworthy AI | Shaping Europe's digital future [Internet]. European Commission; 2019 Apr [cited 2021 Jul 6] p. 41. Report No.: B-1049. Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

31. Kent C. Medical Device Network. 2020 [cited 2021 Jul 8]. Artificial consent: should doctors be telling patients more about AI? - Verdict Medical Devices. Available from:

https://www.medicaldevice-network.com/features/artificial-consent-should-doctors-be-telling-patients-more-about-ai/

32. Wenger F, Jaquet-Chiffelle DO, Kleine N, Weber K, Morgan G, Gordijn B, et al. Canvas White Paper 3 – Attitudes and Opinions Regarding Cybersecurity [Internet]. Rochester, NY: Social Science Research Network; 2017 Oct [cited 2021 Sep 16]. Report No.: ID 3091920. Available from: https://papers.ssrn.com/abstract=3091920

33. Mudroch V. The future of interdisciplinarity: The case of Swiss universities. Stud High Educ. 1992 Jan;17(1):43–54.

34. Darbellay F. The Gift of Interdisciplinarity: Towards an Ability to Think across Disciplines. Int J Talent Dev Creat. 105AD Dec;3(2):201–11.

35. Cohen IG. Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? [Internet]. Rochester, NY: Social Science Research Network; 2020 May [cited 2021 Jul 6]. Report No.: ID 3529576. Available from: https://papers.ssrn.com/abstract=3529576

36. Goldhill O. Quartz. 2020 [cited 2021 Jul 8]. When AI in healthcare goes wrong, who is responsible? Available from: https://qz.com/1905712/when-ai-in-healthcare-goes-wrong-who-is-responsible-2/

37. Kent C. Medical Technology. 2020 [cited 2021 Jul 8]. A race to the bottom: how AI encodes racial discrimination within medicine. Available from: https://medical-technology.nridigital.com/medical_technology_sep20/ai_racial_discrimination_medicine

38. Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. 'Brilliant AI Doctor' in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment. Proc 2021 CHI Conf Hum Factors Comput Syst. 2021 May 6;1–18.

39. Légaré F, Ratté S, Gravel K, Graham ID. Barriers and facilitators to implementing shared decision-making in clinical practice: Update of a systematic review of health professionals' perceptions. Patient Educ Couns. 2008 Dec 1;73(3):526–35.

40. Academy of Medical Royal Colleges. Artificial Intelligence in Healthcare [Internet]. 2020. Available from: https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf

Chapter 5

It Takes a Pirate to Know One:

Ethical Hackers for Healthcare Cybersecurity

Citation: Lorenzini, G., Shaw, D.M. & Elger, B.S. It takes a pirate to know one: ethical hackers for healthcare cybersecurity. *BMC Med Ethics* **23**, 131 (2022). https://doi.org/10.1186/s12910-022-00872-y

It Takes a Pirate to Know One: Ethical Hackers for Healthcare Cybersecurity

Abstract

Healthcare cybersecurity is increasingly targeted by malicious hackers. This sector has many vulnerabilities and health data is very sensitive and valuable. Consequently, any damage caused by malicious intrusions is particularly alarming. The consequences of these attacks can be enormous and endanger patient care. Amongst the already-implemented cybersecurity measures and the ones that need to be further improved, this paper aims to demonstrate how penetration tests can greatly benefit healthcare cybersecurity. It is already proven that this approach has enforced cybersecurity in other sectors. However, it is not popular in healthcare since many prejudices still surround the hacking practice and there is a lack of education on hackers' categories and their ethics. The present analysis aims to comprehend what hacker ethics is and who ethical hackers are. Currently, hacker ethics has the status of personal ethics; however, to employ penetration testers in healthcare, it is recommended to draft an official code of ethics, comprising principles, standards, expectations, and best practices. Additionally, it is important to distinguish between malicious hackers and ethical hackers. Amongst the latter, penetration testers are only a sub-category. Acknowledging the subtle differences between ethical hackers and penetration testers allows to better understand why and how the latter can offer their services to healthcare facilities.

Keywords

Cybersecurity, Hacker Ethics, Health Data, Penetration Tests.

Background

Cybersecurity is a major concern in almost every context nowadays, and our reliance on interconnected technologies leaves companies and institutions extremely vulnerable to hackers' attacks. Recent attacks have made it clear that every system has some vulnerabilities, and it is simply a matter of time until some malicious hacker exploits them. Particularly in the healthcare context, cyber threats are becoming increasingly common and a growing concern (1–3). Healthcare has come, and is greatly encouraged, to rely on digital technologies, such as electronic health records (EHR), wearable devices, and artificial intelligence (AI) tools, which further augment vulnerabilities (2,4). Since the outbreak of the COVID-19 pandemic, cyberattacks on healthcare facilities have intensified and they have put additional strain on the already-overwhelmed healthcare industry (4). Amongst other examples, in March 2020 the Czech Brno University Hospital, a COVID-19 testing facility, was targeted by hackers, forcing its entire IT network to shut down and causing the cancellation of all surgeries; during the same month, the World Health Organization saw the creation of a spoof site that mimicked their own, aiming to steal employees' passwords (4–6). There are numerous other examples, however, many cyberattacks are undetected or unreported and only a minority of them are publicly disclosed (7). Accordingly, it is difficult to precisely assess the prevalence of these attacks and their consequences, as well as to intervene promptly. This underreporting conveys a false sense of security while failing to raise the necessary awareness to take protective measures. However, some jurisdictions have imposed obligations for notifying cyber incidents and data breaches; two examples are the Cyber Incident Reporting for Critical Infrastructure Act in the US and Article 33 of the European General Data Protection Regulation (GDPR) (8,9). Further awareness and cyber-hygiene measures are nonetheless needed (4).

Cyber threats are particularly severe in healthcare for two reasons: the vulnerabilities of this sector and the dramatic consequences that can result from their penetration. Indeed, cyberattacks can negatively affect public trust, damage critical equipment, and threaten human lives (10). When a cyberattack occurs, there is little room for negotiation without putting patients' care at risk (11). Moreover, it can be profitable to target the healthcare industry: financial gain can be significant as health data is very valuable (3,7).

This paper aims to show how penetration tests (pen-tests) conducted by ethical hackers can be beneficial for healthcare cybersecurity. It is already established that this approach has enforced cybersecurity in other sectors, where vulnerabilities have been located and the defence system has been reinforced. Pen-testers "have a valuable role to play in probing hardware, software or websites to look for weaknesses" (12). In addition to contributing a lot through pen-tests, they can help address the labour shortage affecting the sector (13). However, it seems that this service is not commonly provided to healthcare facilities. This could partly depend on the fears and prejudices towards the hacking practice, on the lack of a clear, and official, code of ethics for this profession, and on the limited financial resources that these facilities can devote to cybersecurity. The present analysis addresses the first

difficulty by illustrating how pen-tests are a serious service offered by respectable and reliable companies and professionals. The second difficulty is addressed by presenting hacker ethics: the description of this ethics may contribute to rise the awareness necessary to effectively collaborate with ethical hackers, and pen-testers in particular.

Special status of healthcare cybersecurity

A combination of factors renders healthcare particularly exposed to cyberattacks while being profitable for malicious hackers. In this section, both the vulnerabilities and the value of health data will be explored.

a) Health data

Health data are considered one of the most personal types of information by citizens, so unauthorized access is particularly alarming (14). Health data breaches can have serious consequences for individuals' privacy as they can cause stigma, discrimination, and embarrassment to patients. Furthermore, they can impact patients' jobs, insurance and economic status, and family relations (1). This is especially true in the case of well-known individuals, such as politicians and celebrities, or in the case of already vulnerable and stigmatized populations. Nonetheless, the negative consequences of health data breaches are potentially serious for every individual.

Since health data document intimate personal information that cannot be reset, unlike other types of personal sensitive information (e.g. credit cards), they are interesting and profitable for the black market (15). Often, these records contain enough information to open bank accounts, obtain loans, or acquire an identity document (7). Therefore, health records can be worth more than credit card information and they allow identity thieves to create convincing identities (16). When the records do not contain sufficient personal information, they can anyhow be valuable for the black market as they allow access to prescription drugs (10).

Health data are obviously vital also for healthcare: when hackers launch ransomware attacks and lock access to this data, the entire workflow is halted. Ransomware is a malicious program that encrypts the information stored on the servers, rendering them inaccessible to the staff. Usually, the encryption is followed by a ransom demand for the decryption keys (4). It is not possible to check a patient's blood type, surgeries are cancelled, and everything has to be noted by hand (17). One exemplary case is the University of Vermont (UVM) Medical Center, which was hit by a ransomware attack during the COVID-19 pandemic in October 2020, when an employee opened a phishing email (18). The attack caused the shutdown of all internet connections, precluding access to patients' EHR. Unable to communicate, they sent employees to buy walkie-talkies. For nearly a month, the UVM Medical Center could not use EHR and other digital tools. For days their staff could not access patients' appointments. Many surgeries were rescheduled and cancer patients were re-addressed to other facilities

for radiation treatment (19). When the system is finally accessible again, all the handwritten data noted during the time the system was inaccessible has to be reported back to the computer. This is a time-consuming activity and it can take months before returning to the pre-attack situation. Therefore, protecting health data is essential for ensuring patients' safety.

b) Vulnerabilities

The healthcare industry has many vulnerabilities that can be exploited by malicious hackers. Longstanding insufficient investment in IT is one first factor (10). For example, legacy software, such as Windows XP, is still frequently used although this operating system is no longer supported by security updates. This makes it easier for malicious hackers to exploit vulnerabilities (10). The situation is further exacerbated by the shortage of cybersecurity experts; in order to attract them, companies offer them exceedingly competitive salaries that healthcare organisations often cannot match (7,13,20). Prioritising patient care, the healthcare sector lacks the resources necessary to establish a solid cyber defence. Therefore, in several healthcare facilities cybersecurity has been neglected to various degrees.

A second factor accounting for healthcare vulnerability is the implementation of interconnected technologies that, while enabling remote and distributed access to care, constitute an opportunity for intrusion (21). The cybersecurity issue is one of the challenges posed by the introduction of new technologies in healthcare[3]. For example, a novel concern is the malicious intrusions into medical devices such as pacemakers and insulin pumps: as an ethical hacker demonstrated, it is theoretically possible to remotely hack a Medtronic insulin pump to deliver a lethal dose of insulin, with potentially disastrous consequences for the patient (22). While there are no reports of attacks on insulin pumps, in 2015 there has been a massive cyberattack targeting medical devices, known as MEDJACK (23). Unbeknownst to hospitals' staff, diagnostic equipment (MRI machines and CT scanners), therapeutic equipment (infusion pumps), and life support equipment (ventilators) were compromised (24).

Penetration tests

Cybersecurity experts design their techniques based on assumptions about malicious hackers' behaviour. However, without actual knowledge of the hacking practice and motivations, these assumptions often reveal unrealistic expectations and might overlook important factors (25). This is where penetration tests (often referred to as "pen-tests") can play a crucial role as they simulate malicious hackers' intrusions, thus locating the system's weak points. Pen-tests are authorized attempts

[3] AI could also be used to enforce healthcare cybersecurity by detecting suspicious activity within the facility network, hence addressing the difficulty human experts face in detecting intrusions. Indeed, it often takes weeks, if not months, for humans to detect breaches and contain the damage. Cybersecurity staff are often overwhelmed by threat alerts and report risks. AI could free humans from this burden by automatically checking all the alerts; although humans would still need to evaluate AI's warnings, these tools can render their jobs more efficient. However, monitoring the facility network might have privacy implications if employees are surveilled at their workplace. Therefore, cybersecurity AI should be carefully implemented to avoid undesirable consequences, such as privacy invasions.

to break into a system in the same way malicious hackers would do (26). To gain this knowledge, someone who thinks like a malicious hacker is needed. Even better, hackers themselves should conduct pen-tests, ethical hackers. Pen-tests can help healthcare facilities to be aware of weaknesses and to fix them before a malicious hacker finds them.

It is fundamental to understand that the pen-test is a common service provided by reputable companies. Pen-testers are therefore employed or have a contractual relationship with the company that has been hired. It is true that pen-testers often resort to the same techniques and tools as malicious hackers, which can sometimes be controversial (e.g. social engineering strategies, such as phishing and USB drops[4]) (27). But the goal is always to conduct realistic simulations to efficiently bolster cybersecurity (26). Although technically they would still break into systems, pen-testers would be under the obligations of a contract that explicitly establishes boundaries and prohibited practices. They must have written permission before conducting pen-tests, and during the assessment, they must stay within the scope of the employers' expectations, which must have been previously disclosed and discussed. It is also essential that they transparently communicate all the findings and provide a detailed transcript of their actions (28). Contemporarily, there is explicit consent on the part of the employer. The establishment of *ad hoc* legal contracts is a necessary step to support the employment of pen-testers in healthcare cybersecurity.

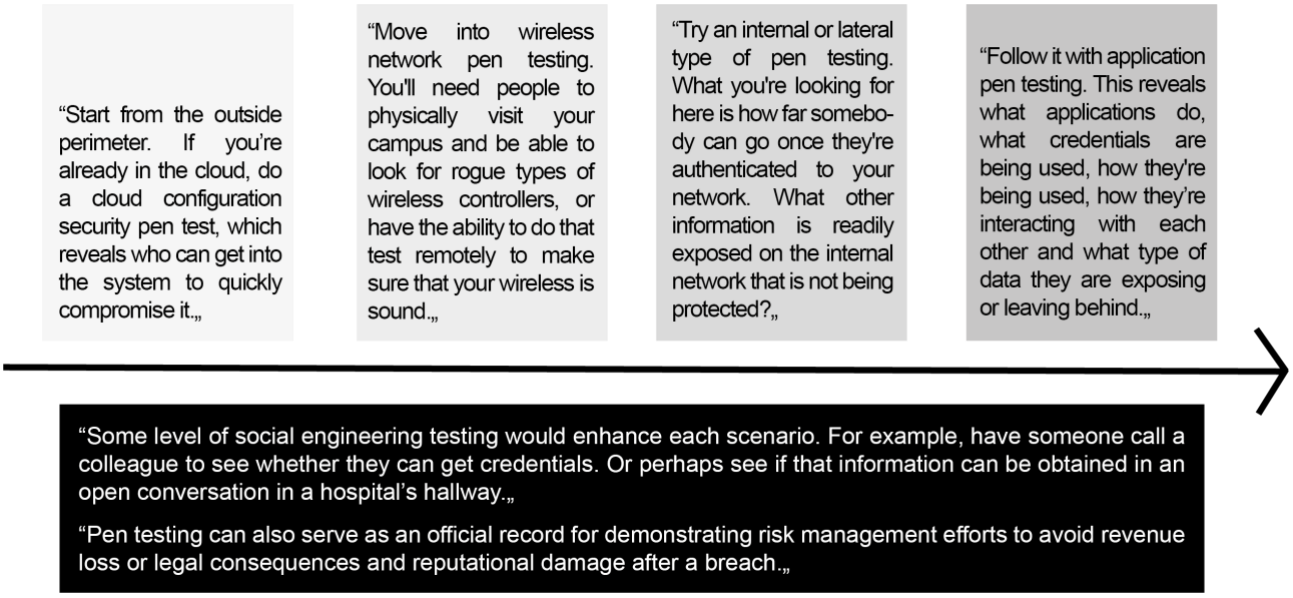


Figure 1: Example of a four-step approach pen-test for healthcare.Figure based on: Gregory, M. Tabletop Exercises and Penetration Testing: What Health Systems Need to Know, in HealthTech Magazine, 2021.

[4] Phising and USB drops are only two examples of the many social engigneering methods that pen-testers can use. Phising refers to emails sent with the intent to deceive the receiver; they usually include malicous links or requests for sensitive and private information. USB drops consist in leaving a USB stick in a common area (e.g. the hospital secretariat); this USBs typically contain malicous softwares that, when plugged into the hospital IT infrastructure, allows access to that system.

Ethical hacker certification exists and is becoming increasingly important (12). At the same time, the possibilities to get certified are also increasing: there are already many institutions that offer certification, both theory and experience-based. Additionally, many universities are offering ethical hacking courses (12). These certificates can be very convenient in the hiring process for pen-testers.

Hiring companies that provide pen-test services is not the same thing as allowing anonymous ethical hackers to try and hack into a healthcare facility. While rewarding and bug bounty programs can work perfectly fine for other sectors, in healthcare this may cause controversies. As previously seen, the healthcare sector is particularly vulnerable and its data is very sensitive. Collaborating with companies seems a better path as this allows to negotiate the details of the service provided and establish specific contracts; this also entails the employer's explicit consent to the pen-test. At the same time, the identity of pen-testers is known and the company can be held liable if damages occur.

There is therefore a difference between pen-testers and ethical hackers that is now necessary to elucidate. While these two terms are often used interchangeably, "ethical hacker" is rather an umbrella term that includes all hacking methodologies and techniques (29,30). It is correct to say that pen-testers are ethical hackers, however not all ethical hackers are pen-testers. Pen-testing is a very planned process that requires all necessary permissions; although it mimics real-life cyberattack scenarios, it is a helpful and non-harmful process (30). Pen-testers are certified professionals hired by companies under legal contracts, instead, this is not a requirement for ethical hackers that usually act on a voluntary basis and through bug bounty programs; these do not involve contractual relationships and informed consent procedures. The range of attack vectors and attack types for pen-testers is limited, whereas this limitation is not present when referring to ethical hacking. While usually adhering to high standards of behaviour (31), the lack of limitations on the available hacking techniques may sometimes cause collateral damage (e.g. computer downtime or data breaches).

Acknowledging the subtle difference between ethical hackers and pen-testers is important for employing them in healthcare cybersecurity. Since pen-testers still fall under the "hackers" category, understanding what hacker ethics is and who ethical hackers are can be relevant when deciding to hire their services.

Understanding hackers and their ethics

Hacking began in the 1970s and is defined as the "unauthorized intrusion into a computer system" (32). This definition includes both the practice of malicious hacking and ethical hacking. The difference lies in the hacker's intent: if the purpose of hacking is just for the challenge, the thrill, and finding (and reporting) leaks in the security, but without stealing money or disseminating data, then that hacker is called a "white hat", or an ethical hacker. Note that hacking for the thrill or challenge alone

would not constitute ethical hacking in the absence of reporting. There is a grey area where someone is neither an ethical hacker nor a malicious one; they are what can be called a "grey hat": morally ambiguous hackers that do not fully adhere to ethical hackers' principles but whose actions are not fundamentally guided by malicious intentions (28,33). Instead, if the aim is the hacker's financial gain and disruption, we are faced with a "black hat", namely a malicious hacker. Nonetheless, it is not always so simple to differentiate and the same hacker can sometimes act in both ways or later "convert" to ethical hacking. Some argue that it is wrong to fit hackers into a moral binary, in which they are either heroes or villains (34).

Besides the difficulty of categorizing every hacker with certainty, there is the issue that even hacking "for the good" can be punishable. This causes many ethical hackers to avoid reporting vulnerabilities for fear of legal repercussions[5] (35). It also contributes to hackers' willingness to work in the shadows and consequently creates a distorted perception of hacking practices. Media coverage that portrays them in dark hoodies in dark rooms at night further contributes to this misconception. The term "hacker" itself presents negative and pejorative connotations, stigmatizing a widely varied group (28,36). In an effort to collaborate with ethical hackers (and to professionalize them as pen-testers, particularly in the healthcare sector), a first fundamental step is to acknowledge the prejudices and narratives surrounding their practice. A second step would entail recognizing the existence of different categories of hackers. Lastly, better understanding hacker ethics could address some controversies and concerns.

Considering hacker ethics is useful for better understanding ethical hackers and their values. Sooner or later, hackers are confronted with ethics. Even if hacking is not primarily an ethical issue, most hackers come to a point where they have to face some ethical questions, hence, there is a certain connection between hacking and ethics (37). Hacker ethics is a type of personal ethics, therefore every hacker has a unique understanding of its values. In fact, some claim that "there is no hacker ethics. Everyone has his own" (34). Despite the lack of unitary hacker ethics, the many hacker codes present numerous similarities (38,39). Indeed, they all somehow share liberalistic ideals. For example, they endorse open-source projects and are very privacy-aware. What differs is how they interpret and defend

[5] Some companies (e.g. Google, Microsoft, Facebook) and governments (e.g. in the Netherlands) already adopted approaches and legislation that allow ethical hackers to report vulnerabilities, while protecting them form legal repercussion; this protection is crucial to continuing collaboration with ethical hackers to bolster cybersecurity. Lacking this legal protection, many vulnerabilities could go unreported out of fear of legal repurcussions, hence leaving open a possibility for malicous hackers to exploit said vulnerability. The discussion pertaining to the boundaries and modalities of the practice is still ongoing; it appears that as long as no harm is done to the system and the data, nothing is leaked, and the vulnerability is appropriately reported, the ethical hacker should not be prosecuted. It nonetheless remains difficult to assess what "appropriately reported" means and whether a copy of the data was made. In the meanwhile, new regulations are being issued (e.g. in May 2022 the US Department of Justice ruled that ethical hackers will no longer be prosecuted under the Computer Fraud and Abuse Act) that aim to tackle this delicate situation.

these ideals. This difference can be well-illustrated by the positive and the negative understanding of "freedom".

In its positive connotation, freedom invokes free and open access to information with the pedagogical goal of equally allowing humans to educate themselves (34). What matters is to advance human knowledge, make sure that it is available to everyone, and encourage cooperation (37). From this perspective, mechanisms to privatize and monetize information and software constitute a barrier and are considered unethical (34). Copyright laws corrupt freedom since information is not ownable property. Sharing information would then be a moral imperative. However, this does not call for the elimination of all barriers: it is important to maintain and enforce privacy measures. This is a freedom that values learning, community, sharing, and equal opportunities. It aims to advance human knowledge and bridge the current information gap. For this reason, the focus is on the liberalization of knowledge and open-source software, rather than on the notion of privacy, although deemed extremely important.

The negative sense of freedom stands close to anarchistic ideals and can be intended as "freedom from everything". It greatly values privacy and often leads to acts of civil disobedience to protect it (34). It is antagonistic to institutionalization and surveillance measures. The focus is on self-determination and non-interference of others. Its primary values are individual autonomy, self-reliance, and, of course, individual privacy. While positive freedom emphasizes community welfare, negative freedom is focused on individuality.

Hacker ethics is neither dichotomic nor unitary; it entails a different, and sometimes contradictory, understanding of values. However, as has been previously observed, there are commonalities and similarities. It is noteworthy that it revolves around two values: freedom and privacy. Although distinctively interpreted, they constitute the core of hacker ethics. Hackers' actions often emanate from different interpretations of these two values. However, adherence to hacker ethics does not imply that their actions would be deemed morally good by society: some hackers may advocate their ethics by stealing confidential information and disseminating it[6]. Although black hat hackers' actions are generally unquestionably unethical, with grey hat hackers the morality of some actions can be debatable (for example, grey hat hackers that report vulnerabilities often threaten the owner of the hacked system to publicly reveal it, hence enormously exposing the system to malicious hackers' attacks, in case it will not be timely patched (28)). Therefore, for more safely employing pen-testers in healthcare cybersecurity, it is necessary to re-think hacker ethics, and in particular the understanding that ethical hackers have of it, as something else than just a personal ethics that is subject to an immense variety of strands and interpretations.

[6] This is often the case of hacktivists, that pursue ideological objectives (be they political, religious, or pertaining personal values and motivations) careless of their means' morality, repercussions and adequatness.

Pen-testers comply with a specific interpretation of hacker ethics, namely the one that includes and prioritizes respect for individuals' privacy. This entails that pen-testers do not disseminate or leak data. They also do not intend to cause damage when hacking into systems, nor do they download, modify, or disseminate the data. Their intent is rather to find vulnerabilities and appropriately report them. Therefore, they work towards the establishment of a safer cyber environment. They are institutionalized (through regular employment contracts and certifications) and confine their activity within the law (28). It is true that pen-tests often resort to the same techniques and tools as malicious hackers, but the goal is always to conduct realistic simulations to efficiently bolster cybersecurity without disrupting the workflow (26). Following the present description of pen-testing practice, it seems possible to consider their ethical hackers' ethics as a sort of professional ethics, beyond that of personal ethics. This would allow for a two-fold benefit: it would be possible (and recommended) to draft an international code of ethics that can less arbitrarily define and describe moral principles, standards, expectations, and best practices; also as a consequence, it would facilitate the regulation of their practice and allow punitive measures when said code is disrespected. When professionals disregard their code of ethics they lose the right to practice. Equally, when pen-testers are intentionally violating privacy norms by, for example, breaching data or damaging infrastructures, they could be excluded from the cybersecurity field. This can be enforced with pen-testers as they are regularly employed, and controlling and sanctioning their behaviour can be simpler than with ethical hackers in general. However, as of now, there is no official ethics of conduct for pen-testers. At a professional level, the absence of an ethical code is surprising. A similar code would be a great advantage for further promoting the service of pen-testing, particularly in sensitive sectors such as healthcare. Following the present conceptualization of hacker ethics, it seems possible to consider the employment of ethical hackers as pen-testers in healthcare cybersecurity, with the recommendation of establishing an official code of ethics for their practice.

Conclusion

Cyber threats to healthcare are an unavoidable new reality. However, there are ways to strengthen healthcare cybersecurity. For this sector, cybersecurity is not only about protecting data: health data is particularly sensitive and protecting it equals maintaining patients' safety, privacy, and trust (7). While pen-tests alone will not, and cannot, solve the cybersecurity vulnerabilities of healthcare, they surely can constitute a further measure to bolster it. In this paper, it has been shown how pen-tests are compatible with the healthcare sector and can be advantageous. Other cyber-hygiene steps are needed, among them: removing legacy software like Windows XP and promoting best practices.

Pen-tests can greatly contribute to cybersecurity. It seems that the best way to employ ethical hackers as pen-testers in healthcare is to hire a company providing pen-test services. Hacker ethics, in general, is not particularly relevant for identifying ethical hackers; rather, a particular understanding of this ethics, emphasising privacy and data protection, could help set professional standards for pen-

testers. Therefore, it is recommended to work on an official, national or international, code of ethics for this profession. In addition, considering hacker ethics can raise awareness of the prejudices about the hacking practice and address narratives of fear. For this reason, is it important to acknowledge the variety of hackers and their ethics.

Accepting ethical hackers, especially pen-testers, into our society can bring significant benefits. Firstly, they can greatly contribute to cybersecurity in general, and particularly in the healthcare sector with pen-testers. However, relying solely on pen-tests will not solve the cybersecurity issues of healthcare: other cyber-hygiene measures still need to be implemented and improved. Secondly, they could help address the labour shortage affecting the cybersecurity industry. Again, this is valid not only for healthcare cybersecurity; however, the limited financial resources of this sector constitute a limitation in the employment of pen-testers. Eventually, the prime goal of healthcare is to protect human life and health; balancing financial resources to reinforce the cybersecurity of this sector can contribute to that goal.

Abbreviations

AI	Artificial intelligence
EHR	Electronic health records
Pen-tests	Penetration tests
Pen-testers	Penetration testers
Pen-testing	Penetration testing
UVM	University of Vermont

Declarations

- Ethics approval and consent to participate

Not applicable.

- Consent for publication

Not applicable.

- Availability of data and materials

Not applicable.

- Competing interests

There are no competing interests for any author.

- Funding

The present research is funded by the Swiss National Science Foundation (SNSF) as a part of the National Research Project 77 (NRP 77), project number 187263. The original title of the project is "Ethical and Legal issues of Mobile Health-Data – Improving understanding and eXplainability of digitaL transformation and data technologies using artificial IntelligeNce (EXPLaiN)".

- Authors contributions

GL drafted the majority of the paper. GL, DMS, and BSE contributed to the planning, design, drafting, and critical revision of the paper.

- Acknowledgements

We would like to thank Dr. Markus Christen and Dr. Melanie Knieps from the Digital Society Initiative of the University of Zürich (Switzerland). We are deeply grateful for their support and feedback on the paper. Additionally, we would like to thank the reviwers for their constructive comments that contributed to the quality of the analysis.

References

1. Hellsten H. Cyber risk management in the Finnish healthcare sector [Internet]. [Tampere, Finland]: Tampere University; 2018 [cited 2022 May 5]. Available from: https://www.semanticscholar.org/paper/Cyber-risk-management-in-the-Finnish-healthcare-Hellsten/567b836f752c4afabe95750f222c83b9df9ab357

2. Luna R, Rhine E, Myhra M, Sullivan R, Kruse CS. Cyber threats to health information systems: A systematic review. Technol Health Care Off J Eur Soc Eng Med. 2016;24(1):1–9.

3. Blanke SJ, McGrady E. When it comes to securing patient health information from breaches, your best medicine is a dose of prevention: A cybersecurity risk assessment checklist. J Healthc Risk Manag. 2016;36(1):14–24.

4. Muthuppalaniappan M, Stevenson K. Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health. Int J Qual Health Care. 2021 Jan 1;33(1):mzaa117.

5. Cimpanu C. ZDNet. 2020 [cited 2022 May 12]. Czech hospital hit by cyberattack while in the midst of a COVID-19 outbreak. Available from: https://www.zdnet.com/article/czech-hospital-hit-by-cyber-attack-while-in-the-midst-of-a-covid-19-outbreak/

6. Satter R, Stubbs J, Bing C. Exclusive: Elite hackers target WHO as coronavirus cyberattacks spike. Reuters [Internet]. 2020 Mar 23 [cited 2022 May 12]; Available from: https://www.reuters.com/article/us-health-coronavirus-who-hack-exclusive-idUSKBN21A3BN

7. Martin G, Martin P, Hankin C, Darzi A, Kinross J. Cybersecurity and healthcare: how safe are we? BMJ. 2017 Jul 6;j3179.

8. Desai S, Roberson JE, Serafino MC, Coulter HM. Holland&Knight. 2022 [cited 2022 Jun 10]. Cyber Incident Reporting Requirements for Critical Infrastructure Sectors Signed into Law | Insights | Holland & Knight. Available from: https://www.hklaw.com/en/insights/publications/2022/03/cyber-incident-reporting-requirements-for-critical-infrastructure

9. General Data Protection Regulation (GDPR) [Internet]. 2016 [cited 2021 Jul 5]. General Data Protection Regulation (GDPR) – Official Legal Text. Available from: https://gdpr-info.eu/

10. Coventry L, Branley D. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas. 2018 Jul 1;113:48–52.

11. Wagner D. Health IT Outcomes. 2018 [cited 2022 May 12]. Why Healthcare Is A Top Target For Hackers. Available from: https://www.healthitoutcomes.com/doc/why-healthcare-is-a-top-target-for-hackers-0001

12. Caldwell T. Ethical hackers: putting on the white hat. Netw Secur. 2011 Jul 1;2011(7):10–3.

13. Fazzini K. CNBC. 2019 [cited 2022 May 5]. Why some of the world's top cybersecurity hackers are being paid millions to use their powers for good. Available from: https://www.cnbc.com/2019/05/17/cybersecurity-hackers-are-paid-millions-to-use-their-powers-for-good.html

14. Wenger F, Jaquet-Chiffelle DO, Kleine N, Weber K, Morgan G, Gordijn B, et al. Canvas White Paper 3 – Attitudes and Opinions Regarding Cybersecurity [Internet]. Rochester, NY: Social Science Research Network; 2017 Oct [cited 2021 Sep 16]. Report No.: ID 3091920. Available from: https://papers.ssrn.com/abstract=3091920

15. WEDI. The Rampant Growth of Cybercrime in Healthcare. Workgroup for Electronic Data Interchange; 2017 Aug.

16. Rubenfire A, Conn J. Modern Healthcare. 2017 [cited 2022 May 11]. Building a better cyberdefense. Available from: https://www.modernhealthcare.com/reports/cybersecurity

17. Landi H. Fierce Healthcare. 2021 [cited 2022 Jun 22]. Memorial Health cancels surgeries, reverts to paper records as it responds to cyberattack. Available from: https://www.fiercehealthcare.com/tech/memorial-health-cancels-surgeries-reverts-to-paper-records-as-it-responds-to-cyberattack

18. Bergal J. PEW. 2022 [cited 2022 Jun 22]. Ransomware Attacks on Hospitals Put Patients at Risk. Available from: https://pew.org/3li9O8z

19. Weiner S. AAMC. 2021 [cited 2022 Jun 22]. The growing threat of ransomware attacks on hospitals. Available from: https://www.aamc.org/news-insights/growing-threat-ransomware-attacks-hospitals

20. Arbel N. Forbes. [cited 2022 Jun 22]. The Widening Cybersecurity Talent Gap And Its Ramifications In 2022. Available from: https://www.forbes.com/sites/forbestechcouncil/2022/01/28/the-widening-cybersecurity-talent-gap-and-its-ramifications-in-2022/

21. Perakslis ED. Cybersecurity in health care. N Engl J Med. 2014 Jul 31;371(5):395–7.

22. Parmar A. Hacking wireless insulin pumps [Internet]. MedCity News. 2012 [cited 2022 May 12]. Available from: https://medcitynews.com/2012/03/hacker-shows-off-vulnerabilities-of-wireless-insulin-pumps/

23. Storm D. Computerworld. 2015 [cited 2022 Jun 22]. MEDJACK: Hackers hijacking medical devices to create backdoors in hospital networks. Available from: https://www.computerworld.com/article/2932371/medjack-hackers-hijacking-medical-devices-to-create-backdoors-in-hospital-networks.html

24. Newman LH. Medical Devices Are the Next Security Nightmare. Wired [Internet]. [cited 2022 Jun 22]; Available from: https://www.wired.com/2017/03/medical-devices-next-security-nightmare/

25. Ceccato M, Tonella P, Basile C, Coppens B, De Sutter B, Falcarin P, et al. How Professional Hackers Understand Protected Code while Performing Attack Tasks. In: 2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC). 2017. p. 154–64.

26. Engebretson P. The Basics of Hacking and Penetration Testing: Ethical Hacking and Penetration Testing Made Easy. Elsevier; 2013. 223 p.

27. Allen J. Social Engineering Penetration Testing: Attacks, Methods, & Steps [Internet]. PurpleSec. 2019 [cited 2022 Jul 1]. Available from: https://purplesec.us/social-engineering-penetration-testing/

28. Jaquet-Chiffelle DO, Loi M. Ethical and Unethical Hacking. In: Christen M, Gordijn B, Loi M, editors. The Ethics of Cybersecurity [Internet]. Cham: Springer International Publishing; 2020 [cited 2022 May 25]. p. 179–204. (The International Library of Ethics, Law and Technology). Available from: https://doi.org/10.1007/978-3-030-29053-5_9

29. Irwin L. IT Governance. 2021 [cited 2022 Jun 22]. Ethical hacking vs penetration testing: what's the difference? Available from: https://www.itgovernance.eu/blog/en/ethical-hacking-vs-penetration-testing-whats-the-difference

30. Singh R. How Penetration Testing is Different from Ethical Hacking? [Internet]. Indusface. 2020 [cited 2022 Jun 22]. Available from: https://www.indusface.com/blog/how-penetration-testing-is-different-from-ethical-hacking/

31. Denning DE. Concerning Hackers Who Break into Computer Systems. In: Proceedings of the 13th National Computer Security Conference, Washington, DC. Washington (DC): Information Systems Security; 1990. p. 653–64.

32. European Crime Prevention Network (Eucpn). Cybercrime: A theoretical overview of the growing digital threat. Brussels: European Commission; 2016 p. 39.

33. Falk C. Gray Hat Hacking: Morally Black and White. Center for Education and Research in Information Assurance and Security, Purdue University, Lafayette; 2014. Report No.: 2004–20.

34. Coleman EG, Golub A. Hacker practice: Moral genres and the cultural articulation of liberalism. Anthropol Theory. 2008 Sep 1;8(3):255–77.

35. Malone M. Centre for International Governance Innovation. 2021 [cited 2022 Sep 30]. Ethical Hackers Deserve Plaudits, Not Punishment. Available from: https://www.cigionline.org/articles/ethical-hackers-deserve-plaudits-not-punishment/

36. Auray N, Kaminsky D. The professionalisation paths of hackers in IT security: The sociology of a divided identity. Ann Télécommunications. 2007 Nov 1;62(11):1312–26.

37. Vadén T. gnu.org. 2002 [cited 2022 May 5]. The Hacker Community and Ethics: An Interview with Richard M. Stallman. Available from: https://www.gnu.org/philosophy/rms-hack.html

38. The Mentor. Phrack. 1986 [cited 2022 Jun 22]. Hacker's manifesto. The Conscience of a Hacker. Available from: http://www.phrack.org/issues/7/3.html#article

39. Chaos Computer Club. Chaos Computer Club. [cited 2022 Jun 22]. Hacker Ethics. Available from: https://www.ccc.de/en/hackerethics

Figure reference

1. Gregory, M. Tabletop Exercises and Penetration Testing: What Health Systems Need to Know, in HealthTech Magazine, 2021. https://healthtechmagazine.net/article/2021/12/tabletop-exercises-and-penetration-testing-what-health-systems-need-know. Accessed on 30 Sep 2022.

Chapter 6

The ‘Magical Theory’ of Artificial Intelligence in Medicine:

A Thematic Narrative Analysis

Citation: Lorenzini G., Arbelaez Ossa L., Milford S., Elger B. S., Shaw D. M., De Clerq E.. The ‘Magical Theory’ of Artificial Intelligence in Medicine: A Thematic Narrative Analysis. *Journal of Medical Internet Research*. 2024;3:e49795.

The ‘Magical Theory’ of AI in Medicine: A Thematic Narrative Analysis

Abstract

Background: The discourse surrounding medical artificial intelligence (AI) often focuses on narratives that either hype the technology's potential or predict dystopian futures. AI narratives have a significant influence on the direction of research, funding, and public opinion and thus shape the future of medicine.

Objective: The paper aims to offer critical reflections on AI narratives, with a specific focus on medical AI. The aim of this article is to raise awareness as to how people working with medical AI talk about AI and discharge their “narrative responsibility”.

Methods: Qualitative semi-structured interviews were conducted with 41 participants from different disciplines who were exposed to medical AI in their profession. The research represents a secondary analysis of data using a thematic narrative approach. The analysis resulted in 2 main themes, each with 2 other subthemes.

Results: Stories about the AI-physician interaction depicted either a competitive or collaborative relationship. Some participants argued that AI might replace physicians, as it performs better than physicians. However, others believed that physicians should not be replaced and that AI should rather assist and support physicians. The idea of excessive technological deferral and automation bias was discussed, highlighting the risk of “losing” decisional power. The possibility that AI could relieve physicians from burnout and allow them to spend more time with patients was also considered. Finally, a few participants reported an extremely optimistic account of medical AI, while the majority criticized this type of story. The latter lamented the existence of a “magical theory” of medical AI, identified with techno-solutionist positions.

Conclusions: Most of the participants reported a nuanced view of technology, recognizing both its benefits and challenges and avoiding polarized narratives. However, some participants did contribute to the hype surrounding medical AI, comparing it to human capabilities and depicting it as superior. Overall, the majority agreed that medical AI should assist rather than replace clinicians. The study concludes that a balanced narrative (that focuses on the technology’s present capabilities and limitations) is necessary to fully realize the potential of medical AI while avoiding unrealistic expectations and hype.

Keywords

Artificial intelligence, medicine, physicians, hype, narratives, qualitative research

Introduction

Background

Artificial intelligence (AI) technologies are steadily emerging and intertwining with humans' everyday lives and practices. Their applications are broad and diverse: in the field of health care, AI tools are supporting administrative tasks, predicting patients' prognoses, monitoring health through wearable devices, reading computed tomography scans, accelerating drug discovery and development, and many more applications [1]. Particularly relevant for the present analysis are AI-enabled wearable devices (eg, smartwatches) and clinical decision support systems (CDSSs). CDSSs are AI-based tools that provide diagnostic and treatment suggestions based on patient data and test results [2,3]. They bear the potential to impact physicians' clinical judgment, decision-making process, and their relationship with patients [4]. Lately, CDSSs are being combined with machine learning and deep learning techniques, thus generating hopes for faster and more accurate medical decisions and diagnoses [5]. Machine learning and deep learning are types of AI that continuously learn from the data they are fed [6]. Both wearables and CDSSs are artificial narrow intelligence as they are designed to perform only specific tasks. On the contrary, humans have general intelligence: they can excel in speech recognition, pattern recognition, decision-making, and creating. This is also the goal of AI research: with artificial general intelligence, the aim is to apply the same tool to different areas with similar satisfactory results and performance [7]. As artificial general intelligence is not currently a possibility, this paper focuses on artificial narrow intelligence applied in the medical context as CDSSs or wearable devices.

Our work rests on 2 pillars: the first is medical AI, and the second is the creation and perpetuation of AI narratives by people exposed to AI in their profession. It is in the nature of humans to make sense of things, events, and situations. One way of doing this is through the construction of narratives that link together complex and multifaceted realities while assigning roles, identities, and values. Narratives are, therefore, stories we tell about our lives in a nuanced meaning-making effort [8]. It is important to analyze narratives because they reveal our attitudes, opinions, relationships, and emotions [9]. There is a multitude of general AI narratives (Figure 1), which come mainly from news outlets, science fiction accounts, the technology industry, and academic research. Prominent general AI narratives extensively concentrate on the struggle between humans and machines on different levels (ie, comparing their performances, worrying about job displacement, and wondering to which extent humans will relent control to AI). On the one hand, envisioning a world where AI takes over routine and tedious chores can be uplifting. On the other hand, it seems impossible to put to rest the underlying fear that it will take over everything else too, including more enjoyable and creative tasks [10]. Consequently, job displacement narratives are created based on the preoccupation that AI will render many jobs obsolete, particularly the ones revolving around menial tasks that could easily be automated [11]. This worry is exacerbated by the relentless comparison between humans' and AI's performances, as a means to validate AI's capabilities [12]. In this human-machine struggle, AI is depicted as a superefficient tool

at the service of a heartless capitalistic system [10]. At the same time, AI is appreciated exactly because it holds the potential to simplify humans' lives: it is designed to help humans accomplish more with less effort. AI's achievements are often publicly praised; this is continuously underlined when its performance excels humans' capabilities. Accordingly, positive emotions and optimism are prevalent in social media posts about AI, also when the authors are experts in the field [13]. However, what is not acknowledged as much is that these successes are confined to very specific tasks: an AI that can excel in facial recognition will not automatically perform better than humans in driving cars. The lack of generalizability in AI means that human control and oversight are still pretty much needed. Having said that, narratives on AI taking control of human lives and societies are vastly popular [14]. What is usually incorrectly implied behind these narratives is that AI shares the human desire for greediness and its survival instincts, thus attributing these qualities to anthropomorphized machines [10,15].

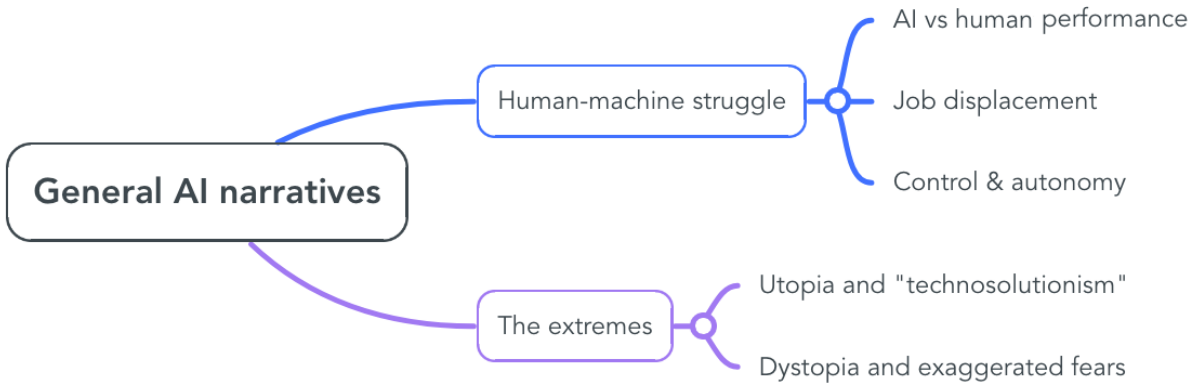


Figure 1: Summary of general artificial intelligence (AI) narratives identified in the literature and pertinent to this analysis.

Dominant AI narratives are often mistrusted or criticized in light of their extremism: they frequently depict either utopian or dystopian futures, light-years away from the complex and mundane reality, that misrepresent the present state of the technology [16]. For example, the way AI fails in the real world is far less epic and catastrophic from Hollywood conceptions: these failures usually happen when the AI does what it is programmed to do but with unintended consequences, that is, a robot trained to behave in ways that would meet humans' approval pretending to be doing something useful [14,17]. The perpetuation of unrealistic AI failures inflates implausible fears while failing to address the real ways in which AI could fail [14]. The debate about AI is very polarized, and as opposed to apocalyptic predictions, there are overly optimistic accounts. The idea of AI being a "master technology" that would be able to unlock all sorts of useful technologies, including those that could help humanity achieve immortality, is common [18]. This leads to the imagination that AI could be considered a form of "holy grail" that bears the potential not only to provide for humanity's needs but also to fulfill its wildest desires and dreams [18].

Narratives can have different functions for different authors and in different situations; in this analysis, the focus is on how narratives could influence medical AI development and uptake and particularly how they could foster a climate where medical AI supports physicians. Indeed, narratives

on AI have the power to influence the further development of these technologies, the availability of funding, the directions of research, and the opinions and expectations of both experts and the public. They influence how new sociotechnical realities are accepted and address both the concerns and the hopes surrounding AI [19]. Therefore, they form the background against which AI is being developed, interpreted, and assessed [16]. While general AI narratives are widely studied and debated, particularly in the Western world [14,19-21], little data are available on AI applications in specific sectors. The lack of research on medical AI narratives, coupled with the perception of AI being particularly promising in the field of health care [22,23], calls for more attention to the topic. Humans have a "narrative responsibility" [24]: there is a duty to make sense of medical AI and to do it responsibly because these sensemaking processes concretely impact its development, implementation, and uptake. Since the stories humans tell about medical AI shape the future of health care, narratives cannot be conceived as normatively neutral. Narratives that support how we wish medicine to be for the years to come should be preferred [8].

Objective

This paper offers a critical reflection on the existing literature on AI narratives. It is one of the first studies to examine the stories told by people who are professionally exposed to medical AI about its applications. This study compares these stories with the existing dominant general AI narratives so as to uncover meaningful similarities and differences. This study aims to raise awareness of how we talk about medical AI and how this can shape the future experiences of both patients and physicians. It is expected that some general AI narratives will be present in medical AI narratives. However, as this medical AI is implemented in a specific sector, namely, health care, with its particular features and challenges, some narratives will be unique for this context. The goal is to understand these similarities and differences to better evaluate medical AI narratives. Consequently, this study aims to recommend a more ethical approach when creating and perpetuating these narratives, considering their impact on physicians' jobs and the physician-patient relationship.

Methodology

Overview

The data used for this manuscript are part of a larger research project titled Ethical and Legal issues of Mobile Health-Data: Improving Understanding and Explainability of Digital Transformation and Data Technologies Using Artificial Intelligence (EXPLaiN), which aims to clarify the legal and ethical issues that need to be resolved for the collection, use, and analysis of health data with AI methods. The project is funded by the Swiss National Science Foundation. The first part of the study consisted of 41 semistructured interviews with participants who are exposed to medical AI. These participants were from a range of disciplines: medicine, philosophy, law, ethics, public health, and computer science. The

interviews focused on the barriers and facilitators for the implementation of AI in clinical settings, particularly regarding CDSSs and wearable devices. The original study aimed to examine the current views, attitudes, knowledge, and barriers to using AI models in the analysis of health data and to support physicians and patients in their decision-making.

This analysis is a secondary analysis of these data and focuses on a subset of the data collected. While coding the data, it became apparent that narratives were often discussed. This justified a secondary analysis that was attentive to this aspect of the data. A second code tree was created based on the narratives identified in the literature, and the interviews were recoded. Of the 41 interviews, 30 (73%) were selected for the secondary analysis based on the presence of narrative elements about AI in health care. This selection, inherent to the secondary nature of the analysis, resulted in incomplete saturation in 1 subtheme, namely, "welcoming the holy grail."

The data subset was analyzed using a thematic narrative approach that identified and reported stories participants told about medical AI [25,26]. This approach was chosen for its flexibility and ability to allow large data sets to be managed and reduced into themes [25]. The topic and the format of the data are not conducive to a structural narrative approach, as the narrative segments were relatively short and lacked common narrative characteristics (eg, characters with roles, a narrator, a complication, a resolution, and a coda) [27,28]. Therefore, a narrative thematic analysis was chosen, as it enabled single units of meaning, primarily phrases, and short paragraphs to be formed into themes and interpreted narratively [29]. With a narrative thematic approach, we could better describe how people exposed to AI in their profession experienced and understood medical AI, as well as how they made sense of it. This allowed for the analysis of underlying assumptions and values [27,30].

Participants

Participants were purposively sampled and came from various disciplines and backgrounds: medicine, bioethics, public health, philosophy, psychology, economy, law, and computer science. Inclusion criteria, other than being exposed to medical AI in their profession, were the holding of a senior position, either in academia or in the private sector, hence excluding PhD students, interns, and junior professionals. Participants' profiles were, for example, full professorship at a university, chief executive officer of a company working with AI, or a data protection officer at a hospital. Participants were recruited internationally; however, there was a focus on European and Swiss participants since the EXPLaiN project aimed to especially explore their attitudes. Participants were recruited because they were working with medical AI through projects, products, research, and development. Identification of participants occurred through publications or affiliations with companies working in the field of medical AI. Their email addresses were found on the web through their institution or company's website. At the end of the interview, participants were asked if they knew someone meeting the inclusion criteria who would be interested in participating (snowball sampling).

First contact with participants was through email where they were invited to be interviewed by introducing the project and explaining the aims and the implications of their participation (eg, time commitment, voice recording, the method of transcription, and data pseudonymization format).

Data Collection and Analysis

LAO and GL, who recruited the participants and conducted the one-on-one semistructured interviews, did not personally know the participants. LAO has a background in medicine and public health, while GL studied philosophy with a focus on ethics and philosophy of science. At the time of the data collection, both were PhD candidates in bioethics at the Institute of Biomedical Ethics of Basel.

Data were collected from November 2021 to April 2022 (therefore preceding some breakthrough such as ChatGPT; it could be hypothesized that after the most recent novelties in the AI field, such as natural language processing tools, narratives about AI might be different also in the health care sector) through semistructured interviews that lasted an average of 50 minutes. All the interviews of this subset were conducted on the web and recorded directly via Zoom (Zoom Video Communication, Inc). The original interview guide was composed of 13 questions, each with several prompts or follow-ups. The interview guide made use of 3 vignettes to better clarify and contextualize the questions. The questions were divided into 6 blocks: introductory questions (about the experience of the participant), general questions about using AI in medical practice, context-related questions about AI-patient relationships (vignette 1 involving a wearable device), context-related questions about physician-patient relationships with AI (vignette 2 involving CDSSs), context-related questions about private-public relationships (vignette 3), and closing questions. The more significant questions (reported in Textbox 1) for this analysis were questions numbered 3 and 4, as well as 3 prompts for question 8. However, relevant data were found elsewhere in the data set as the interviews were semistructured, and participants had some freedom in guiding the topics of the interview.

Question number	Question
3	I would like to start discussing clinical usability. What do you think about using AI in clinical practice?
4	What would you consider the biggest challenges of using AI in healthcare?
8.6	How important it is that the doctor understands AI?
8.8	Would AI have an impact on the doctor-patient relationship?
8.9	Would AI challenge the traditional model of shared decision-making?

Textbox 1: Relevant questions and prompts from the interview guide.

The interviews were transcribed verbatim by LAO, GL, and 2 students at the University of Basel using MAXQDA (VERBI GmbH), a software application designed to assist with qualitative analysis methods. LAO and GL checked all the transcripts and compared their correctness with the audio of the interviews. All data were securely stored on the server of the University of Basel and pseudonymized. Potentially reidentifiable information was removed from the transcripts.

After the original inductive coding, conducted equally by GL and LAO, GL reorganized the relevant coded sections for the secondary analysis. Upon consulting existing literature to identify dominant narratives, a new code tree was created, and the selected segments were deductively recoded. The selected data subset was interpreted through the lens of the existing categories of general AI narratives [31]. The new code tree composed the dominant AI narratives found in the literature. GL then selected the most significant codes and grouped them into themes.

Ethical Considerations

All methods were approved by the Ethics Committee of Northwest and Central Switzerland, under Switzerland's Human Research Act (HRA) Article 51 [32]. The methods were carried out in accordance with the relevant HRA guidelines and regulations. After revision, the Ethics Committee of Northwest and Central Switzerland concluded that interviewing AI professionals falls outside the HRA and requires only verbal consent at the beginning of an interview (declaration of no objection AO_2021-00045).

All personal data were pseudonymized and safely stored on the server at University of Basel. The key is accessible only to the research team. Potentially reidentifiable data were omitted from publication. No compensation was offered to participants.

Results

Overview

For this analysis, we used 30 interviews and reported at least 1 quote from each. We selected this subset because narratives were not prominent in all interviews. It was challenging to categorize participants into disciplines, as AI is notoriously an interdisciplinary field. More often than not, participants had mixed backgrounds and were dealing with medical AI from different points of view. In categorizing participants, we picked their main expertise: 9 (30%) participants had a background in medicine, 6 (20%) in bioethics, 6 (20%) in law, 3 (10%) in computer science, 2 (7%) in public health, 2 (7%) in philosophy, 1 (3%) in psychology, and 1 (3%) in economy. The vast majority of the selected participants (21/30, 70%) were male (female participants: n=9, 30%). Only 5 (17%) participants were located outside Europe: 3 (10%) in the United States, 1 (3%) in Canada, and 1 (3%) in South Africa (for more details on the participants, please refer to the Multimedia Appendix 1).

Our analysis identified 2 main themes and various subthemes (Figure 2). Representative anonymized quotes were taken from the interviews to illustrate the reported results. Participants are identified with the abbreviation of their main expertise and a number: bioethics (BE), computer science (CS), economy (EC), law (LW), medicine (ME), public health (PH), philosophy (PL), and psychology (PS).

Figure 2: Themes and subthemes that emerged from the thematic narrative analysis. AI: artificial intelligence; MAI: medical artificial intelligence.

1. *Medical AI as a Game Changer*

With regard to physicians and medical AI relationships, attitudes fell into 2 main groups. Some participants depicted a rather competitive relationship and compared the performances and capabilities of medical AI to those of physicians. The majority, however, emphasized how AI can support clinicians, thus outlining a more collaborative relationship and focusing on the benefits of this cooperation. Nevertheless, these 2 groups shared the underlying idea that AI would be a game changer for medicine, and both emphasized how it could be useful in health care.

1.a.) A Competitive Relationship

Medical AI Outperforming Doctors

Some participants described AI as a competitor to physicians and argued that not only clinicians are dependent on AI but also they could even be replaced by it. Medical AI was said to outperform clinicians in pattern recognition and data processing. AI was believed to notice aspects that physicians would miss, hence emphasizing the limitations of human capabilities and describing AI as being faster, more accurate, and less costly than human physicians:

> *AI is able to grasp so many ideas within a very small time interval [and] also integrate information that doctors might even oversee that this might be even like more precise than physicians. And I think this is also an advantage.*[ME7]

> *The AI tool uncovers a pattern that the clinician did not pick up or maybe could not have picked up within a human limited abilities.*[BE5]

> *[Medical AI is] very inexpensive to use. In principle, like once you've trained the system for let's say a diagnosis, you can basically use these things on a*

regular laptop or smartphone even. […] It doesn't come for free, but it is rather inexpensive and easy to get.[BE4]

Comparing physicians and AI performance, abilities, and costs sometimes resulted in claims about the obsolescence of physicians since AI would be better in many aspects of a physician's role, while also being faster and cheaper, and it seemed to be preferable to delegating tasks to AI. It was rarely implied that physicians as a whole would be replaced. More commonly, it was suggested that some specific tasks could be carried out by AI. A common limitation was that AI could not interact with patients as at present it lacked the necessary skills. Presuming that AI capacities would steadily improve, a few participants wondered whether in the future medical AI might be able to assume all physicians' duties:

Nowadays there are certain things that might not be outsourced to machines in terms of human interactions. But on the other side, I think, if we wait long enough, we can basically outsource everything to machines.[PH1]

I'm pretty sure that the doctor will be quite cautious, at least in the beginning, when they know that they use these kinds of products[medical AI], but maybe with time, you know, when they are used to it, in like 5 years, 10 years, 15 years, maybe with time they could lose probably autonomy.[LW6]

You will actually have better outcomes if you don't involve humans.[ME9]

Things were different for image recognition: several participants mentioned that medical AI gave outstanding results in radiology. This led to the possibility of outsourcing routine cases to AI while consulting radiologists only for peculiar cases:

I think one of the big places where it's already implemented is in radiology. Meaning, recognition of patterns in pictures; machines are better at it than we are.[ME1]

What people have been doing in radiology, I think it's also awesome. […]. *The machine can give you feedback right away and maybe you just use the humans for very specific cases*.[CS1]

And as more people use that tool, there might be the temptation that therefore maybe we don't need as many dermatologists. Or as many specialists in certain areas, like radiology. Because we have very good AI that is able to detect cancer from X-rays. Or covid from X-rays of lungs.[PS1]

The Risk of Technological Deferral

While pondering the idea of a more autonomous medical AI, many participants worried about the risks of excessive technological deferral (giving too much power to technology). Automation bias, namely, the tendency to over-rely on automatic decision-making tools, was mentioned as an issue in areas of practice that are time-sensitive:

In the long run [doctors] end up with them just following what the machine says.[PL1]

There is a very real risk, especially in areas of practice that have time pressure, that we will see automation bias, that we will see AI systems that formally were advisory, actually being the ones who decide treatment choices.[BE2]

This tension on who holds the final decision-making responsibility was framed as an actual conflict with potentially detrimental consequences if the humans were to 'lose' their decision-making power. Physicians might also be intimidated by this outstanding tool and therefore would struggle to override its decisions even when they did not agree with it:

Can you even win, so to speak? So, that might be the bigger danger, where you say like "well, the machine says that, so therefore it is correct".[PL2]

The recurrent mentioned consequences of deferring decision-making powers to medical AI were dependency on technology, with fewer and fewer specialists trained and a gradual loss of autonomy for physicians. Many participants in this group worried about physicians' autonomy being endangered by medical AI and described the technology as authoritarian or tyrannical:

Well, if the algorithms prove to be better than doctors then you would have to change the role of doctors from decision-makers to more just like people, in the end, giving injections.[CS2]

1.b.) A Collaborative Relationship

The Question of Irreplaceability

For many participants, physicians are not to be replaced by AI, rather, AI enables them and supports their daily practice and decision-making activities:

> *It should go in the direction that the systems are not seen as a competitor to the doctors but more as a cooperation between both. And I think what it's worthwhile, what it's important, it's that the cooperation leads to better results*.[EC1]

> *The use of technology is going to assist the physician and not harm because in the end it's called a clinical decision support tool, not a clinical decision maker tool*.[ME3]

Some interviewees noted how humanity is irreplaceable, others described medical AI as an assistive tool that is not designed to replace physicians but to empower them. Participants in this group emphasized the idea of medical AI "assisting", "helping", "empowering", and "supporting" clinicians rather than comparing their ability, accuracy, and cost.

When emphasizing physicians' irreplaceability, participants referred to the sensibility, emotivity, and empathy that are needed in medical decision-making. Given the current state of the art, medical AI is unable to grasp the complex totality of the patient's situation. Many participants also questioned patients' willingness to relinquish the physician-patient relationship in favour of an AI-patient relationship. They argued that communication with AI would not be authentic as it would not consider patients' personhood. Therefore, these participants concluded that medical AI should never override physicians' decisions, rather, it should promote and preserve physicians' agency:

> *The patient needs a person he can talk [to], a person that can read their emotions, feelings*.[LW4]

> *I think medicine has a certain degree of nuances, that only a person might catch and not a computer. And you can't let these computers or AI run autonomously*.[ME4]

Independently from AI capabilities (whether it outperforms physicians or not and whether it is limited or not), physicians remain essential: medical AI should always be considered in the light of

physicians' clinical judgment and never left unsupervised. According to this description of medical AI, physicians should always keep an active role in decision-making:

> *I think the one who has the responsibility to make decisions is, or will always be, the doctor.*[LW4]

> *I expect that the technology will help you give an assessment, but that you will still have a clinician that will evaluate further that kind of technical assessment by software. So it's not fully replacing an intervention as such. It's helping, supporting a development.*[LW5]

Medical AI Freeing Physicians

The collaborative relationship narrative does not depict physicians as dependent on their tools; rather, it suggests that medical AI could constitute an important resource. The relationship is described as a fruitful partnership, and the outcome would be a general improvement to both physicians' practice and their work conditions. Medical AI could free physicians from burdensome tasks, hence relieving them from burnout and allowing them to spend more quality time with patients:

> *[Medical AI] could improve the doctor-patient communication [...] So I am kind of hoping that, in that way, because of AI certain aspects of healthcare could be simplified and automated, but that equally should generate room for more empathy between doctors and patients.*[PS1]

> *[Medical AI] is helping doctors to really focus on, or be able to have more time for patients and less to spend with tools*.[CS3]

> *What I hope it'll do it's improve the relationship between the patient and the physician. What I mean by this is [that] the physician is going to be relieved from the burnout.*[ME3]

2. *The Power of Medical AI*

Most of the participants were optimistic about the future of medicine when AI was involved and reported an overall positive impact, or potential, of this technology. While a large part of the answers balanced medical AI's advantages with the challenges it introduces, some focused only on the benefits of the technology. At the same time, many interviewees identified a hype-type narrative of medical AI

and problematized it. In this context, hype is understood as an exaggeratedly optimistic rhetoric about an emerging technology [33].

2.a.) Welcoming the Holy Grail

In a few interviews, medical AI was discussed mainly in positive terms. These participants did not see any negative aspects or concerns about the technology. Medical AI was deemed always useful, and if it was not useful for something yet, it surely would be in the future. It encapsulated so many opportunities for health care that 1 participant referred to it as "the holy grail." Consequently, medical AI was expected to solve a wide range of problems:

> *I basically don't see any negative effects, like, I can't really see any negative effects.*[LW1]

> *So, it seems to me that it's both inevitable and good that we have it [medical AI].*[BE2].

> *What do you think about using AI or machine learning in clinical practice?[GL] I think it's the Holy Grail.[CS1]*

2.b.) Medical AI is Not Magical

A significant part of the participants addressed the romanticization of this technology and highlighted the importance of promoting a more truthful narrative. "Truth" and "reality" were terms often mentioned when discussing the medical AI hype: it was deemed untruthful, unhelpful, and unrealistic, and this was judged problematic:

> *The problem is that this enthusiasm is so uncritical and then we build into this. This is not giving us the truth and not helping us to generate probabilities. This is the problem that I hugely see.*[BE3]

According to the participants, one of the consequences of the hype around medical AI is that it is impossible to live up to the expectations that it builds. Therefore, some participants were profoundly critical toward overhyped accounts of the capabilities of medical AI:

> *There is so much hype in this field [medical AI] and this builds narratives and expectations. And to live up to those expectations is always challenging.*[ME2]

> *So that has, probably now backwards looking, not been so clever to phrase it as the silver bullet solution to everything, to patient autonomy, or patient empowerment, to more efficient and better healthcare*[BE1].

The outcome of this ideology is that medical AI is portrayed as the appropriate means to tackle every pressing issue of health care: AI is the hammer that fixes everything. Techno-solutionist narratives would misunderstand AI and promote a representation of the technology as if it were some kind of magical tool:

> *The hype around the technology at the moment, you know, that people think that it can solve everything. It's like they have a hammer and everything is a nail*.[PH2]

> *I think a lot of people and a lot of doctors kind of have the magical theory of machine learning, where you just kind of throw the numbers in the hopper, shake it out, and you get the results by magic*.[BE6]

> *The major problem of deep learning today are the people doing deep learning because they think they will solve everything with that and the ignorant people because they don't understand what is deep learning and they think it's magic that will solve everything*.[ME6]

Discussion

Principal Findings

The accounts of medical AI that emerged from these interviews are more realistic and less influenced by science fiction narratives than the general discourse on AI. Dystopian futures were not reported, and only a few participants described AI as a utopian technology that would address all challenges faced by the health care sector. While general AI narratives are usually polarized, describing AI as either the milestone of a better future for humanity or the cause of all evils [19,34,35], study participants often found a middle path between medical AI's promises and risks, thus avoiding alignment with extreme positions and providing instead a more nuanced depiction of the technology. We hypothesize that people exposed to AI in their profession are less prone to exaggerated and polarized narratives, while lay people tend to be more susceptible to these narratives as they feel they have less control over the technology [36]. The lack of a strongly polarized discourse in medical AI can be

regarded as positive: the contradictions present in narratives that are diametrically opposed and irreconcilable hinder a nuanced and sophisticated understanding of the technology [21].

However, our study sample was not exempt from hype narratives that uncritically focused on the expected benefits of medical AI. This confirmed the existence of hype narratives, which are already reported in the literature as well as the conceptualization of AI as a "holy grail" technology [22,23,34].

Claims about superiority are very popular in AI narratives, not only in fictional and media narratives but also in the scientific discourse, as researchers frequently compare AI with humans' capabilities and performances as a means of validating the technology [12,37]. The physician-AI juxtaposition ends with depicting the classical human-machine struggle panorama, where physicians are menaced by an authoritarian machine that outperforms them and that leaves humans dependent on it, no longer in control, and stripped of their agency [12,14,19,35]. Indeed, 1 participant described this struggle as a real win-lose situation.

While a few participants hyped medical AI, the majority recognized both the advantages and the challenges introduced by AI in health care. Therefore, stronger than the hype narrative were the cautionary tales of avoiding a "myopic techno-solutionism" and the criticism of this hype [34]. Techno-solutionism is the ideology in which every kind of problem (technical, social, economic, political, psychological, or physical) can be ameliorated with an "appropriately designed" technological solution [38]. Attributing magical properties to AI, meaning that it can somehow address every problem, reveals a shallow understanding of the technology. This requires better education, which can be achieved through the establishment of a more balanced narrative that realistically assesses medical AI's current capabilities and shortcomings [37].

Participants confirmed the idea that medical AI narratives can sometimes be detached from the everyday reality of the technology and that the hyping of AI leads to unrealistic expectations and overpromising while obscuring technological bottlenecks [19,21]. Therefore, our findings demonstrate that the current dominant narratives can mislead the understandings of medical AI, even in people working with it. Instead, "narratives should focus on the realities of AI's present capabilities" [34] and take into account the narrative responsibility that is always entailed when the future of medicine with AI is imagined. Every story we tell about medical AI shapes its development, adoption, and perception in health care in ways that are not normatively neutral.

Accordingly, almost all participants recognized the limitations of AI. There is a risk that by failing to acknowledge the potential problems and shortcomings of medical AI, the hype narrative might further exacerbate these hidden specters. The need for a more realistic narrative that returns the image of the actual state of the art is commonly present both in the interviews and in the debate about AI narratives [14,19,34].

With the exception of a few participants, there was a general agreement that AI could not and would not replace human clinicians. This finding is present in the literature about the future of medicine with AI; for example, patients appeared less prone to seek medical assistance if AI provides it, even if it was better than a human expert [39]. When it comes to this topic, there is an alignment between different narratives that appear to share similar moral codes according to which medical AI cannot entirely replace the physician's role or human interaction [40]. Therefore, this could be regarded as the "proper narrative" of the AI-physician relationship, and, as such, it might take the form of a collective narrative or "imaginary," judged true without a need for further justification [41]. The prevailing idea remains: "patients will always need human physicians" [42].

Having determined that medical AI is to assist clinicians, it remains to be assessed whether it will have an impact on the physician-patient relationship. Some participants believed that medical AI would ameliorate their relationships, for example, by allowing physicians to spend more time with patients. This is also a popular idea in the literature to the extent that some claim that medical AI could be an opportunity to make physicians more human and empathetic [43-47]. However, as with many things about AI, opinions are divided, and this idea is also widely criticized. It could be that physicians will visit more patients in the time AI saved, thus maintaining the status quo or worsening care provisions [12,48,49]. Consequently, medical AI might not necessarily have a positive impact on the physician-patient relationship as either the participants in our study or many prominent voices in research think.

Limitations

There is a clear prevalence of a Western perspective in our study. Hence, it remains questionable whether our findings are valid in other contexts.

The interview guide we used for this study focused on certain applications of AI in medicine, namely CDSS and wearable devices (e.g. smartwatches). This may have limited the discourse on possible outcomes and futures. Moreover, question 8.8. discusses the "traditional" model of shared decision-making; this wording could be considered nonneutral and leading.

Prior to this paper, we have conducted theoretical research on the ethical issues of medical AI. This led us to publications where we took a position on the role of AI in healthcare and the doctor-patient relationship. We concluded that medical AI is currently, and should continue to be, an assistive tool that should support doctors' and patients' decision-making. We acknowledge that this belief was already sedimented at the time of data collection and analysis, thus possibly shaping the way in which we presented the results.

Conclusions

Through the establishment of a more realistic and nuanced medical AI narrative, it is easier to describe AI tools as assistive. The discourse about their benefits, risks, and possible applications is less spectacular. Narratives that support the idea of AI augmenting humans' capabilities, rather than substituting them, should be preferred as these narratives better correspond to the current reality of the technology [34]. It is also fundamental to raise awareness of the narrative responsibility that humans have to make sense of, interpret, and narrate medical AI in a way that shapes a positive future for medicine. Similarly, humans are responsible for scrutinizing the dominant narratives and evaluating them [24]. Everyone has this responsibility when talking about medical AI, including researchers, since we all can impact the future of technology, although to different degrees. Failing to exercise this narrative responsibility would entail relinquishing our sense-making task to other narrators (eg, big tech, transhumanists, governments, etc). The consequence would be a world in which we live in the narrative created by others for us. This world would be one in which the majority of humanity delegated the construction of our future to a few, in that they did not participate in the process that would shape what mattered most in the present [24,50].

Disproportionate fears and expectations could halt the development of medical AI, for example, by generating opposition or disillusionment when the technology does not live up to its promised expectations [19,21]. Medical AI narratives shape the role of AI in societies in ways that are ethically and politically relevant and can influence the perceptions of citizens, policy makers, politicians, health care personnel, and researchers [8,16]. Therefore, narratives have a constitutive role that is more than strictly descriptive: it is performative. Narratives have the power to decide the future of medical AI [51,52]. We argue that it is important to recognize the role that narratives of technologies play for humanity and reflect on which type of narrative is dominant in medical AI. This is a fundamental ethical issue that cannot be overlooked. It must be addressed so as to shape our desired future for medicine.

Acknowledgements

The authors thank the people at the Institute for Biomedical Ethics of Basel for their support of this study, particularly, Dr Tenzin Wangmo and Dr Michael Rost for their precious expertise and kind advice. This paper and the research on which it is based are funded by the Swiss National Science Foundation (SNF) in the context of the Ethical and Legal issues of Mobile Health-Data: Improving Understanding and Explainability of Digital Transformation and Data Technologies Using Artificial Intelligence (EXPLaiN) project (National Research Projects 77; grant or award number 407740_187263/1).

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

GL prepared the original draft for this paper. GL, EDQ, and DMS participated in the conceptualization, in the methodology choice, and in the secondary data analysis. GL and LAO conducted the data collection and together with SM, and DMS participated in the original data analysis. BSE was responsible for funding acquisition. All the authors contributed to the review and editing of the article.

Conflicts of interest

None declared.

Multimedia Appendix 1

Research participant	**Country**	**Sex**	**Expertise**
ME1	Africa	Male	Medicine
ME2	Europe	Male	Medicine
ME3	North America	Male	Medicine
ME4	North America	Male	Medicine
ME5	Europe	Female	Medicine
ME6	Switzerland	Male	Medicine
ME7	Europe	Female	Medicine
ME8	Switzerland	Male	Medicine
ME9	Switzerland	Female	Medicine
BE1	Europe	Female	Bioethics
BE2	Europe	Male	Bioethics
BE3	Switzerland	Female	Bioethics
BE4	Europe	Male	Bioethics
BE5	Europe	Male	Bioethics
BE6	North America	Female	Bioethics
LW1	Switzerland	Male	Law
LW2	Switzerland	Female	Law
LW3	Europe	Male	Law
LW4	Switzerland	Male	Law
LW5	Europe	Male	Law
LW6	Switzerland	Male	Law
CS1	Europe	Male	Computer science
CS2	Switzerland	Male	Computer science
CS3	Europe	Male	Computer science
PH1	Switzerland	Female	Public health
PH2	Europe	Female	Public health
PL1	Europe	Male	Philosophy
PL2	North America	Male	Philosophy
PS1	Europe	Male	Psychology
EC1	Europe	Male	Economy

References

1. Basu K, Sinha R, Ong A, Basu T. Artificial intelligence: how is it changing medical sciences and its future? Indian J Dermatol. 2020;65(5):365-370. [FREE Full text] [CrossRef] [Medline]
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. Feb 6, 2020;3:17. [FREE Full text] [CrossRef] [Medline]
3. Berner ES, La Lande TJ. Overview of clinical decision support systems. In: Berner E, editor. Clinical Decision Support Systems: Theory and Practice. Cham, Switzerland. Springer; 2007:1-17.
4. Lorenzini G, Arbelaez Ossa L, Shaw DM, Elger BS. Artificial intelligence and the doctor-patient relationship expanding the paradigm of shared decision making. Bioethics. Jun 25, 2023;37(5):424-429. [CrossRef] [Medline]
5. Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. "Brilliant AI doctor" in rural clinics: challenges in AI-powered clinical decision support system deployment. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan. [CrossRef]
6. Du Y, McNestry C, Wei L, Antoniadi AM, McAuliffe FM, Mooney C. Machine learning-based clinical decision support systems for pregnancy care: a systematic review. Int J Med Inform. May 2023;173:105040. [FREE Full text] [CrossRef] [Medline]
7. Kaplan A, Haenlein M. Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz. Jan 2019;62(1):15-25. [FREE Full text] [CrossRef]
8. Coeckelbergh M. Time machines: artificial intelligence, process, and narrative. Philos Technol. Oct 23, 2021;34:1623-1638. [CrossRef]
9. Wertz FJ, Charmaz K, McMullen LM, Josselson R, Anderson R, McSpadden E. Five Ways of Doing Qualitative Analysis: Phenomenological Psychology, Grounded Theory, Discourse Analysis, Narrative Research, and Intuitive. New York, NY. Guilford Publications; 2011.
10. Dihal K. Enslaved minds: artificial intelligence, slavery, and revolt get access arrow. In: AI Narratives: A History of Imaginative Thinking about Intelligent Machines. Oxford, UK. Oxford University Press; 2020:189-212.
11. Guenduez AA, Mettler T. Strategically constructed narratives on artificial intelligence: what stories are told in governmental artificial intelligence policies? Gov Inf Q. Jan 2023;40(1):101719. [CrossRef]
12. Ostherr K. Artificial intelligence and medical humanities. J Med Humanit. Jun 11, 2022;43(2):211-232. [FREE Full text] [CrossRef] [Medline]

13. Manikonda L, Kambhampati S. Tweeting AI: perceptions of lay versus expert Twitterati. In: Proceedings of the International AAAI Conference on Web and Social Media. 2017. Presented at: ICWSM-17; May 15-18, 2017; Montreal, Quebec. [CrossRef]
14. Recchia G. The fall and rise of AI: investigating AI narratives with computational methods. In: Cave S, Dihal K, Dillon S, editors. AI Narratives: A History of Imaginative Thinking about Intelligent Machines. Oxford, UK. Oxford University Press; 2020:382-408.
15. Singler B. Artificial intelligence and the parent–child narrative. In: Cave S, Dihal K, Dillon S, editors. AI Narratives: A History of Imaginative Thinking about Intelligent Machines. Oxford, UK. Oxford University Press; 2020.
16. Cave S, Dihal K, Dillon S. Introduction: imagining AI. In: Cave S, Dihal K, Dillon S, editors. AI Narratives: A History of Imaginative Thinking about Intelligent Machines. Oxford, UK. Oxford University Press; 2020.
17. Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei A. Deep reinforcement learning from human preferences. Preprint posted online June 12, 2017. [FREE Full text]
18. Cave S. AI: artificial immortality and narratives of mind uploading. In: Cave S, Dihal K, Dillon S, editors. AI Narratives: A History of Imaginative Thinking about Intelligent Machines. Oxford, UK. Oxford University Press; 2020.
19. Cave S, Craig C, Dihal K, Dillon S, Montgomery J, Singler B, et al. Portrayals and perceptions of AI and why they matter. The Royal Society. Nov 2018. URL: https://royalsociety.org/-/media/policy/projects/ai-narratives/ai-narratives-workshop-findings.pdf [accessed 2024-07-23]
20. Cave S, Dihal K. The whiteness of AI. Philos Technol. Aug 06, 2020;33:685-703. [CrossRef]
21. Vicsek L. Artificial intelligence and the future of work – lessons from the sociology of expectations. Int J Sociol Soc Policy. Oct 06, 2020;41(7/8):842-861. [CrossRef]
22. Cameron D, Maguire K. Public views of machine learning: digital natives. The Royal Society. Oct 2017. URL: https://royalsociety.org/-/media/policy/projects/machine-learning/digital-natives-16-10-2017.pdf [accessed 2024-07-23]
23. Frost EK, Carter SM. Reporting of screening and diagnostic AI rarely acknowledges ethical, legal, and social implications: a mass media frame analysis. BMC Med Inform Decis Mak. Dec 10, 2020;20(1):325. [FREE Full text] [CrossRef] [Medline]
24. Coeckelbergh M. Narrative responsibility and artificial intelligence: how AI challenges human responsibility and sense-making. AI Soc. Dec 30, 2021;38(6):2437-2450. [CrossRef]
25. McAllum K, Fox S, Simpson M, Unson C. A comparative tale of two methods: how thematic and narrative analyses author the data story differently. Commun Res Pract. Nov 25, 2019;5(4):358-375. [CrossRef]
26. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. Jan 2006;3(2):77-101. [CrossRef]

27. Bruner J. The narrative construction of reality. In: Beilin H, Pufall PB, editors. Piaget's Theory: Prospects and Possibilities. Mahwah, NJ. Lawrence Erlbaum Associates, Inc; 1992:229-248.
28. Clarke V, Braun V. Thematic analysis. J Posit Psychol. Dec 09, 2016;12(3):297-298. [CrossRef]
29. Ross JA, Green C. Inside the experience of anorexia nervosa: a narrative thematic analysis. Couns Psychother Res. Jul 22, 2010;11(2):112-119. [CrossRef]
30. Fisher WR. Narration as a human communication paradigm: the case of public moral argument. Commun Monogr. Jun 02, 2009;51(1):1-22. [CrossRef]
31. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol. Feb 13, 2022;9(1):3-26. [CrossRef]
32. Federal Act on research involving human beings (Human Research Act, HRA), chapter 9: research ethics committees. Fedlex. 2011. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en#chap_9 [accessed 2024-07-23]
33. van Lente H, Spitters C, Peine A. Comparing technological hype cycles: towards a theory. Technol Forecast Soc Change. Oct 2013;80(8):1615-1628. [CrossRef]
34. Chubb J, Reed D, Cowling P. Expert views about missing AI narratives: is there an AI story crisis? AI Soc. Aug 25, 2022:1-20. [FREE Full text] [CrossRef] [Medline]
35. Klarmann N. Artificial intelligence narratives: an objective perspective on current developments. Preprint posted online March 18, 2021. [FREE Full text]
36. Borup M, Brown N, Konrad K, Van Lente H. The sociology of expectations in science and technology. Technol Anal Strateg Manag. Jul 2006;18(3-4):285-298. [CrossRef]
37. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? J Educ Eval Health Prof. Jul 03, 2019;16:18. [FREE Full text] [CrossRef] [Medline]
38. Gardner J, Warren N. Learning from deep brain stimulation: the fallacy of techno-solutionism and the need for 'regimes of care'. Med Health Care Philos. Sep 1, 2019;22(3):363-374. [CrossRef] [Medline]
39. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res. Dec 2019;46(4):629-650. [CrossRef]
40. Berkhout F. Normative expectations in systems innovation. Technol Anal Strateg Manag. Jul 2006;18(3-4):299-311. [CrossRef]
41. Konrad K. The social dynamics of expectations: the interaction of collective and actor-specific expectations on electronic commerce and interactive television. Technol Anal Strateg Manag. Jul 2006;18(3-4):429-444. [CrossRef]

42. Mittelman M, Markham S, Taylor M. Patient commentary: stop hyping artificial intelligence-patients will always need human doctors. BMJ. Nov 07, 2018;363:k4669. [CrossRef] [Medline]
43. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York, NY. Basic Books; 2019.
44. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. JAMA. Jan 02, 2018;319(1):19-20. [CrossRef] [Medline]
45. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. JAMA. Aug 13, 2019;322(6):497-498. [CrossRef] [Medline]
46. Mittelstadt B. The impact of artificial intelligence on the doctor-patient relationship. Council of Europe. URL: https://www.coe.int/en/web/bioethics/report-impact-of-ai-on-the-doctor-patient-relationship [accessed 2023-04-14]
47. Liu X, Keane PA, Denniston AK. Time to regenerate: the doctor in the age of artificial intelligence. J R Soc Med. Apr 2018;111(4):113-116. [FREE Full text] [CrossRef] [Medline]
48. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak. Apr 20, 2023;23(1):73. [FREE Full text] [CrossRef] [Medline]
49. Sparrow R, Hatherley J. High hopes for "deep medicine"? AI, economics, and the future of care. Hastings Cent Rep. Jan 18, 2020;50(1):14-17. [CrossRef] [Medline]
50. Sanz Menéndez L, Cabello C. Expectations and learning as principles of shaping the future. Unidad de Políticas Comparadas (CSIC). Feb 2000. URL: https://digital.csic.es/bitstream/10261/1491/1/expectations_learning.pdf [accessed 2023-04-26]
51. Guice J. Designing the future: the culture of new trends in science and technology. Res Policy. Jan 1999;28(1):81-98. [CrossRef]
52. van Lente H. Navigating foresight in a sea of expectations: lessons from the sociology of expectations. Technol Anal Strateg Manag. Sep 2012;24(8):769-782. [CrossRef]

Abbreviations

AI: artificial intelligence

CDSS: clinical decision support system

EXPLaiN: Ethical and Legal issues of Mobile Health-Data: Improving Understanding and Explainability of Digital Transformation and Data Technologies Using Artificial Intelligence

HRA: Human Research Act

Chapter 7

Discussion

7.1. Discussion

The exact extent of medical artificial intelligence (MAI) impact on the daily routine of clinical practice, its consequences for doctors and patients' autonomy, communication, safety, and decision-making, is not yet sure(1–3). This is partly because of the novelty of MAI applications; particularly clinical decision support systems (CDSS), which could introduce the most changes in the doctor-patient relationship, are yet not widely implemented(1). As the deployment of MAI is nascent, there is still room to decide how to routinely implement these systems to try and attain the most benefits and avoid pitfalls.

Despite the uncertainty that surrounds the introduction of MAI in healthcare, there is a wide effort to predict the changes, challenges, and advantages it will bring along. There seems to be a tendency to rely on MAI to solve an immense variety of longstanding issues in healthcare(4). These range from helping overworked doctors to remember the interaction between drugs or a patient's vulnerability to a drug side effect(5), to rendering the workflow more efficient and less costly(1,6). Some desired effects are highly debated. For example the "gift of time" for doctors, namely the idea that by automatizing some tasks with MAI doctors will have more time to devote to their relationship with patients(4). Here opinions vary and, while nobody argues that dedicating more time to patients is undesirable, those sceptics about this possibility are just as loud as the supporters(6–8).

Opinions are ambivalent also regarding the doctor-patient relationship. MAI could both allow for more empathy and meaningful communication between doctors and patients, as well as requiring doctors to pay more attention to computers thus reducing their engagement with patients(9). Similarly, MAI entertains hopes to empower patients: they could be better informed and educated about their health, diagnosis, prognosis, and treatment alternatives, hence being better positioned to actively take part in the decision-making process. Therefore, MAI could hold the potential to allow for patients' preferences and values to be heard and respected. At the same time, however, it seems that MAI could limit their autonomy(10,11): patients may simply blindly rely on technology, without (being motivated to) understand how to properly use it for their gain or disregarding their preferences when confronted with MAI suggestions.

Not all the discussions about MAI focus on the expected benefits. There are concerns about the lack of transparency, biased algorithms, and the actual accuracy of the systems. Many worries seem to concern the implications of MAI for doctors' professional autonomy. For example, CDSS could render redundant some doctors' skills, thus generating uncertainty on the future of medical education and profession(8). It is common to compare doctors' and MAI performances and wonder whether, or sometimes when, doctors will become redundant as a whole and therefore replaced(5). At the same time, many claim the irreplaceability of human doctors, either because some skills cannot be delegated to a

machine (e.g. empathy, understanding of social determinants of health, consideration of contextual information) or because it is not desirable (e.g. patients prefer to be taken care by humans)(12–16).

Implementing MAI in healthcare entails as many challenges as the benefits it promises. With a focus on CDSS, the following sections summarize the Chapters' recommendations on how to introduce these technologies in healthcare while trying to preserve and promote good practices, such as shared decision-making (SDM), patients' empowerment and safety, and doctors' professional autonomy. The risks that MAI poses to patients' care must not be overlooked but described, addressed, and studied. With these recommendations, we aim to promote a usage of MAI and CDSS that empowers patients, allows doctors to provide better care, and fosters SDM, all while maintaining that these tools are meant to assist and support doctors.

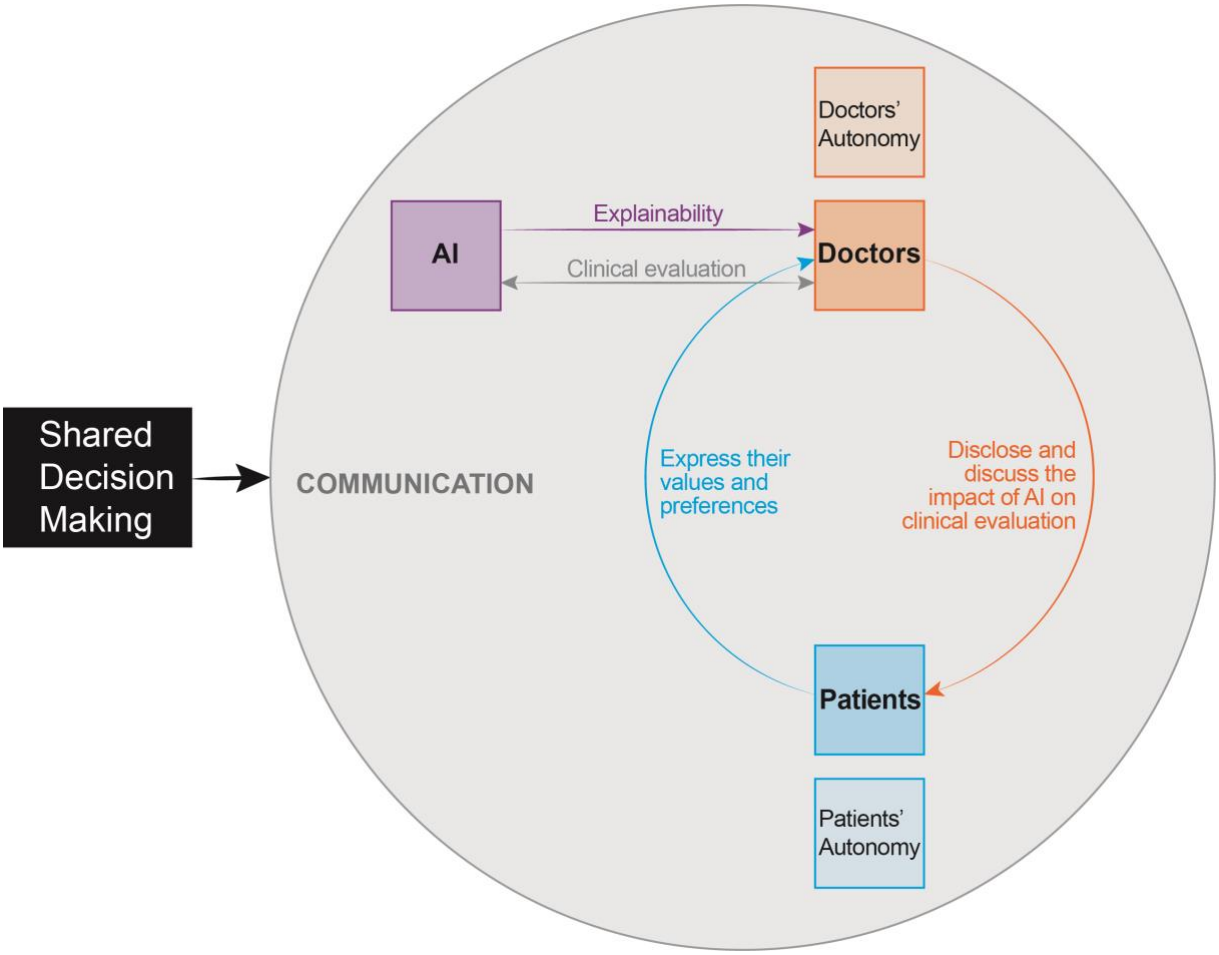
7.2. The Outlook of Medicine

The digitalisation of healthcare bears the potential to fundamentally change the medical profession and the patients' experience(11,12,17). More and more attention is given to data: electronic health records (EHR) collect and store a huge amount of information that can be easily shared and transferred with other healthcare professionals(18). Furthermore, data gathered with EHR can be used to train machine learning and deep learning algorithms to be used as CDSS. As EHR and CDSS are expected to improve patients' care, much effort and funding are put into further advancing these technologies(19). However, medicine is not only health data and state-of-the-art technology. For caring for patients, doctors not only need accurate diagnoses and effective therapies but also honest and empathetic communication. Communication is essential both for better understanding the medical history and social context of patients but it also constitutes the base of a trustful doctor-patient relationship(20,21). When this relationship is no longer cultivated, and priority is given to more technical aspects of medicine, the risk could be the erosion of trust hence undermining the ground on which patients feel comfortable sharing sensitive personal information. If the focus on health data takes over other, more humanistic, sides of medicine, the risk is to perceive data representations of patients as an objective measure of health, thus reducing the importance of the patient as a socially embodied person(1).

Were the paradigm of medicine to shift more towards hard sciences, like mathematics and computer science, at the expense of its humanistic half, the doctor-patient relationship would greatly lose its centrality(22). This is not to say that the innovations introduced by computer science in healthcare will, by default, impair the doctor-patient relationship. However, the possibility for a paradigm change, as envisioned in Chapter 3, is real(23). Care must be taken to steer the direction of this change to avoid damaging the relationship and, instead, further cultivating it. While science is an important, fundamental, and indispensable part of medicine, it does not constitute the whole discipline: medicine is also an art, at its core there is the healing relationship that takes care of human lives, not

only organs and bodies(22). MAI can contribute to both souls of medicine only if this duality is maintained and balanced.

The paradigm shift we envisioned in Chapter 3 for the future of medicine is still based on SDM. This means that careful consideration must be paid to coming to a final shared decision not only between doctors and patients but also with AI, which could enter the relationship as a third actor. It is implied that both doctors and patients remain a fundamental part of the relationship, which turns from a dual partnership to a triadic one. By entering the doctor-patient relationship as a third actor that is only a part of the shared decision, and not the sole decision-maker, MAI bears the potential to play a positive for doctors and patients(4).



Claims about MAI replacing doctors led to serious concerns that such changes could result in the death of the doctor-patient relationship(24). Throughout the thesis, we reject this idea on the basis that doctors shall remain essential actors and mediators in the healing relationship. Only by preserving and valuing the doctors' role in the relationship, we can hope for human empathy, connection, listening and compassion(13). It seems that CDSS in particular can be a double-edged sword for the doctor-patient relationship: it could make space for better connection and care or it could add another layer of complexity that would require further explanation and could undermine trust and autonomy(24).

It is too early to assess the impact of CDSS on the doctor-patient relationship, however, it is the right time to raise awareness on the best strategies to implement them within the existing framework of SDM. It seems imperative to not sacrifice the human factor and therefore not attempt to replace human doctors(4). Eventually, the highest and best contribution of CDSS is exactly what is in their name: supporting doctors' work, optimizing and improving what they already do(25). The focus should not be on comparing doctors and CDSS performances, but rather on how CDSS can augment doctors(5). The role that doctors play in the SDM relationship, and in healthcare in general, is unique and indispensable, even when CDSS can read X-rays better than radiologists. So while many tasks can be delegated to CDSS, doctors will have the responsibility to supervise them and make the final decision(26). Were doctors not to hold this final responsibility, their professional autonomy would be questioned, and this could detrimentally impact their relationship with patients. The following section explores the specific recommendations for implementing SDM within the MAI-doctor-patient relationship.

7.3. The Spectre of a New Paternalism

Paternalism has been defined in Chapter 3 as the opposite paradigm of what we wish to promote, which is instead SDM. In a paternalistic relationship, patients' autonomy is hindered and the final decision is taken by the doctor based on what they consider the patient's best interest(27). It is an undesirable paradigm that risks being (re) introduced with MAI. This new form of paternalism would be a double paternalism: not only patients' autonomy would be limited, but also doctors' professional autonomy could be endangered(11,12,28). On the one hand, doctors could end up following the recommendations of a paternalistic CDSS, for example, without being able to understand and evaluate them(29). They would not have the responsibility of the final decision-making, but would rather follow the directions suggested by the system as it is supposed to be more accurate and reliable than human doctors. In turn, it would be difficult to discuss the CDSS outcome with patients, to properly inform them and to empower them to participate in a decision-making process where in the first place doctors themselves did not partake.

It is therefore fundamental to promote both doctors and patients' autonomy when MAI is involved in the relationship. Doctors' professional autonomy is necessary for patients to practice their own autonomy. Communication is important also within the MAI-doctor relationship: doctors should be competent in understanding and evaluating MAI suggestions and limitations. MAI should therefore be explainable, although not in an exhaustive way (e.g. being fully transparent or training doctors on the technical inner workings of the algorithm). Explainability should be contextual and include multiple explanations such as assumptions, limitations, and usability(29). This kind of explainability could allow doctors to freely exercise their clinical judgment with the support of MAI, rather than despite it or instead of it. When doctors are both competent to conscientiously integrate MAI in their clinical practice and in charge of the final decision-making, SDM can be achieved. Doctors can form a partnership with MAI

that can draw force from the strengths of both parties and deliver better care than doctors or MAI alone(11).

The second step of the triadic SDM relationship requires including patients in the shared decision-making process. To accomplish this, some sort of informed consent might be necessary. It remains to be defined which informed consent procedure should be implemented: formalized informed consent can serve as a proxy for autonomy and SDM, but it also risks distorting SDM into a stagnant transactional event (i.e. signing a consent form). Although implementing an informed consent requirement for MAI is impractical, it could be beneficial for the MAI-doctor-patient relationship and patients' empowerment(9,30). Moreover, the practical difficulties of an informed consent requirement appear to be contingent and not structural, hence allowing for the possibility to overcome them. Not only these obstacles could be overcome, but MAI itself may present new and unforeseen opportunities to improve informed consent(31).

When reviewing the expected advantages of MAI, patients' empowerment, independence, and liberation are often cited(10,32,33). However, we must be careful to avoid the independence from healthcare professionals offered by MAI does not transform into a dependence on technology(34). Similarly to doctors, also patients require some kind of contextual explanation in order to use MAI tools in a way that fosters participation in their own care. In the clinical context, where doctors are using MAI, discussing it with patients can help better position them to evaluate MAI recommendations together with the doctor and to ensure that their values and preferences are included in the final shared decision-making. In Chapter 4 we argued that informing patients about MAI enables them to better participate in SDM, hence promoting their autonomy and valuing their preferences. This transparent communication is part of the requirements for the triadic SDM paradigm. If patients are not informed about MAI, it will be more challenging for them to express their preferences, ask for further clarification, and participate in their own care. The final decision risks to be shared solely between MAI and doctors as patients were not included in this process. A trustworthy MAI-doctor-patient relationship requires a certain degree of transparency, and that includes both communication and explainability(9). Despite the practical difficulties of informing patients about MAI, there is the ethical consideration that only transparent communication enables patients to actively participate in the decision-making process with their own expertise, which is important for the success of care. Doctors have the role of mediating between MAI and patients, to communicate in a meaningful way MAI outputs to patients, to ensure that the latter are informed and can actively participate in the decision-making(1).

Avoiding a new form of paternalism that would undermine both doctors' and patients' autonomy is important when routinely implementing CDSS(32,35). The risk for both is to put complete trust in the system hence making it more difficult to overlook its recommendations. This is usually referred to as automation bias, namely the phenomenon where doctors follow MAI's outputs without looking for further confirmatory evidence(36,37). Secondly, it is important to be competent to evaluate these

recommendations; this implies explainability and informed consent. Lastly, all the parties should be able to participate in the final decision-making as this is the only way to ensure that patients' values and preferences are considered and respected. There are, however, ideas to allow for patients' values to be considered by the system to guarantee even more their inclusion(35,38). A paternalistic MAI would mean that its outputs are put into practice without evaluating and challenging them, on the basis of the perceived MAI's authority. Instead, within SDM, MAI should allow for transparency and plurality(4).

7.4. Priority Should Be Given to Patients' Safety

The risks introduced by MAI are not only the direct ones of a new paternalism, discrimination, bias, and lack of transparency(39). There are also indirect risks such as the increased vulnerability of healthcare cybersecurity, as seen in Chapter 5. It is an indirect consequence as it is not the use of MAI tools per se that is causing this issue, rather, the introduction of interconnected and cloud-based technology (amongst which lies MAI) augments the opportunities for malicious intruders to break into the system(40,41). These intrusions constitute a serious risk to patients' safety and privacy as the stolen medical information can be sold on the dark web (for identity theft purposes, hence causing a dangerous intermingling of medical information) or publicly posted on the internet(42). Other than the dissemination of highly sensitive and personal data, the concern is that healthcare professionals cannot do their job properly when they do not have access to patients' data(43). This puts at risk human lives and slows down the intervention times. Strengthening cybersecurity is therefore fundamental to preventing cyberattacks cause tangible harm to human health(1). MAI promise of improved care will be broken if patients' safety could be exposed to greater risks(42).

Healthcare is a particular field, consequently, cybersecurity issues here assume a specific importance(44). In healthcare, cybersecurity is not only about securely storing secretive and important data, ensuring the operations flow, and safeguarding the system's integrity. Healthcare cybersecurity is tasked also with the duty to protect the lives of vulnerable patients, to ensure that their care is not undermined by cyber threats. It is about protecting patients' safety, privacy, and trust(45,46).

An unexplored approach to healthcare cybersecurity is penetration testing (pen-test), which is instead widely employed in other fields such as finance. There, it has proven effective in identifying and patching the system's vulnerabilities, which could no longer be exploited by malicious hackers. In Chapter 5, we considered the role of penetration testers (pen-testers) and concluded that it is compatible with the needs and peculiarities of healthcare. When employed through serious pen-test companies and with the creation of a code of conduct for this profession, the sub-category of ethical hackers that are pen-testers could greatly contribute to healthcare cybersecurity. They should be encouraged to find vulnerabilities and become an integral part of a wider collaborative effort towards stronger healthcare cybersecurity(47).

Indeed, penetration tests (pen-tests) should be an approach amongst other cyber-hygiene measures. Many other issues cannot be addressed with pen-tests alone as healthcare is a widely complex and multi-layered field that requires a mixed approach to cybersecurity. Further measures should be: replacing legacy software, raising doctors' awareness of cybersecurity and encouraging them to educate patients, and promoting best practices (e.g. frequently changing passwords)(47). It could even be imagined that informed consent could address data ownership and privacy issues to further educate patients on the matter while empowering them to make decisions about the handling of their health data(47). Eventually, both doctors and patients should understand the potential cybersecurity risk associated with MAI and make informed decisions when integrating it into their daily routine, be it a wearable device or CDSS(47).

There is an ethical question of resource allocation when it comes to cybersecurity in healthcare: it requires a careful balancing of risks and benefits when deciding how much budget to allocate. Surely, cybersecurity can improve patients' safety. At the same time, priority is usually given to those activities that directly impact patients' well-being such as treatments, new surgical tools, hygiene, training of healthcare professionals, etc. This limits the possibilities for healthcare facilities to opt for pen-test, particularly the smaller ones that may have less budget to allocate for cybersecurity.

Despite these limitations, pen-test practices should be extended to healthcare. Awareness of hackers' practice and variety is important to understand how pen-testers could contribute to healthcare cybersecurity. As professionals hired by serious companies, pen-testers are held responsible for their actions, they cannot resort to questionable or dangerous practices to try and break into the system, and they are bound to the highest standards of ethics(48). Unfortunately, although there are many unofficial codes of ethics for ethical hackers, there lacks an official code of conduct for professional pen-testers that further holds them accountable for their actions. Such a code could constitute the basis to prevent the further exercise of this profession in case they would engage in activities that would break the code. An official code of conduct would constitute a useful measure for more confidently employing pen-testers in healthcare. Therefore, it is recommended not only to consider bolstering healthcare cybersecurity with pen-testers but also to synergetically join forces to create an official code of conduct, at least on a national level.

7.5. Redimensioning the Hype

Artificial intelligence (AI) has seen many breakthroughs in the last decade and, consequently, it is a very popular topic that attracts the hopes and expectations of those who believe new technologies can solve every problem or can do more than what they achieve. This hype, namely the tendency to exaggerate AI benefits while failing to properly acknowledge its shortcomings, can steer funding into its development and research. It is therefore instrumentalised, at times, since there are financial interests in keeping the hype up and feeding it(3). At the same time, this hype creates a distorted representation

of what AI is, what it is supposed to do, how to correctly implement and use it, and its present capabilities. In Chapter 6, we investigated how experts talk about MAI to conclude that it is not exempt from a hype tendency. In MAI narratives, the hype is present mainly in the form of comparing doctors and MAI performances. In these comparisons, it is always the machine that comes out as a winner: a certain superiority of MAI over human doctors seems to be asserted. This, in turn, raises concerns about future doctors' role and the doctor-patient relationship(24). It also generates an exaggerated perception of what MAI can achieve, thus feeding both disproportionate hopes and fears.

Hype narratives constitute the basis of the idea that MAI will eventually replace doctors. The lack of evidence on the possibility of developing systems that can take over all of the doctors' tasks and responsibilities renders unrealistic, or at least improbable, the claims humans have been replaced or that are obsolete(5,49). Instead, MAI is being introduced as a support for doctors, to assist them and optimize their work(5,22,25). More realistic narratives, that acknowledge the current state and possibilities of the technologies, are essential to better understand how MAI can contribute to healthcare, how to integrate it efficiently in the clinical routine, which are the most needed and most promising directions of research, and what are the risks, challenges, pitfalls, and obstacles of introducing these systems. These nuanced and truthful narratives are beneficial for MAI future development as they recognize both the benefits and dangers of the technology and balance them. They allow also for more accurate and proportionate expectations, concerns, and understanding of MAI both for experts and laypeople. It seems that promoting narratives of MAI-doctor collaboration could be beneficial instead: MAI development would focus more on this supportive role while experts, lay people, and media outlets would be more prone to reflect on MAI-doctor collaboration rather than comparison.

Narratives can impact the future direction and implementation of MAI as their role is not only descriptive but performative: they can decide the future of medicine(50,51). There is therefore a responsibility when writing and talking about MAI that requires us to reflect on the narratives we are perpetuating. To different degrees, we are all responsible for preferring hype-free narratives that promote more truthful expectations when discussing MAI(3). It is what can be called a narrative responsibility, that encourages us to be aware of the impact that our narratives have in the real world while also urging us to exercise this responsibility by choosing to sustain narratives that we deem worthy(52). In a similar way, we are responsible for evaluating and assessing the current narratives to avoid them being imposed on us. It is important to recognize harmful hype narratives that polarize the discourse about MAI and prevent a proper understanding of the reality of the technology. Consequently, we have to decide how we want to impact the real world: we should prefer those narratives that describe the future we think would be the best for medicine.

7.6. Further Research

Some of the recommendations presented in the preceding Chapters, and summarized here, already indicate further research directions. Regarding the implementation of CDSS in hospitals, two issues should be addressed in the near future. It should be further investigated how CDSS can be introduced in the doctors' daily clinical work without disrupting it but, instead, supporting and optimizing it. This requires careful consideration of the reality of clinical work with its time and resource limitations. However, such considerations are out of the scope of the present research, which is mainly theoretical and focused on normative discussions. The second issue is likewise practical and has been previously mentioned: there is a need to understand how informed consent for MAI applications could be introduced. Bearing in mind that the ethical conclusions demonstrate the benefits of such a requirement for patients' autonomy and SDM, it is worth exploring how the healthcare system could accommodate this. It could also be that eventually informed consent for MAI will not be implemented because it would be included in the practices that are already in place. Similarly, it might be sufficient to discuss MAI with patients without introducing a proper specific informed consent procedure. But it could be the case, instead, that dedicated informed consent practices will be implemented when MAI becomes more routinely intermingled with clinical work. Only with further research, it will be possible to identify which alternative would better promote patients' self-determination while avoiding overburdening healthcare professionals. Together with practical considerations, a legal analysis will probably be needed to assess the present-day situation and understand which basis is there to implement formal informed consent for MAI. The synergy of practical and legal considerations could indicate the best way forward.

Regarding the need for further normative analysis, what we deem of primary importance is to create a code of conduct for pen-testers. This is needed not only for employing pen-tests services in healthcare: it would benefit all the fields that currently use this approach or are considering adopting it in the near future. It would also give more credibility to these professionals by strengthening trust and accountability. Considering the increasing numbers of cyberattacks and risks, there is a high-priority need to enforce cybersecurity (not only in healthcare, although it is a particular sector and requires special consideration and protection). In order to produce a sound code of conduct, the needs of special fields such as healthcare and finance should be considered, as well as the ethical hackers' codes already available. Particular attention should be paid to privacy, data integrity, safety, and confidentiality.

7.7. Limitations

Below are presented the limitations of the overall study. Limitations of the single Chapters are discussed in the respective Chapters.

The reported empirical results cannot be generalized as it was a widely heterogeneous sample. Thus, it is not representative of either national or international experts' opinions and attitudes or experts

from a certain field (e.g. medicine or philosophy). Still, it met the research project's objective to explore thoughts, feelings, risks, benefits, obstacles, and facilitators of MAI. The results offer an incomplete but meaningful overview of the various attitudes that MAI experts from various disciplines hold.

It should also be noted that most of the participants were not native English speakers. Although they were all proficient in English, this could have limited their ability to express their opinions. It could have also influenced the depth in which they could have articulated and understood complex and sometimes challenging information. Their choice of words might have not been completely accurate and this could have in turn biased the analysis. The same limitation applies to L. Arbelaez Ossa and I, who conducted the interviews. However, whenever misunderstanding or uncertainty was spotted, further clarification was attempted.

Lastly, my background in Philosophy and ethics might have impacted the planning and conducting of the research, as well as the report of the results. It could have moved the focus toward a more normative approach and introduced evaluative elements in the supposedly impartial phases of the analysis (e.g. coding). Indeed, the coding process included mixed inductive and deductive codes. Yet, this perspective was counterbalanced by the expertise, experience, and contributions of my colleagues who had diverse backgrounds (Medicine, Public Health, Theology, and Psychology).

7.8. Conclusions and recommendations

This thesis placed the doctor-patient relationship at the core of medicine: it constitutes the basis on which care can be provided, treatments chosen, and prognoses discussed. Cost-saving and efficiency considerations should come only after it has been assessed that this relationship is strong and healthy. Therefore, when we think about MAI and how it could impact healthcare, it is of foremost importance to understand how it could change the doctor-patient relationship. Introducing MAI in healthcare can be challenging from many points of view: there are concerns about lack of transparency, accuracy of the algorithms, and regulations. However, as healthcare is primarily dealing with human lives, the first step should be to ensure that patients' well-being, values, and self-determination are safeguarded from any potential adverse consequences. The following bullet points summarize the conclusions of the present research and prioritize once again human interaction and patients' safety.

- MAI should not replace doctors but only provide tools to support and improve their work. Doctors should be competent to supervise, evaluate, and overview MAI recommendations. The quality and quantity of human interaction should not be sacrificed.

- MAI should be pursued only for the good of the patients, not for the sake of cost-saving.

- MAI should not constitute an obstacle to a good doctor-patient relationship, rather, it could be a third actor that, when appropriately introduced, could foster SDM and empower patients to participate in the decision-making process with their values and preferences.

- A collaborative relationship between MAI, doctors, and patients should be established: by joining forces, rather than competing, they can achieve better care.

- MAI bears the potential to foster the doctor-patient relationship and improve SDM if correctly implemented. Otherwise, it bears the risk of introducing a new form of paternalism: a double paternalism where doctors could be stripped of their autonomy.

- Properly implemented informed consent can protect patients' self-determination and better positions them to take part in SDM. Adequate information-sharing tactics should be put into practice.

- Priority should be given to empathy and human interaction. This could help avoid the time saved with MAI will be used to increase the throughput of patients.

- Cybersecurity is not only an IT issue but more attention should be paid to it as MAI will further exacerbate the present vulnerabilities and this could negatively impact patients' safety and trust.

- Misrepresentations of MAI promote false expectations and fears while promoting the idea that it can be the silver bullet solution for complex and longstanding problems. It is not only necessary to debunk hype narratives but also to develop and promote a more truthful and nuanced MAI narrative.

In conclusion, MAI will, in one way or another, impact healthcare and the doctor relationship. The changes it will introduce, either radical or superficial, will not necessarily be for the worse: when carefully implementing MAI tools, such as CDSS, they bear the potential to positively impact the doctor-patient relationship and foster SDM. MAI could be a third actor in the medical relationship and this could either empower doctors' and patients' autonomy or threaten it. Eventually, it all depends on how MAI will be implemented, how much attention will be paid to patients' rights and values, and how well doctors will use these tools to assist their daily clinical work.

7.9. References

1. Mittelstadt B. The Impact of Artificial Intelligence on the Doctor-Patient Relationship [Internet]. Council of Europe; 2021 Dec [cited 2023 Apr 14]. Report No.: F-67075. Available from: https://rm.coe.int/inf-2022-5-report-impact-of-ai-on-doctor-patient-relations-e/1680a68859

2. Vento DD, Fanfarillo A. Traps, Pitfalls and Misconceptions of Machine Learning applied to Scientific Disciplines. In: Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) [Internet]. New York, NY, USA: Association for Computing Machinery; 2019 [cited 2023 Aug 8]. p. 1–8. (PEARC '19). Available from: https://dl.acm.org/doi/10.1145/3332186.3332209

3. Leufer D. Why We Need to Bust Some Myths about AI. Patterns. 2020 Oct 9;1(7):100124.

4. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak. 2023 Dec;23(1):1–14.

5. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019 Oct 4;7:e7702.

6. Nittas V, Daniore P, Landers C, Gille F, Amann J, Hubbs S, et al. Beyond high hopes: A scoping review of the 2019–2021 scientific discourse on machine learning in medical imaging. PLOS Digit Health. 2023 Jan 31;2(1):e0000189.

7. Topol EJ. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. First Edition. New York: Basic Books; 2019. 378 p.

8. Sparrow R, Hatherley J. High Hopes for "Deep Medicine"? AI, Economics, and the Future of Care. Hastings Cent Rep. 2020;50(1):14–7.

9. Astromskė K, Peičius E, Astromskis P. Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. AI Soc. 2021 Jun 1;36(2):509–20.

10. De Proost M, Segers S. We need to talk about disruption in bioethics: a commentary on Rueda, Pugh and Savulescu. Trends Biotechnol. 2023 Jun 1;41(6):741–2.

11. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan;28(1):31–8.

12. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. JAMA. 2018 Dec 4;320(21):2199.

13. Liu X, Keane PA, Denniston AK. Time to regenerate: the doctor in the age of artificial intelligence. J R Soc Med. 2018 Apr;111(4):113–6.

14. Academy of Medical Royal Colleges. Artificial Intelligence in Healthcare [Internet]. 2020. Available from: https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf

15. Yun JH, Lee EJ, Kim DH. Behavioral and neural evidence on consumer responses to human doctors and medical artificial intelligence. Psychol Mark. 2021;38(4):610–25.

16. Mittelman M, Markham S, Taylor M. Patient commentary: Stop hyping artificial intelligence—patients will always need human doctors. BMJ. 2018 Nov 7;k4669.

17. Braun M, Hummel P, Beck S, Dabrock P. Primer on an ethics of AI-based decision support systems in the clinic. J Med Ethics. 2021 Dec 1;47(12):e3–e3.

18. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy. 2011 Dec 31;4:47–55.

19. Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. 'Brilliant AI Doctor' in Rural China: Tensions and Challenges in AI-Powered CDSS Deployment. Proc 2021 CHI Conf Hum Factors Comput Syst. 2021 May 6;1–18.

20. Ha JF, Longnecker N. Doctor-patient communication: a review. Ochsner J. 2010;10(1):38–43.

21. Chandra S, Mohammadnezhad M, Ward P. Trust and Communication in a Doctor-Patient Relationship: A Literature Review. J Healthc Commun [Internet]. 2018 Jul 19 [cited 2021 Apr 21];3(3). Available from: https://healthcare-communications.imedpub.com/abstract/trust-and-communication-in-a-doctorrnpatient-relationship-a-literature-review-23072.html

22. Niel O, Bastard P. Artificial Intelligence in Nephrology: Core Concepts, Clinical Applications, and Perspectives. Am J Kidney Dis. 2019 Dec 1;74(6):803–10.

23. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digit Med. 2020 Feb 6;3(1):1–10.

24. Aminololama-Shakeri S, López JE. The Doctor-Patient Relationship With Artificial Intelligence. Am J Roentgenol. 2019 Feb;212(2):308–10.

25. Norman A. Your future doctor may not be human. This is the rise of AI in medicine. [Internet]. Futurism. 2018 [cited 2023 Aug 10]. Available from: https://futurism.com/ai-medicine-doctor

26. Van Cauwenberge D, Van Biesen W, Decruyenaere J, Leune T, Sterckx S. “Many roads lead to Rome and the Artificial Intelligence only shows me one road”: an interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. BMC Med Ethics. 2022 Dec;23(1):1–14.

27. Jauhar S. When Doctors Need to Lie. The New York Times [Internet]. 2014 Feb 22 [cited 2021 Apr 21]; Available from: https://www.nytimes.com/2014/02/23/opinion/sunday/when-doctors-need-to-lie.html

28. Taddeo M, Floridi L. How AI can be a force for good. Science. 2018 Aug 24;361(6404):751–2.

29. Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. Digit Health. 2022 Jan 1;8:20552076221074490.

30. Lysaght T, Lim HY, Xafis V, Ngiam KY. AI-Assisted Decision-making in Healthcare. Asian Bioeth Rev. 2019 Sep 1;11(3):299–314.

31. Michalski A, Stopa M, Miśkowiak B. Use of Multimedia Technology in the Doctor-Patient Relationship for Obtaining Patient Informed Consent. Med Sci Monit Int Med J Exp Clin Res. 2016 Oct 26;22:3994–9.

32. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46:205–2011.

33. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach. 2018;28(4):689–707.

34. Segers S, Mertes H. The curious case of “trust” in the light of changing doctor–patient relationships. Bioethics. 2022;36(8):849–57.

35. McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics. 2019;45:156–60.

36. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 2019 Mar 1;28(3):231–7.

37. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. J Am Med Inform Assoc JAMIA. 2003;10(5):478–83.

38. Birch J, Creel KA, Jha AK, Plutynski A. Clinical decisions using AI must consider patient values. Nat Med. 2022 Jan 31;1–3.

39. Char DS, Abràmoff MD, Feudtner C. Identifying Ethical Considerations for Machine Learning Healthcare Applications. Am J Bioeth AJOB. 2020 Nov;20(11):7–17.

40. Luna R, Rhine E, Myhra M, Sullivan R, Kruse CS. Cyber threats to health information systems: A systematic review. Technol Health Care Off J Eur Soc Eng Med. 2016;24(1):1–9.

41. Muthuppalaniappan M, Stevenson K. Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health. Int J Qual Health Care. 2021 Jan 1;33(1):mzaa117.

42. Jarrett MP. Cybersecurity—A Serious Patient Care Concern. JAMA. 2017 Oct 10;318(14):1319–20.

43. Riggi J. American Hospitals Association. [cited 2023 Aug 18]. The importance of cybersecurity in protecting patient safety. Available from: https://www.aha.org/center/cybersecurity-and-risk-advisory-services/importance-cybersecurity-protecting-patient-safety

44. Wenger F, Jaquet-Chiffelle DO, Kleine N, Weber K, Morgan G, Gordijn B, et al. Canvas White Paper 3 – Attitudes and Opinions Regarding Cybersecurity [Internet]. Rochester, NY: Social Science Research Network; 2017 Oct [cited 2021 Sep 16]. Report No.: ID 3091920. Available from: https://papers.ssrn.com/abstract=3091920

45. Coventry L, Branley D. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas. 2018 Jul 1;113:48–52.

46. Wagner D. Health IT Outcomes. 2018 [cited 2022 May 12]. Why Healthcare Is A Top Target For Hackers. Available from: https://www.healthitoutcomes.com/doc/why-healthcare-is-a-top-target-for-hackers-0001

47. Das S, Siroky GP, Lee S, Mehta D, Suri R. Cybersecurity: The need for data and patient safety with cardiac implantable electronic devices. Heart Rhythm. 2021 Mar 1;18(3):473–81.

48. Singh R. How Penetration Testing is Different from Ethical Hacking? [Internet]. Indusface. 2020 [cited 2022 Jun 22]. Available from: https://www.indusface.com/blog/how-penetration-testing-is-different-from-ethical-hacking/

49. Berkhout F. Normative expectations in systems innovation. Technol Anal Strateg Manag. 2006 Jul 1;18(3–4):299–311.

50. Guice J. Designing the future: the culture of new trends in science and technology. Res Policy. 1999 Jan 1;28(1):81–98.

51. Van Lente H. Navigating foresight in a sea of expectations: lessons from the sociology of expectations. Technol Anal Strateg Manag. 2012 Sep;24(8):769–82.

52. Coeckelbergh M. Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. AI Soc [Internet]. 2021 Dec 30 [cited 2023 Feb 22]; Available from: https://link.springer.com/10.1007/s00146-021-01375-x

Appendix

Interview guide experts

a) Introductory questions

	Questions	Follow up probes
	Can you tell me a little bit about yourself?	● What is your background? ● Where are you located? ● What is your role? ● Can you tell me more about your current projects? ● Do you work with AI? ● How did you become interested in AI? ● What is your experience with AI?

b) General questions about using AI in clinical practice

	Questions	Follow up probes
	I would like to start discussing using AI in clinical practice. What do you think about using AI in healthcare?	● In which cases do you think is useful and NOT useful? ● Has it been beneficial so far?
	How/Where do you think AI can/should be implemented in clinical practice?	● Is this a high priority need in clinical practice? ○ Is there any specific application that you think would be the most useful?
	What do you think about using AI for supporting doctors and patients in clinical decisions?	● Do you have any concerns? ● Do you see any benefits? ● Could it affect ○ Doctor-patient relationship? ○ Trust?

		○ Autonomy? ○ Informed consent? ● Would understanding AI be relevant in this context? ○ What type of knowledge? ● How would you define explainability?
	What would you consider the biggest challenges of using AI in healthcare ?	● What are the biggest obstacles to overcome? ● What challenges have you faced using AI in clinical practice? ● Do you have any ethical or regulatory concerns? ● Do you think AI can cause concerns for physicians? ● Do you think AI can raise concerns for patients?
	What would you consider the biggest challenges of using AI for supporting doctor and patients in clinical decisions?	● How do you think policy makers should be involved? ● How do you think clinicians should be involved? ● How do you think computer scientists should be involved? ● How do you think ethicists should be involved? ● How do you think patients should be involved?
	I would like to know how you think we could advance the implementation of AI for the analysis of health data?	● How do you think policy makers should be involved? ● How do you think clinicians should be involved? ● How do you think computer scientists should be involved? ● How do you think ethicists should be involved? ● How do you think patients should be

		involved?
	Which regulatory aspects are important for the implementation of AI in healthcare?	● Which regulations do you think are important? ○ Would you prefer a strong regulatory framework, where different usages and situations are defined and there is a quite clear procedure, or a soft regulatory framework where general usages are defined but the particular practice can still be rather discretional? ● Do you know any regulatory frameworks that are applied? ● Which aspects need a lot of consideration? ● Are there any major concerns that should be addressed by regulations?

c) Vignettes

Cardiology cases (comparison if patients' basal risk change answers - context related questions):

If the interviewee has mentioned another case. Ask them to expand on it and ask the same questions/probes related to their example.

Let's consider a fictional scenario where someone owns a smartwatch. This smartwatch uses artificial intelligence to check the functioning of the heart (like heart rate, respiration rate, saturation, …).

Scenario 1: Jane is 40 years, has no previous diseases, feels healthy.

Scenario 2: Max is 70 years, lives with hypertension and diabetes and feels healthy.

One day there is a pop-up message saying that they have a change in the rhythm of their heart (cardiac arrhythmia - atrial fibrillation).

Additional medical detail (if needed):

Jane has no family history of cardiac disease. Jane has no symptoms. Previous visit to the doctor all results were in normal standards.

Max takes medication for his hypertension and diabetes and has already some signs in previous visits of deteriorating renal function.

The smartwatch that they use has been validated to take ECG (single-lead) to diagnose atrial fibrillation.

	How do you think these patients should react to the pop-up message?	• Do you think that they should believe the pop-up message? What are your thoughts on trust? • What criteria are important for you to trust the results? • Who should trust the results? • How important is this concept for you? Why? • What information or facts would you need to evaluate the suggested diagnosis? • Should Jane and Max visit the doctor? Emergency or request a normal consultation?
	If Max/Jane decides to book an appointment with the doctor, what do you think they should say to their doctor?	• How should they share the data of the smartwatch with the doctors? • Should they mention that the reason for consultation is the smartwatch pop-up message? • How do you think Max/Jane feel about using technology to provide them more information about their health and have the capacity to say this to their doctor?

Vignette to compare other case where the clinical decision is based only on information (comparison of positions between patients and doctors and if type of disease change answers-context related questions):

Now we have a patient named Ruth, 67 years old and she is feeling dizzy. She decides to go to the clinic for a check-up appointment.

The hospital she visits is implementing AI technology and the doctor will be using it during the consultation to support the diagnosis. During the check-up, the doctor adds all the symptoms. The doctor mentions that there is a chance of diabetes and that she is at risk for complications.

Additional medical detail (if needed):

During the medical interview, Ruth mentions the triad of diabetes (thirst, polyuria and increased appetite). Ruth's blood sugar levels are high, also her haemoglobin A1C is high.

	How do you think the doctor should communicate the usage of AI?	• Do you think that the doctor should mention the use of AI to support the diagnosis? • What should the doctor disclose regarding the usage of AI? • How should the doctor explain AI to Ruth? • What other information should the doctor ask, mention or share with Ruth?
	What do you think about Ruth's consent to use AI?	• Do you think consent is necessary? When? • What do you think about using her data? Do you have any concerns about health data handling? • If the context of consent changes and the AI would be suggesting an invasive procedure (e.g. surgery to remove the appendices) would that change any of your previous answers?
	How would you feel about the doctor using AI to support the diagnosis?	• Would it be necessary for the doctor to understand AI? To what degree? • Would it be necessary for the patient to understand AI? To what degree? • What does explainability mean to you? • How would you evaluate explainability?

		• How important is this concept for you? Why? • *Additional medical probes:* - *How would you handle a disagreement between your clinical judgement and the machine's suggestions?* - *How do you feel if the context of this case will be in an emergency situation? Would that change any of your previous answers?* - *How do you think AI should handle medical uncertainty?* - *Would you feel supported by the usage of AI?*
	Do you think using AI during the consultation will affect in any way Ruth's relationship with her doctor?	• How do you think patients will react to the knowledge of doctors using technology to support their decisions? • What would be the advantages and disadvantages to the doctor-patient relationship of using AI during the consultation? • *Additional medical probes:* • *How would you feel about telling patients you are receiving AI support to make clinical decisions?*

Vignette private-public relationship

Let's consider the two scenarios discussed and add that in both cases, the AI/AI used to analyse the data was a private company. For example, Apple would be the one analysing the data with the apple watch.

	What would be your opinion regarding the involvement of private companies?	• What do you think about sharing data from public hospitals with private companies? • Do you have any ethical or legal concerns regarding their

		involvement?

d) Closing questions

	Are there any other practical, medical, and ethical issues that you think are important for the scenario that we have not discussed?	● Is there some concern that you have that was not addressed? ● Do you have any recommendations?

Curriculum Vitae of Giorgia Lorenzini

- Personal information

Via alle scuole 15,
6516 Cugnasco-Gerra (TI)
079 312 16 88
giorgia.lorenzini@unibas.ch or giorgia906@gmail.com
19.12.1996

- Education

2020-2023	University of Basel, Institute for Biomedical Ethics, Basel (CH)
	PhD in Bioethics, legal medicine and health policy
2018-2020	Catholic University, Milano (IT)
	Master in Ethics, bioethics & anthropology
2019	KU Leuven, Leuven (BE)
	Erasmus program, Faculty of arts
2015-2018	Catholic University, Milano (IT)
	Bachelor in Philosophy
2017	Utrecht University, Utrecht (NE)
	Summer School, Project management & Intercultural communication

- Work experience

2020-2023	University of Basel, Institute for Biomedical Ethics, Basel (CH)
	Research assistant
2022-2023	Scuole speciali cantonali del sopraceneri
	Teacher
2020	Sezione degli Enti Locali
	Intern Scientific Collaborator

- Teaching at University of Basel

FS 2020-2021 Contemporary debates: Ethics of medical artificial intelligence

- Publications in peer-reviewed journals

1) Arbelaez Ossa L, Starke G, **Lorenzini G**, Vogt JE, Shaw DM, Elger BS. (2022) Re-focusing explainability in medicine. *DIGITAL HEALTH*. doi:10.1177/20552076221074488
2) **Lorenzini G**, Shaw DM, Arbelaez Ossa L, Elger BS. (2022) Machine learning applications in healthcare and the role of informed consent: Ethical and practical considerations. *Clinical Ethics*. https://doi:10.1177/14777509221094476
3) **Lorenzini G**, Shaw DM & Elger BS. (2022) It takes a pirate to know one: ethical hackers for healthcare cybersecurity. *BMC Med Ethics* 23, 131. https://doi.org/10.1186/s12910-022-00872-y
4) Arbelaez Ossa L, Rost M, **Lorenzini** G, Shaw DM, Elger BS. (2023) A smarter perspective: Learning with and from AI-cases. *Artificial Intelligence in Medicine* 135. https://doi.org/10.1016/j.artmed.2022.102458
5) **Lorenzini G**, Arbelaez Ossa L, Shaw DM, Elger BS. (2023) Artificial Intelligence and the Doctor- Patient Relationship: Expanding the paradigm of shared decision-making. *Bioethics*. 37, 424-429. https://doi.org/10.1111/bioe.13158
6) Arbelaez Ossa L, **Lorenzini G**, Milford SR, Shaw DM, Elger BS, Rost M. (2024). Integrating ethics in AI development: a qualitative study. *BMC Med Ethics* 25, 10 https://doi.org/10.1186/s12910-023-01000-0
7) Thouvenin F, Elger B, Shaw D, **Lorenzini G**, Arbelaez Ossa L, Mätzler S. (2024) Aufklärung beim Einsatz von KI-Systemen in der medizinischen Behandlung. *Jusletter 29* 10.38023/786b14a4-8ce8-452d-a15e-8bd557b0cbcd
8) Milford SR, **Lorenzini G**. (2024). The Hostile Hospital: Exploring Hospitality, Violence, and the Doctor-Patient Relationship. *Journal of Multidisciplinary Healthcare*, 17, 3971–3979. https://doi.org/10.2147/JMDH.S389728
9) **Lorenzini G**, Arbelaez Ossa L, Milford S, Elger B, Shaw D, De Clercq E. (2024) The "Magical Theory" of AI in Medicine: Thematic Narrative Analysis. *JMIR AI 2024*;3:e49795 https://ai.jmir.org/2024/1/e49795

- Presentations at scientific conferences

1) Artificial Intelligence Techniques for Tackling the COVID-19 Pandemic. Herbst-Seminar Medizinethik SGBE, Bigorio (CH). 18-20.11.2021.
2) It Takes a Pirate to Know One: Ethical hackers for healthcare cybersecurity. Postgraduate Bioethics Conference, Bristol (UK). 14-15.07.2022.
3) It Takes a Pirate to Know One: Ethical hackers for healthcare cybersecurity. World Bioethics Conference, Basel (CH). 20-22.07.2022.

4) Artificial Intelligence and the Doctor-Patient Relationship: Expanding the paradigm of shared decision-making. Workshop EPFL: To Trust or Not To Trust?, Lausanne (CH), 07-09.09.2022.
5) Artificial Intelligence and the Doctor-Patient Relationship: Expanding the paradigm of shared decision-making. Human-Machine Collaboration, Paris (FR). 01-02.12.2022.

- Other information

Languages	Italian	C2
	English	C2
	French	C1
	German	B1

IT Skills
MS Office
Adobe Acrobat, Illustrator, Photoshop, Spark, Premiere & InDesign
MAXQDA
Zotero
Social media

Giorgia Lorenzini, September 2023