Machine Learning-Induced Epistemic Injustice in Medicine and Healthcare

Pozzi, G.

DOI
10.4233/uuid:78be5850-2df9-40fe-973d-e537d0d172c0

Publication date
2024

Document Version
Final published version

Citation (APA)
Pozzi, G. (2024). *Machine Learning-Induced Epistemic Injustice in Medicine and Healthcare*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:78be5850-2df9-40fe-973d-e537d0d172c0

Machine Learning-Induced Epistemic Injustice in Medicine and Healthcare

Giorgia Pozzi

Machine Learning-Induced Epistemic Injustice in Medicine and Healthcare

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus Prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Thursday 11 April 2024 at 12:30 o'clock

by

Giorgia Pozzi

Master of Arts in Philosophy,
Ludwig-Maximilians-Universität München, Germany,
born in Cesena, Italy.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof. dr. M. J. van den Hoven	Delft University of Technology, promotor
Dr. J. M. Durán	Delft University of Technology, copromotor

Independent members:

Prof. dr. S. Roeser	Delft University of Technology
Prof. dr. F. Russo	Utrecht University
Prof. dr. ir. I. R. van de Poel	Delft University of Technology
Dr. K. R. Jongsma	University Medical Center Utrecht
Prof. dr. ir. B. Taebi	Delft University of Technology (reserve member)

Keywords: ethics of AI, epistemology of AI, machine learning-induced epistemic injustice, trustworthy AI, medical machine learning, automated hermeneutical appropriation

Printed by: Ridderprint www.ridderprint.nl

Cover picture: Lustrator/Shutterstock.com

A Raffaella, Giuseppe e Sofia - miei genitori amorevoli e sorella adorata

An Stefan - meine große Liebe und Partner fürs Leben

Contents

Acknowledgements

I feel incredibly lucky to have had the opportunity to dedicate the past years to topics of great interest to me while working in an extremely stimulating research environment. I have many people to thank who contributed to rendering my Ph.D. trajectory a worthwhile and profoundly enriching time both from a professional and personal point of view.

First of all, I would like to express my gratitude to my supervisors Juan M. Durán and Jeroen van den Hoven. Juan, thank you so much for believing in me and for your encouragement during my Ph.D. Thank you also for transmitting to me your enthusiasm for the philosophy of science and epistemology of ML and for teaching me so much about these topics during these years. Jeroen, thank you for your guidance and for being an inspiring example to follow. I am very grateful to you for always finding the time to meet up with me and for giving me a reassuring feeling that I was on the right path.

I would like to express my sincere gratitude to the members of my defence committee, Karin Jongsma, Sabine Roeser, Federica Russo, Ibo van de Poel, and Behnam Taebi. I greatly appreciate your time to read my dissertation and share your expertise with me.

A warm thank you belongs to all the amazing colleagues within the Ethics/ Philosophy of Technology Section and particularly the members of the Delft Digital Ethics Centre. I am especially grateful to Jonne Maas and Ryan Timms for being my paranymphs. A very special thank you belongs to Jonne—you have been the best companion and my first reference throughout my Ph.D. journey. Thank you for always being there for me. I am also thankful to Lavinia Marin for being a great Ph.D. buddy. A big thank you goes to Émile

Chappin for being an exceptionally supportive mentor. I am also very thankful to the secretaries of our section, Lissy Mayer, Monica Natanael, and Sophia Wessels, for their help. I am particularly grateful to Nathalie van den Heuvel for being incredibly supportive in any organizational matter from the very beginning of my Ph.D. I would also like to thank all the colleagues within the SoBigData++ research infrastructure I had the opportunity to work with during my Ph.D. I am particularly grateful to Francesca Pratesi, who has been a great support during my collaboration with the project. A big thank you goes also to Michiel De Proost. Thank you, Michiel, for being a great collaborator and co-author. I profited very much from the many discussions we had on epistemic injustice in medical AI.

I am extremely thankful to my former Master's supervisor, Fiorella Battaglia. Fiorella, I will never forget our time together in Munich and how much you taught me about topics that later became central to my Ph.D. research. Thank you for being an important reference in my life and for having always shown me the best sides of academia and research.

I am also very grateful to my dear friend and former roommate, Yasmine. You managed to make the first year of my Ph.D. working from home in Munich special. Even though we are now geographically separated, I am extremely thankful for our friendship.

Finally, my absolutely deepest gratitude belongs to the four most important people in my life who fill it every day with love and support beyond measure. I am profoundly grateful to my parents, Raffaella and Giuseppe, and my baby sister, Sofia. You have welcomed every life choice I made, supporting me in the pursuit of my dreams, even if this has meant being (at least physically) apart from each other. Because we are so incredibly close, I will never get used to being away from you. But I feel the warmth of your affection every day despite the distance between us. Mamma e babbo, you are my most significant source of inspiration, and making you happy by achieving this goal means everything to me. Thank you for being so wonderful. Sofia, you are my best friend in the

whole world. Thank you for allowing me to lean on your shoulder every time I need to. This achievement is also yours.

Words cannot express my gratitude to my soon-to-be-husband, Stefan. We shared every second of our Ph.D. journey with each other working side by side in very small spaces, first in Munich and later in Delft. With your loving way of caring for me, you have managed to always dissipate my concerns. Every joy along the way did not feel like one until I could share it with you. Thank you for being my most profound reason to be happy.

Summary

As knowledge-intensive fields, medicine and healthcare have been central to the artificial intelligence (AI) revolution in recent decades. Many tasks that are part of physicians' daily activities require collecting, elaborating, and analyzing a large amount of information. The advancement of data science technologies, such as machine learning (ML) systems, can improve healthcare provision through the application of such technologies in evaluating health risks and providing, among others, reliable treatment recommendations, diagnoses, and prognoses. ML systems can recognize patterns in vast amounts of data points at high speed. Thus, they can be used to efficiently fulfill relevant epistemic tasks, such as classifications and predictions, which are essential activities in modern-day medical practice.

As many ML systems are black boxes, the decision-making logic underlying the outputs they produce is inaccessible to the human investigator (e.g., a physician deciding about whether to follow a system's recommendation in a particular case). Epistemic limitations of ML systems, particularly in terms of their epistemically opaque nature, have given rise to a series of ethical issues that are at center stage in the current debate revolving around the ethics and epistemology of ML. The issues are related to responsibility attribution, moral justification for acting upon the recommendation of an ML system, algorithmic fairness and discrimination, privacy, and whether these systems are worthy of our trust, among many others. These concerns are particularly imperative given the moral salience of medical interactions in which vulnerabilities and a clear set of moral values and demands are in place.

Other than the urgency of the concerns mentioned above, there is another central issue at the intersection of the ethics and epistemology of ML that has been largely neglected. This pertains to the careful scrutiny of how ML systems can degrade individuals' epistemic standing as receivers and conveyors of knowledge and, thereby, perpetrate epistemic injustice. This phenomenon can assume different forms, including silencing, marginalization, unfair distrust, and withholding credibility, among others. A prolific body of literature in social epistemology broadens Miranda Fricker's original framework (comprehending two forms of epistemic injustice—testimonial and hermeneutical injustice) and applies it to the field of medicine and healthcare. The discussions have, so far, been analyzed exclusively in human-human contexts (e.g., patient-physician interactions).

However, the increasingly relevant role of ML systems in healthcare urges us to investigate the role they play in bringing about various forms of epistemic injustice. Since ML systems are powerful epistemic entities that are not easily contestable, and their decision-making rationale is often inaccessible, it is crucial to consider their role in creating imbalances in patients' disfavor and the ways to mitigate these imbalances. This is especially important when it comes to interactions between patients and physicians, in which questions of credibility, trust, and understanding are central.

Against this background, the overarching purpose of this dissertation is to fill the discussed research gap by providing a framework to identify and mitigate epistemic injustices that are *ML-induced*, i.e., that emerge specifically due to the role that ML systems play in patient-physician interactions. Simply applying Fricker's original framework of epistemic injustice to ML cases would not be enough to account for the novel challenges these systems raise. Therefore, in this dissertation, I have adopted a unique approach by extending the framework of epistemic injustice to include ML as an additional epistemic entity supporting central medical activities (e.g., diagnoses, prognoses, treatment recommendations, and mental health support). In more general terms,

I show that ethical issues in ML extend beyond the widely discussed issues of privacy, bias, and responsibility attribution. The problems related to epistemic injustice deserve separate scrutiny since they are not reducible to any of the issues that are currently receiving extensive attention in debates on the ethics and epistemology of AI. Accordingly, my framework demonstrates that there are currently overlooked and highly problematic forms of injustice that ML systems entail, which should be tackled to increase the epistemic, moral, and social acceptability of ML systems in medicine.

Epistemic injustice can have different forms—some forms are more explicitly recognizable, while others lurk in the background of our social practices. In addition to the various theoretical specifications that are discussed in the course of this dissertation, epistemic injustice generally pertains to all practices that constrain the epistemic activities of individuals, e.g., practices that limit the possibility of transmitting and receiving information and pose disadvantages in making their lived experiences accessible to others. As I explicate in this dissertation, ML systems can exacerbate these forms of injustice. Members of disadvantaged social groups often experience such ethically problematic epistemic harms due to discriminatory practices and prejudices related to their social identity (e.g., race, gender, ability, and social status). Since the unjust mechanisms for unrecognition of what an individual can contribute to an epistemically relevant enterprise are fueled by power relations and structural inequalities, they often tend to go unnoticed.

Consider a simple case: a medical interaction in which a patient communicates her symptoms to a physician. In this situation, the patient plays the role of an epistemic subject because she conveys knowledge and information to her physician that is relevant for further course of action. If the physician dismisses this patient's testimony as irrelevant based on unfounded prejudices related to her gender (e.g., that women are apt to complain, more so than men), the patient is wronged in her status as a knowing subject (i.e., as a receiver and conveyor of knowledge and information). In fact, it seems undeniable that

her part as a conveyor of information and knowledge requires her to be taken seriously in her self-reporting of her symptoms. In this case, she is the victim of epistemic injustice (more precisely, of a testimonial injustice), according to the standard framework, because she was constrained in her participation in the medical interaction with her physician on account of her testimony (i.e., the self-report of her symptoms) being dismissed on unfair grounds.

Now, consider a situation in which an ML system produces an output (e.g., in the form of a risk score evaluating patients' likelihood of drug abuse) that contradicts the patient's testimony. Assume further that, on the basis of this risk score, the physician attributes, by default, more credibility to the ML output than to what the patient is telling them. Consequently, the physician dismisses the patient's testimony and takes action based on the ML risk score, for example, by withholding the prescription of a certain medication. Regardless of whether the score is accurate in a particular case, the system is central to how the physician forms an assessment of the patient's credibility. The epistemic authoritative role ML systems play in medical encounters between patients and physicians can, thus, be disadvantageous to patients by devaluing their testimony. This sketchy situation points to the importance of analyzing how relevant epistemic attitudes (such as credibility and trust) in medical relationships are affected by ML systems. The concrete case of an ML system outputting patients' risk of misusing opioids hinges exactly on similar dynamics. Later in the dissertation, I examine this case in detail and provide a framework to capture what I have labeled as *ML-induced* testimonial and hermeneutical injustice.[1]

In light of the above considerations, the main research question that this dissertation aims to answer is: *In which ways does machine learning-induced epistemic injustice in medicine and healthcare emerge, and how can it be mitigated?* This question is tackled in two steps, which are presented in the two main parts of this dissertation. Below, I briefly summarize the contents of

[1] Chapters 4 and 5 are dedicated to the development of the mentioned framework.

both parts and also describe the main focus of the relevant chapters.

Part 1 consists of Chapters 2 and 3 and presents a discussion of fundamental concepts at the intersection of the ethics and epistemology of ML geared toward making explicit two main research gaps (that emerge, respectively, in Chapters 2 and 3). I show that these research gaps lead to overlooking problems of epistemic injustice in ML and thus need to be addressed in a timely manner.

Chapter 2 provides a literature review of a concept that figures prominently in the analysis of epistemic injustice: trust. This chapter lays the groundwork for the conditions for diverse forms of trustworthy AI. Notably, from this chapter, it emerges that the specific question of whether it is possible and/or desirable to *directly* trust ML has received considerable scholarly attention. However, and crucially in relation to the overall project of this dissertation, it also emerges that the role ML plays in *mediating* trust relationships between patients and physicians has been largely neglected. This is the first research gap this thesis aims to bridge. Inquiring into the role ML systems play, as an additional epistemic authority entering patient-physician encounters and affecting trust relations between them, is central to an analysis of epistemic injustice. Hence, Chapter 2 lays the groundwork for identifying epistemic injustice in ML characterized by trust dysfunctions (this is explored in detail in Chapter 4).

Chapter 3 highlights a second research gap at a more general level of analysis, i.e., considering how the debate treats the general relationship between the epistemology and ethics of ML. I show that considering these two dimensions as unrelated leads to overlooking the role that ethical properties should play in regulating central epistemological functions in ML (such as explanations). In turn, this creates conditions for the emergence of epistemic injustice, in the form of epistemic objectification, in medical ML. The phenomenon of epistemic objectification in ML differs from the conceptualization of epistemic objectification in the standard epistemic injustice framework. Epistemic objectification in ML will be tackled in two steps. First, I argue that the current

approach to the ethics and epistemology of ML is unidirectional. Second, and as a consequence, I explain that a medical ML cannot pick up on a patient's values, thus constraining their contribution to the medical discourse. In Chapter 3, I address both points by providing the analytic tools to capture epistemic objectification, specifically in medical ML. This is a fundamental step in the development of a framework for epistemic injustice in medical ML. I dedicate the second part of this dissertation to its further expansion.

The second part of the thesis includes Chapters 4 to 7. In Chapter 4, I advance the conceptualization of ML-induced testimonial injustice. This analysis is crucial to show that ML systems can decrease patients' credibility for epistemically invalid reasons, thus lowering the relevance of their testimonies and inflicting morally significant harm on them. As previously indicated, I apply my framework to the concrete and worrisome case of an ML system in medicine currently implemented throughout the USA to predict patients' likelihood of developing opioid addiction or misuse (PDMPs). I analyze the role of the system in mediating the patient-physician relationship by scrutinizing its impact on physicians' assessments of patients' credibility and trustworthiness. Drawing on the conceptual analysis proposed in Chapter 2, I define ML-induced testimonial injustice, specifically in cases where these systems are treated as markers of trustworthiness.

In Chapter 5, I continue with the expansion of the complementary part of the epistemic injustice framework by conceptualizing ML-induced hermeneutical injustice. I focus on the position of a person who has been mistakenly classified as being at a high risk of opioid misuse by the system previously described in Chapter 4. I formulate three conditions for capturing ML-induced hermeneutical injustice. Of particular relevance is the analysis of how ML systems reshape medically grounded concepts (such as substance use disorder) in a way that eludes human scrutiny. The formulation of these conditions captures the first part of my framework for ML-induced hermeneutical injustice to identify injustices occurring at patients' expense. However, hermeneutical

injustice in ML can also impact epistemically well-positioned agents, such as physicians, and this possibility is usually not considered in standard debates. I coin and advance the novel concept of *automated hermeneutical appropriation* to capture the epistemic harm suffered by these epistemic agents. This represents the second part of my framework of ML-induced hermeneutical injustice advanced in this chapter.

Chapter 6 examines a form of epistemic injustice that is not part of Fricker's original framework: participatory injustice. This concept aims to capture broader forms of epistemic harm that unjustly constrain subjects in central epistemic activities such as conjecturing, making hypotheses, advancing one's understanding of oneself and the world around us, refusing certain beliefs, and creating new ones, among many others. To make the analysis of this form of injustice more graspable and to depict its urgency in the context of conversational AI (CAI), the case of a mental health Chatbot, Karim, is discussed. This chatbot was developed to deliver mental health support to vulnerable populations such as Syrian refugees. This case substantiates the epistemic and ethical concerns arising from the use of mental health applications, specifically among vulnerable populations. Notably, this chapter offers my perspective on ameliorating the form of epistemic injustice discussed. In this regard, I argue that epistemic participation can be conceptualized as a capability that can be accounted for within the framework of Capability Sensitive Design. The merit of this mitigation strategy for participatory injustice in CAI is that it provides designers and developers with the tools needed to inquire into the potential of a CAI to cause this form of injustice. This is thus useful to anticipate and counteract previously undetected consequences deriving from hindering users' epistemic participation.

Drawing on the framework provided in the previous chapters, Chapter 7 explores a further aspect connected to epistemic injustice that is specifically related to patients' social situatedness. This chapter discusses questions related to disease causation that are usually not considered in the social epistemol-

ogy literature on epistemic injustice or in the standard philosophy of science discussions revolving around causality. The study I present in this chapter is novel because I show that the causes sought after in medical practice (i.e., either only biological or bio-social) are connected with the sources of information and knowledge (such as patients' testimony) admitted in the medical discourse. Consequently, accounts of disease causation that consider biological *and* social factors as causally relevant at the individual level are central to permitting patients' testimonial contributions in the medical discourse.

In Chapter 8, I conclude this thesis with my final remarks. Here, I explicate the social relevance of the arguments provided and point out some aspects of the framework presented that require further attention. I also lay out prospects for future research building upon the framework of ML-induced epistemic injustice I proposed in this dissertation.

Samenvatting

Geneeskunde en gezondheidszorg zijn kennisintensieve vakgebieden en staan al tientallen jaren in het middelpunt van de kunstmatige intelligentie (AI)-revolutie. Veel taken die deel uitmaken van de dagelijkse activiteiten van artsen vereisen het verzamelen, uitwerken en analyseren van een grote hoeveelheid informatie. Vooruitgang in data science-technologieën, zoals machine learning (ML)-systemen, kunnen de gezondheidszorg verbeteren door onder andere gezondheidsrisico's te evalueren en betrouwbare behandelingsaanbevelingen, diagnoses en prognoses te doen. ML-systemen zijn in staat om patronen in enorme hoeveelheden gegevenspunten razendsnel te herkennen, waardoor deze systemen efficiënt zijn in het uitvoeren van relevante epistemische taken. Hieronder vallen classificaties en prognoses, die essentieel zijn in de hedendaagse medische praktijk.

Omdat veel van deze systemen echter black boxes zijn, is de besluitvormingslogica die ten grondslag ligt aan de output die ze produceren ontoegankelijk voor de menselijke onderzoeker (bijv. artsen die moeten beslissen of ze de aanbeveling van het systeem in een bepaald geval opvolgen). Kennisbeperkingen van ML-systemen, vooral met betrekking tot hun epistemisch ondoorgrondelijke aard, geven aanleiding tot een reeks ethische kwesties die centraal zijn komen te staan in het huidige debat rond de ethiek en epistemologie van ML. Dit zijn onder andere vragen over het toewijzen van verantwoordelijkheid, morele rechtvaardiging voor het handelen naar een aanbeveling van een ML, algoritmische eerlijkheid en discriminatie, privacy en of deze systemen ons vertrouwen waard zijn. Deze aspecten zijn vooral belangrijk gezien de morele waarde van medische interacties waarin kwetsbaarheden en een duidelijke reeks

morele waarden en eisen aanwezig zijn.

Naast de noodzaak van de bovengenoemde kwesties, wordt een andere centrale kwestie op het snijvlak van ML-ethiek en epistemologie grotendeels verwaarloosd. Het zorgvuldige onderzoek naar hoe ML-systemen de epistemische status van individuen als ontvangers en overbrengers van kennis kunnen aantasten door het plegen van *epistemisch onrecht* is in het huidige debat over het hoofd gezien. Dit fenomeen kan verschillende vormen aannemen en kan onder andere neerkomen op het tot zwijgen brengen, marginalisatie, oneerlijk wantrouwen of het onthouden van geloofwaardigheid. Een grote hoeveelheid literatuur in de sociale epistemologie past het oorspronkelijke raamwerk van Miranda Fricker (dat twee vormen van epistemisch onrecht omvat, namelijk getuigenis- en hermeneutisch onrecht) toe op het gebied van de geneeskunde en de gezondheidszorg. Deze discussies zijn tot nu toe uitsluitend geanalyseerd in een mens-mens context (bijv. interacties tussen patiënten en artsen).

De steeds relevantere rol van ML-systemen in de gezondheidszorg dwingt ons echter om hun rol in het tot stand brengen van vormen van epistemisch onrechtvaardigheid te onderzoeken. Aangezien ML-systemen krachtige epistemische entiteiten zijn, die nauwelijks betwistbaar zijn en vaak ontoegankelijk in hun besluitvormingsprincipes, is het van cruciaal belang om na te denken over hun rol in het creëren van onevenwichtigheden in de afkeer van patiënten en manieren om deze te verminderen; zelfs nog meer als het gaat om interacties tussen patiënten en artsen waarin vragen over geloofwaardigheid, vertrouwen en begrip centraal staan.

Tegen deze achtergrond is het overkoepelende doel van dit proefschrift om deze onderzoeksleemte op te vullen door een raamwerk te bieden voor het identificeren en verminderen van epistemische onrechtvaardigheden die *ML-geïnduceerd* zijn, d.w.z. die specifiek voortkomen uit de rol die ML-systemen spelen in interacties tussen patiënten en artsen. Het simpelweg toepassen van het oorspronkelijke kader van epistemisch onrecht op ML-zaken zou niet genoeg zijn om de nieuwe uitdagingen van deze systemen te verklaren. De

benadering die ik in dit proefschrift hanteer is uniek omdat het de eerste is die het raamwerk van epistemische onrechtvaardigheid uitbreidt met ML als een verdere epistemische entiteit die centrale medische activiteiten ondersteunt (bijv. diagnoses, prognoses, aanbevelingen voor behandelingen, ondersteuning van de geestelijke gezondheid). In meer algemene termen laat ik zien dat ethische kwesties in ML niet alleen neerkomen op veelbesproken problemen in termen van privacy, vooringenomenheid of toewijzing van verantwoordelijkheid. Problemen van epistemische onrechtvaardigheid verdienen een apart onderzoek omdat ze niet kunnen worden teruggebracht tot een van de kwesties die momenteel veel aandacht krijgen in discussies over de ethiek en epistemologie van AI. Mijn raamwerk laat dus zien dat er momenteel over het hoofd geziene en zeer problematische vormen van onrechtvaardigheid bestaan die door ML-systemen worden geproduceerd en die aangepakt moeten worden om de epistemische, morele en sociale aanvaardbaarheid van ML-systemen in de geneeskunde te vergroten.

Epistemische onrechtvaardigheid kan verschillende vormen aannemen, sommige expliciet herkenbaar terwijl andere op de achtergrond van onze sociale praktijken blijven hangen. Naast verschillende theoretische specificaties die in de loop van dit proefschrift besproken worden, verwijst epistemische onrechtvaardigheid in het algemeen naar alle praktijken die individuen beperken in hun epistemische activiteiten (bijv. het beperken van hun vermogen om informatie door te geven en te ontvangen, nadelen bij het toegankelijk maken van hun geleefde ervaringen voor anderen). In dit proefschrift laat ik zien dat ML-systemen deze vormen van onrechtvaardigheid kunnen verergeren. Personen uit achtergestelde sociale groepen ervaren deze ethisch problematische epistemische schade vaak als gevolg van discriminerende praktijken en vooroordelen met betrekking tot hun sociale identiteit (bijv. ras, geslacht, capaciteiten, sociale status). Aangezien de onrechtvaardige mechanismen van het niet erkennen van wat een individu kan bijdragen aan een epistemisch relevante onderneming gevoed worden door machtsverhoudingen en structurele ongelijkheden,

blijven ze vaak onopgemerkt.

Beschouw een eenvoudig geval: een medische interactie waarbij patiënten hun klachten doorgeven aan artsen. In deze situatie spelen patiënten een rol als epistemische subjecten, omdat ze hun artsen kennis en informatie geven die relevant is voor de daaropvolgende behandeling. Stel dat deze hypothetische artsen de getuigenissen van deze patiënten als irrelevant afwijzen op basis van ongegronde vooroordelen over hun geslacht (bijv. dat vrouwelijke patiënten de neiging hebben om meer te klagen dan mannelijke patiënten). In dit geval wordt hun status als kennend subject (d.w.z. als ontvangers en overbrengers van kennis en informatie) onrecht aangedaan. Het lijkt zelfs onbetwistbaar dat een deel van hun rol als overbrenger van informatie en kennis vereist dat men hen serieus neemt in hun zelfrapportage van hun symptomen. In het beschreven geval zijn deze personen het slachtoffer geweest van een epistemisch onrecht (meer precies, van een getuige-ongerecht) volgens het standaardkader. Dit komt omdat ze niet konden deelnemen aan de medische interactie met hun artsen omdat hun getuigenis (d.w.z. zelfrapportage van hun symptomen) ongegrond werd afgewezen.

Beschouw nu een situatie waarin een ML-systeem een uitvoer produceert (bijv. in de vorm van een risicoscore die de waarschijnlijkheid van drugsmisbruik bij patiënten evalueert) die de verklaring van de patiënten tegenspreekt. Stel verder dat de artsen, op basis van de eerder genoemde risicoscore, standaard meer geloofwaardigheid toekennen aan de ML-uitvoer dan aan wat patiënten hen vertellen. De artsen verwerpen dan de verklaring van de patiënten en ondernemen actie op basis van de ML-risicoscore (bijv. door een bepaald medicijn niet voor te schrijven). Ongeacht of de score in een bepaald geval accuraat is of niet, het systeem staat centraal in de manier waarop artsen oordelen over de geloofwaardigheid van patiënten. De epistemische autoriteitsrol die ML-systemen spelen in medische ontmoetingen tussen patiënten en artsen kan patiënten dus benadelen bij het uitdrukken van de waarde van hun getuigenis. Deze schetsmatige situatie laat al zien hoe belangrijk het is om te

analyseren hoe relevante epistemische houdingen (bijv. geloofwaardigheid en vertrouwen) in medische relaties worden beïnvloed door ML-systemen. Het concrete geval van een ML-systeem dat het risico op misbruik van opioïden door patiënten weergeeft, berust precies op een soortgelijke dynamiek. Ik onderzoek dit geval in detail en geef een kader voor wat ik *ML-geïnduceerde* testimonial en hermeneutische onrechtvaardigheid[2] noem, vast te leggen.

In het licht van deze overwegingen luidt de belangrijkste onderzoeksvraag die dit proefschrift wil beantwoorden: *Op welke manieren ontstaat door machine learning veroorzaakte epistemische onrechtvaardigheid in de geneeskunde en gezondheidszorg, en hoe kan deze worden verminderd?* Dit gebeurt in twee stappen die de verdeling van dit proefschrift in twee hoofddelen weerspiegelen. Hieronder geef ik een korte samenvatting van de inhoud van beide hoofdstukken, samen met de belangrijkste aandachtspunten.

Deel 1 bestaat uit de hoofdstukken 2 en 3 en presenteert een discussie van fundamentele concepten op het snijvlak van de ethiek en epistemologie van ML, gericht op het expliciet maken van twee belangrijke hiaten in het onderzoek (die respectievelijk in de hoofdstukken 2 en 3 naar voren komen). Ik laat zien dat deze hiaten in het onderzoek leiden tot het over het hoofd zien van problemen van epistemische onrechtvaardigheid in ML en dus tijdig moeten worden aangepakt.

Hoofdstuk 2 geeft een literatuuroverzicht van een concept dat een prominente rol speelt in de analyse van epistemisch onrecht: vertrouwen. Dit hoofdstuk legt de basis voor de voorwaarden voor diverse vormen van vertrouwen in AI en betrouwbare AI. Uit dit hoofdstuk blijkt met name dat de specifieke vraag of *direct* vertrouwen in ML mogelijk en/of wenselijk is, veel wetenschappelijke aandacht heeft gekregen. Maar, en dat is cruciaal voor het algehele project van dit proefschrift, het blijkt ook dat de rol die ML speelt in het *bemiddelen* van vertrouwensrelaties tussen patiënten en artsen grotendeels is verwaarloosd. Dit is de eerste leemte in het onderzoek die deze dissertatie

[2]De hoofdstukken 4 en 5 zijn gewijd aan de ontwikkeling van dit kader.

wil opvullen. Het onderzoeken van de rol die ML-systemen spelen als een extra epistemische autoriteit die de ontmoeting tussen patiënten en artsen beïnvloedt, staat centraal in een analyse van epistemische onrechtvaardigheid. Vandaar dat in hoofdstuk 2 de basis wordt gelegd voor het identificeren van epistemisch onrecht in ML dat wordt gekenmerkt door vertrouwensdisfuncties (deze studie wordt naar voren gebracht in hoofdstuk 4).

Hoofdstuk 3 richt zich op een tweede onderzoekshiaat op een meer algemeen analyseniveau, namelijk hoe het debat de algemene relatie tussen de epistemologie en ethiek van ML behandelt. Ik laat zien dat het beschouwen van deze twee dimensies als ongerelateerd leidt tot het over het hoofd zien van de rol die ethische eigenschappen zouden moeten spelen bij het reguleren van centrale epistemologische functies in ML (zoals verklaringen). Dit schept op zijn beurt voorwaarden voor epistemisch onrecht, in termen van epistemische objectivering in medisch ML. Epistemische objectivering in ML komt niet neer op hetzelfde fenomeen als geconceptualiseerd in het standaard kader voor epistemisch onrecht. Om aan te tonen dat een ML-systeem epistemische objectivering teweegbrengt, zijn twee stappen nodig. Ten eerste beweer ik dat er sprake is van een eenzijdige benadering van de ethiek en epistemologie van ML. Ten tweede, en als gevolg daarvan, maak ik expliciet dat een medische ML de waarden van patiënten niet kan oppikken, waardoor hun bijdrage aan het medische discours wordt beperkt. Om beide punten aan de orde te stellen, bied ik in hoofdstuk 3 de analytische hulpmiddelen om epistemische objectivering specifiek in medische ML vast te leggen. Dit is een fundamentele stap in de ontwikkeling van een raamwerk voor epistemisch onrecht in ML-geneeskunde en gezondheidszorg. Ik wijd het tweede deel van dit proefschrift aan de verdere uitbreiding hiervan.

Het tweede deel van het proefschrift omvat de hoofdstukken 4 tot en met 7. In hoofdstuk 4 ga ik verder in op de conceptualisering van door ML veroorzaakte getuigenisonrechtvaardigheid. Deze analyse is cruciaal om aan te tonen dat ML-systemen de geloofwaardigheid van patiënten kunnen aantasten om

epistemisch ongeldige redenen, waardoor de relevantie van hun getuigenissen afneemt en dus moreel significante schade wordt toegebracht. Zoals eerder aangegeven, pas ik mijn raamwerk toe op het concrete en zorgwekkende geval van een ML-systeem in de geneeskunde dat momenteel in de hele VS wordt geïmplementeerd om de waarschijnlijkheid te voorspellen dat patiënten een opioïdenverslaving of -misbruik ontwikkelen (PDMPs). Ik analyseer de rol van het systeem in het bemiddelen van de relatie tussen patiënten en artsen door de invloed ervan op de beoordeling door artsen van de geloofwaardigheid en betrouwbaarheid van patiënten te onderzoeken. Op basis van de conceptuele analyse uit hoofdstuk 2 definieer ik ML-geïnduceerde getuigenisonrechtvaardigheid, specifiek in gevallen waarin deze systemen worden behandeld als indicator van betrouwbaarheid.

In hoofdstuk 5 ga ik verder met de uitbreiding van het complementaire deel van het kader voor epistemisch onrecht door het conceptualiseren van door ML veroorzaakt hermeneutisch onrecht. Ik richt me op de positie van personen die door het eerder in hoofdstuk 4 beschreven systeem ten onrechte zijn aangemerkt als personen met een hoog risico op misbruik van opioïden. Om ML-geïnduceerd hermeneutisch onrecht te kunnen vastleggen, formuleer ik drie condities. Van bijzonder belang is de analyse van de manier waarop ML-systemen medisch gefundeerde concepten (zoals stoornissen in het gebruik van middelen) een nieuwe vorm geven, waarbij menselijke controle wordt omzeild. De formulering van deze voorwaarden omvat het eerste deel van mijn kader voor ML-geïnduceerde hermeneutische onrechtvaardigheid om onrechtvaardigheden ten koste van patiënten te identificeren. Echter, hermeneutische onrechtvaardigheid in ML kan ook gevolgen hebben voor epistemisch goed gepositioneerde agents zoals artsen, wat meestal niet wordt meegenomen in standaarddebatten. Ik introduceer het nieuwe concept van *automatische hermeneutische toe-eigening* om de epistemische schade te vatten die deze epistemische agenten lijden. Dit is het tweede deel van mijn raamwerk van ML-geïnduceerd hermeneutisch onrecht dat in dit hoofdstuk

naar voren wordt gebracht.

Hoofdstuk 6 onderzoekt een vorm van epistemisch onrecht die geen deel uitmaakt van het oorspronkelijke raamwerk van Fricker: participatief onrecht. Dit concept is bedoeld om bredere vormen van epistemische schade te vatten die onderwerpen onnodig beperken in centrale epistemische activiteiten zoals gissen, hypothesen opstellen, iemands begrip van zichzelf en de wereld om ons heen verbeteren, bepaalde overtuigingen verwerpen en nieuwe creëren, enzovoort. Teneinde de analyse van deze vorm van onrechtvaardigheid begrijpelijker te maken en de urgentie ervan in de context van conversational AI (CAI) aan te tonen, wordt het geval van een Chatbot voor geestelijke gezondheidszorg, Karim, besproken. Deze chatbot is ontwikkeld om psychische hulp te bieden aan kwetsbare bevolkingsgroepen, zoals Syrische vluchtelingen. Deze casus toont de epistemische en ethische problemen aan die voortkomen uit het gebruik van geestelijke gezondheidstoepassingen, vooral bij kwetsbare bevolkingsgroepen. Met name biedt dit hoofdstuk mijn kijk op het verbeteren van de besproken vorm van epistemisch onrecht. Dit wordt gedaan door te stellen dat epistemische participatie geconceptualiseerd kan worden als een vermogen dat verantwoord kan worden binnen het raamwerk van Capability Sensitive Design. De verdienste van deze verzachtende strategie voor participatieve onrechtvaardigheid in CAI is dat het ontwerpers en ontwikkelaars voorziet van de hulpmiddelen die nodig zijn om te onderzoeken of een CAI deze vorm van onrechtvaardigheid kan veroorzaken. Het is daarom zinvol om te anticiperen op eerder onopgemerkte gevolgen die voortkomen uit het belemmeren van de epistemische participatie van gebruikers, en deze gevolgen tegen te gaan.

Op basis van het raamwerk dat in de vorige hoofdstukken werd aangereikt, onderzoekt hoofdstuk 7 een ander aspect dat verband houdt met epistemische onrechtvaardigheid en specifiek gerelateerd is aan de sociale situatie van patiënten. Dit hoofdstuk bespreekt vragen met betrekking tot ziekteoorzaken die gewoonlijk niet aan bod komen in de sociale epistemologische literatuur over epistemische onrechtvaardigheid, noch in standaard discussies in

de wetenschapsfilosofie over causaliteit. Het onderzoek dat ik in dit hoofdstuk presenteer is nieuw omdat ik laat zien dat de oorzaken waarnaar in de medische praktijk wordt gezocht (d.w.z. alleen biologisch of alleen biosociaal) verbonden zijn met de bronnen van informatie en kennis (zoals de getuigenis van patiënten) die in het medische discours worden toegelaten. Bijgevolg zijn rekeningen van ziekteoorzaken die biologische en sociale factoren als causaal relevant op individueel niveau beschouwen, van cruciaal belang om de getuigenisbijdragen van patiënten aan het medische discours mogelijk te maken.

In hoofdstuk 8 sluit ik deze scriptie af met slotopmerkingen. Hier maak ik de maatschappelijke relevantie van de gegeven argumenten duidelijk en wijs ik op enkele beperkingen van het gepresenteerde kader. Ik schets ook vooruitzichten voor toekomstig onderzoek dat voortbouwt op het raamwerk van ML-geïnduceerde epistemische onrechtvaardigheid dat ik in dit proefschrift heb gegeven.

1

Introduction

1.1. The ethical nature of epistemic activities

Many activities performed on a daily basis involve receiving, elaborating, and conveying information. To advance our knowledge of the world, we often rely on information obtained from people we consider credible. Conversely, we are reluctant to believe information received from someone we deem an untrustworthy informant. Often justifiably so. For instance, if someone has proven deceitful on several occasions, one probably has good reason to doubt their trustworthiness. Since it is desirable to accept information exchanges that are most probably truth-conducive (Lehrer, 2006), the assessment of people's credibility is an important part of the activities that pertain to forming beliefs and acquiring knowledge and understanding, that is to our *epistemic* activities.

Yet, such an assessment requires particular caution. As Miranda Fricker (2007) compellingly points out, unjustifiably depriving someone of credibility can prove exceptionally harmful. The importance of being acknowledged as

a trustworthy communicator and sharer of knowledge varies according to the stakes. In socially sensitive fields, such as judicial settings and the workplace, questions of credibility and trustworthiness acquire a particularly relevant role. For example, in the work environment, it is important that one's ideas are taken into account and not dismissed too hastily without being given due consideration. The same applies also to healthcare settings, which are the contexts of interest to this dissertation. When we share health-related concerns with our physicians, we want to feel heard and understood. This means that there are not only (rather explicit and widely discussed) ethical demands pertaining to, for instance, patients' right to privacy and non-discrimination. There are also ethical demands pertaining to the knowledge patients are entitled to share and receive within medical interactions. Relatedly, situations in which patients have the feeling of not being taken seriously about their health issues can lead to problematic outcomes.

Consider, for instance, a medical interaction in which a woman's pain is underestimated by a physician based on the wrongful belief that women are often oversensitive and apt to complain, more so than men (Carel & Kidd, 2014; Kidd & Carel, 2017). It seems obvious that this is not a convincing reason why this person deserves less credibility when she reports her symptoms. If this hypothetical physician acted upon this ill-grounded belief by refusing to provide her with the needed medication, it would cause feelings of outrage and indignation. Other than the practical consequence of her being undertreated and having to bear (possibly) unnecessary pain, it is problematic that her credibility, and as such, her epistemic standing, is unjustifiably questioned. Outrage and indignation are genuinely moral emotions that intuitively show the close entanglement between epistemology and ethics in situations such as the one described above.

Social epistemology is particularly well-suited for the evaluation of the profoundly ethical nature of epistemic activities. Feminist epistemologists, particularly, teach us that it is paramount to consider epistemic subjects and their

activities as embedded in a specific social context in which social identities (e.g., race, gender, and ability), power relations, conditions of systemic oppression, and knowledge asymmetries play a crucial role (Demir-Doğuoğlu & McLeod, 2023; Kidd et al., 2017). The ethical dimension of epistemic activities comes to the forefront through the recognition that knowledge is *socially situated*: in other words, epistemic activities are not carried out in a vacuum but, rather, in a particular social context. Consequently, the latter has a crucial bearing on individuals' epistemic role because their social identity considerably shapes their possibility of sharing and acquiring knowledge. Often, interactions that entail sharing and conveying knowledge turn out to be profoundly unjust. For example, in the hypothetical medical interaction described above, the patient's social identity (i.e., her gender) is connected to an unfair and, thus, morally problematic withdrawal of credibility. From these considerations emerges the closely intertwined relationship between social positioning (e.g., gender identity), epistemic activities (e.g., conveying and receiving information), and ethical concerns (e.g., injustices).

The exceptionally strong emphasis on the social situatedness of epistemic subjects allows us to look at the bigger picture, in which wrongfully denying a person's epistemic standing (e.g., by unjustifiably undermining her credibility and unfairly withholding trust) amounts to an injustice. Fricker provides the widely discussed theoretical framework of *epistemic injustice* to capture forms of injustice people suffer, specifically in their capacities as knowers, i.e., as receivers and conveyors of knowledge (Fricker, 2007). The author conceptualizes two forms of epistemic injustice—*testimonial* and *hermeneutical* injustice. Testimonial injustice is defined as the unjustified undermining of a person's credibility that occurs at the interpersonal level between two or more interlocutors. The hypothetical medical interaction just described exemplifies this form of injustice. Hermeneutical injustice, on the other hand, occurs at a prior stage and amounts to the unavailability of the conceptual resources needed to make sense of one's social experience. To exemplify this form of injustice, Fricker

considers the situation of a woman experiencing sexual harassment at a time before this concept was collectively available. In this case, this person cannot make sense of her experience or effectively communicate it to others due to a gap in the collective conceptual resources (Fricker, 2007).[1] In the extensive body of literature on epistemic injustice, other forms of it are conceptualized, expanding Fricker's original framework.[2] Above and beyond specifications of different forms of epistemic injustice, this phenomenon generally pertains to all the practices that constrain individuals in their epistemic activities, such as silencing, epistemic violence, exclusion from communicative practices, epistemic marginalization, and unfair distrust, to mention a few (Kidd et al., 2017).

Since the occurrence of epistemic injustice is highly contextual, this phenomenon has been analyzed according to the specificities of different fields. As previously mentioned, epistemic injustices emerging in medicine and healthcare are of particular interest to this thesis. In these contexts, several authors have successfully shown that patients are often prone to epistemic harm due to several factors that characterize healthcare encounters. Examples are gender and racial prejudices that undermine patients' epistemic standing (Carel & Kidd, 2014), and the epistemic privilege assigned to physicians due to existing knowledge asymmetries (Kidd & Carel, 2017). Moreover, conceptual difficulties can emerge due to, for instance, the ineffability of certain experiences that particular categories of patients encounter (Carel & Kidd, 2017); such difficulties create hermeneutical gaps, among other issues. These and other epistemic dysfunctions can illegitimately hinder patients' epistemic contributions to medical decision-making in ethically problematic ways.

Notably, these debates restrict their focus on medical interactions among human agents (e.g., patients and healthcare professionals such as physicians and nursing personnel). Nevertheless, the revolution of artificial intelligence (AI) in the last few decades has shown that medicine and healthcare are fields

[1] I analyze in great detail both forms of epistemic injustice in the following chapters.
[2] See, for example, Hookway's conceptualization of participatory injustice (Hookway, 2010). I discuss this form of epistemic injustice in depth in Chapter 6.

no longer limited to human expertise. AI-based technologies play an increasingly relevant role in crucial medical practices, such as diagnoses, prognoses, treatment recommendations, drug discovery, and others (Esteva et al., 2019; Topol, 2019). As such, AI systems have entered the medical sphere as entities mediating delicate medical practices and interactions between patients and physicians (Pozzi & Van den Hoven, 2023). As I further substantiate in this thesis, AI systems are epistemically authoritative, difficult to contest, and a substantial part of medical decision-making, and thus, they considerably affect patients' possibilities of sharing, receiving, and elaborating information and knowledge. That is, they have the potential to considerably impact patients' vulnerability to epistemic injustices. Ultimately, effectively showing that AI systems perpetuate epistemic injustices in medical contexts is imperative, even though it has been largely neglected in debates revolving around the ethics and epistemology of AI in medicine and healthcare.

In this dissertation, I aim to address this research gap by inquiring into forms of epistemic injustice that are brought about by AI systems in medicine and healthcare. Against this background, I propose a framework to capture and mitigate epistemic injustice that is specifically tailored to cases in which AI systems support crucial medical activities (such as diagnoses and treatment recommendations). The study I present is novel, as it represents the first effort to systematically grasp and ameliorate epistemic injustices brought about by AI systems in medical settings. This line of research also illustrates that the standard treatment of privacy, bias, responsibility, and trust, among others, has failed to address issues of epistemic injustice. This confirms the epistemic and ethical value of these issues and that they require treatment of their own right. It is then urgent that problems of epistemic injustice are put on par with other ethical concerns that are receiving an overwhelming amount of attention.

In the next section, I discuss the role of AI systems in medicine and healthcare more specifically. This should further highlight the urgent need for a framework that can account for their epistemological and ethical impact in

terms of epistemic injustice. The characteristics of the framework for epistemic injustice in AI are described in more detail in Section 1.3, in which I also outline an overview of each chapter of this dissertation.

1.2. Machine learning in medicine: between ethics and epistemology

In recent years, the AI revolution has strongly impacted the knowledge-intensive fields of medicine and healthcare. Many tasks that are constitutive of physicians' daily activities require collecting, elaborating, and analyzing a large amount of information and organizing it meaningfully. Thus, the advancement of data science technologies, such as machine learning (ML) systems, for implementation in healthcare is increasing rapidly. These systems have revolutionized healthcare provision by providing evaluations of health risks with a high degree of precision and making reliable treatment recommendations and diagnoses, among other tasks.

Before discussing ML in medicine and the interconnected nature of ethics and epistemology in its analysis, a terminological clarification is needed. AI can be understood as an umbrella term encompassing different data science techniques and methodologies.[3] For example, under AI fall ML systems, deep learning systems (Deep Neural Networks)[4], and also Large Language Models, which have received extensive attention in very recent debates due to the introduction of systems such as ChapGPT, among others. ML is, thus, a subcategory of AI. Specifically, ML pertains to systems that have the ability to recognize patterns in vast amounts of data points at high speed and provide

[3] There is a great amount of philosophical literature on whether the term artificial intelligence entails ascribing human-like traits to these systems, such as rationality, intentionality, and creativity. These debates are not relevant to the discussions presented in this thesis.

[4] Since Deep Neural Networks (DNNs) are a subcategory of ML, when I talk about ML, this on occasion includes DNNs. Further specifications are reported in the course of the dissertation.

solutions for relevant tasks (e.g., classifications and predictions), as opposed to following a set of predetermined rules (Alpaydin, 2014). Above and beyond more fine-grained differences of various data science methods, of particular interest to this thesis are black box systems, i.e., systems whose decision-making rationale is not directly accessible to the human investigator. These systems raise considerable epistemological and ethical questions that will receive closer scrutiny in the next chapters. In the remainder of the dissertation, I mostly refer to ML systems with this definition in mind. In cases in which I use the term AI instead, I consider the two terms to be synonymous, unless otherwise specified.[5]

The definition of ML provided above hinges on the premise that AI and ML are epistemic technologies. This is the case because, as Alvarado (2023) argues, "AI can be uniquely positioned as an epistemic technology in that it is primarily designed, developed and deployed to be used in epistemic contexts such as inquiry, it is explicitly deployed in such contexts to manipulate epistemic content such as data, and it manipulates such content specifically through epistemic operations such as inferences, predictions or analysis." (Alvarado, 2023, p. 32).[6] The epistemic power of ML systems makes their use in high-stakes domains, such as medicine and healthcare, particularly promising (Topol, 2019). At the time of writing, we are in the era of an ongoing healthcare crisis resulting from the COVID-19 pandemic in which valuable resources (such as physicians' time) are limited. Using ML systems in healthcare can potentially decrease the burden of human practitioners by supporting them in relevant tasks. For example, ML systems providing X-ray screenings for detecting COVID-induced pneumonia have been used to support physicians in highly sensitive and ethically demanding decision-making contexts such as pa-

[5] For example, in Chapter 2, I often refer to trust in AI (or trustworthy AI). This is because most literature on this topic treats trust in AI from a general perspective, without diving into different applications, thus sidelining how trust in ML could be different from trust in AI. Therefore, trust in AI and ML are used interchangeably in that chapter.

[6] Alvarado's analysis builds on Humphreys' definition of technical artifacts as epistemic enhancers, see Humphreys (2004).

tient triage. At first, this seemed to promise a viable solution amid a healthcare emergency characterized by a paucity of crucial medical resources, as reported by Hao, 2020. However, a later reassessment of the effectiveness and benefits of integrating ML systems in the triage of COVID patients showed much less promising results (Heaven, 2021). Nevertheless, it is worth noting that in a crisis situation, considerable resources were invested in developing ML systems to be introduced in clinical settings. This underpins the assumption that these systems will play an increasingly relevant role in healthcare settings in the years to come.

Even though the potential uses of ML in medicine abound, systematic inclusion of these systems in clinical practices faces considerable hurdles.[7] These are related to epistemological limitations (e.g., constraints in terms of explainability (Durán, 2021), transparency (Creel, 2020), and contestability (Venkatasubramanian & Alfano, 2020)) and ethical concerns (e.g., how to allocate responsibility and assess the systems' trustworthiness (Durán & Jongsma, 2021; Grote & Berens, 2020; Sand et al., 2021)). While exploiting the benefits of these systems is desirable, it is also paramount to recognize that they are strongly limited in several respects and can negatively affect patient-physician interactions.

Asymmetrical knowledge exchanges characterize medical encounters between (usually) non-expert patients and expert physicians. How knowledge and information flow from one side to the other, how credibility and trust are attributed, and how beliefs are formed are all epistemic activities that shape patient-physician interactions and are increasingly mediated by ML systems as powerful epistemic entities. While the relevance of epistemic considerations has been recognized in standard debates in ethics, the literature on the ethics of ML often considers epistemic challenges as decoupled from moral problems. This occurs to the extent that it is often assumed that once the epistemol-

[7] For example, Van de Sande et al. (2021) show that the majority of AI models for implementation in intensive care units (ICUs) remain in the testing and prototyping phase while only very few make it to being introduced in clinical settings.

ogy of ML is deemed suitable (say, a certain level of accuracy is secured), an ethical analysis emptied of epistemologically relevant considerations can be performed. For example, Mittelstadt et al. (2016) argue that an analysis of ML systems' unfair outcomes can be carried out while leaving epistemological aspects aside. Examining problems in terms of epistemic injustice in the context of ML shows that this way of conceiving the relationship between ethics and epistemology is too narrow. In fact, it does not do justice to the complexity and subtleties of issues that can emerge at the intersection of these two dimensions and need a synergy between them to be successfully captured.[8]

I would also like to point out that, more often than not, debates on the ethics of AI tend to formulate principles or conditions that need to be striven for in the development and deployment of ML systems following a top-down approach. For example, to secure the ethical acceptability of a given ML system, we need to ensure that it is explainable and does not infringe on the privacy of the stakeholders affected by its outputs. This approach is in line with many regulatory documents and ethical guidelines that aim to guarantee the fulfillment of certain ethical standards for AI systems. For example, the European Commission High-Level expert group widely discussed guidelines for trustworthy AI proceed by individuating a set of overarching principles that AI systems need to fulfill to be deemed trustworthy (European Commission, 2019).[9] While this approach is useful in providing guidance on central ethical and epistemological demands, it carries the risk of oversimplifying the social dynamics, often characterized by systemic injustices, in which AI systems are embedded. It is also noticeable that, compared to other widely discussed AI principles and values such as explainability, transparency, privacy, and responsibility, among others, justice as a central ethical ideal has received considerably less attention (Le Bui & Noble, 2020).

[8] I dedicate Chapter 3 to the detailed analysis of the implications of seeing the epistemology and ethics of AI as decoupled dimensions.

[9] I think this approach is problematic in many respects. A critique of it can be found in Chapter 2.

Against this background, in this dissertation, I use a bottom-up approach by starting from the analysis of conditions of epistemic injustice that are exacerbated or newly created by medical ML systems. It should be mentioned here that bottom-up approaches are already present in the literature. These often aim to tackle the unfair outcomes of ML systems by highlighting the disparate effect they have on socially disadvantaged populations. For example, it is widely recognized in the current debate that ML systems can accurately predict relevant outcomes for populations better represented in the set of training data (Goodman & Flaxman, 2017; Mittelstadt et al., 2016). This means that individuals belonging to underrepresented groups risk being mistakenly unrecognized or, even worse, penalized by these systems. Considerable research efforts are, justifiably, being invested in analyzing the fairness of algorithmic systems and their possibly discriminatory effects (Aquino et al., 2023). On occasion, AI systems turned out to be clamorously biased against population sub-groups, resulting in ethically unacceptable results. For instance, a system used in the US to estimate people's urgency with regard to receiving kidney transplants resulted in systematic discrimination of black people, largely lowering their overall chances of being considered eligible for a transplant (Simonite, 2020).

Tackling the discriminatory effects of algorithmic systems is important in the individuation of epistemic injustices. However, the latter cannot be reduced to issues of bias and discrimination. As I explain in more detail in the following chapters, epistemic injustice encompasses wider forms of harm that deserve separate scrutiny. The approach I take in this dissertation is the first to extend the framework of epistemic injustice to include ML as an additional entity entering medical encounters between patients and physicians. To address the more comprehensive issues of epistemic injustice caused by ML systems, there is a dire need for a suitable epistemological and ethical framework. Efforts toward the development of such an epistemological-ethical framework are particularly evident in Chapters 4, 5, 6, and 7. As discussed

earlier, the current debate does not offer a clear conceptualization of epistemic injustice tailored to the challenges raised by ML in medical decision-making. Therefore, the goal of this dissertation is, precisely, to provide such an epistemological and ethical framework needed to identify, analyze, and mitigate epistemic injustice when it is *ML-induced* in medicine.

1.3. Research questions and thesis overview

This thesis is divided into two parts. The first part includes Chapters 2 and 3. This part of the thesis presents a discussion of fundamental concepts at the intersection of the ethics and epistemology of ML geared toward making explicit two main research gaps. Here, I show how these research gaps lead to overlooking problems of epistemic injustice in ML and thus need to be addressed in a timely manner.

The first research gap concerns how the literature treats the concept of trust in AI. To this aim, I start by providing a literature review to corroborate the existence of this research gap.[10] This chapter displays the conditions for different forms of trustworthy AI. Moreover, and crucially for the general purpose of the dissertation, this chapter highlights that the specific question of whether AI systems are appropriate objects of our trust, and relatedly, whether *directly* trusting AI is possible and/or desirable, has received considerable attention. However, the literature review provided illustrates that the role that AI systems play in *mediating* relations of trust among relevant stakeholders (e.g., between patients and physicians in decision-making scenarios that involve AI systems) has been largely neglected thus far. Focusing on this research gap is essential to carry out an analysis of epistemic injustice since unjustifiably with-

[10] I would like to clarify that even though some chapters are based on co-authored papers, I use the first person singular throughout this introduction exclusively as a matter of consistency. Of course, the ideas presented in some chapters are the result of collaborative efforts with my co-authors. I specify which chapters are co-authored in a footnote at the beginning of each chapter.

holding trust can be detrimental to a person's epistemic standing. I analyze these issues in detail in Chapter 4 and maintain that, considering the increasing role that AI systems are playing in medicine, we need to evaluate how these systems have a bearing on physicians' assessments of patients' credibility and trustworthiness. Thus, Chapter 2 lays the groundwork for the development of a framework to identify testimonial injustice in ML that is later presented in Chapter 4.

The second research gap, which is addressed in Chapter 3, is related to how the general relationship between the ethics and epistemology of ML is conceptualized in the current debate. I show that considering the ethics and epistemology of ML as unrelated dimensions leads to the role of ethical properties in regulating central epistemic functions (such as explanations) in ML being overlooked. As I show later in the thesis, neglecting the regulatory role of the ethics of ML in its epistemology creates the conditions for epistemic injustice, in particular epistemic objectification, to emerge. Epistemic objectification in ML cannot be captured by simply applying the standard framework of epistemic injustice to ML cases for two main reasons. First, to determine epistemic objectification in ML, one needs to show that the approach to the ethics and epistemology of ML is unidirectional and, thus, constrains the possibility of ethically relevant properties (e.g., patients' values) affecting relevant epistemic functions (e.g., an explanation). The second and related reason is that a medical ML system cannot pick up on a patient's values and, therefore, constrains their contribution to the medical discourse. In Chapter 3, I tackle both points, thus providing the conceptual tools to capture epistemic objectification in medical ML.

So, while Chapter 2 *zooms into* a particularly relevant concept in the ethics and epistemology of AI literature, i.e., trust and trustworthy AI, to bridge the gap to epistemic injustice, also Chapter 3 plays a foundational role in this thesis, albeit through a different analytical lens. This chapter *zooms out* of particular issues and analyzes how the approach available in the literature

conceives of the general relationship between the ethics and epistemology of ML and how this contributes to issues of epistemic injustice being overlooked. Moreover, the chapter provides a conceptualization of epistemic objectification tailored to medical ML, thus expanding the concept beyond its standard formulation.

The second part of the dissertation includes Chapters 4 to 7. Here, I adopt a case-based, bottom-up approach. As previously stated, epistemic injustices are highly contextual and need to be discussed based on their specificity. Accordingly, I further develop my framework for ML-induced testimonial injustice, and I apply my considerations to two main cases of medical AI. The application of my framework to these cases depicts the societal relevance and urgency of addressing issues of epistemic injustice in ML. The first case is of an ML system currently deployed in the USA to predict patients' risk of developing opioid addiction or misuse (Oliva, 2022; Szalavitz, 2021). The second case is of a conversational AI (CAI) system used to provide mental health support to Syrian refugees (Solon, 2016). While there are also tangential references to other cases of ML in medicine and healthcare, the analysis of these two cases and the problems of epistemic injustice that arise through their use represent the backbone of the dissertation. The chapters in this part of the thesis provide the conceptual tools needed to recognize an epistemic injustice when it is *ML-induced* and account for its main features. In these chapters, I also offer my perspective on how to mitigate the forms of epistemic injustice analyzed (see, in particular, Chapters 6 and 7).

Overall, both parts of the dissertation contribute to answering my main research question:

Main research question:

In which ways does machine learning-induced epistemic injustice in medicine and healthcare emerge, and how can it be mitigated?

I now present a more detailed description of each chapter of this thesis. Chapter 2 focuses on the following research sub-question:

RQ 2

What are the conditions for relations of trust in connection with medical AI?

This chapter provides a map-out of the literature on trustwothy AI to analyze how the central concepts of trust and trustworthiness have been treated in the current debate. Following Hawley's standard conceptualization of trust as reliance plus some "extra factor" (Hawley, 2019), I structure the debate on trust in AI according to this definition. This conception of trust is in line with philosophical accounts that clearly distinguish between mere reliance and a morally rich notion of trust.

The aim of the literature review is twofold. First, it aims to show that in the literature on trustworthy AI, securing the reliance of AI systems is often taken for granted. Instead, much more emphasis is placed on whether AI can be the target of our trust in a morally relevant sense. However, I argue that assessing whether a system is reliable is much more demanding than often assumed, and I do this by analyzing different approaches available in the literature. The second goal of this chapter is to point out that the specific question of whether AI systems are appropriate objects of our trust has overshadowed a more pressing issue. This is the issue of the role AI systems play in *mediating* trust relationships between human healthcare professionals and patients. This literature review, thus, provides a critical analysis of the concept of trust in AI. Moreover, it lays the groundwork for my consideration of physicians' lack of trust in patients' testimony due to the role played by a medical ML involved in medical decision-making. I analyze this problem as the most immediate manifestation of testimonial injustice in medical ML in Chapter 4. Ultimately, the aim of Chapter 2 is to motivate the analysis of

questions of epistemic injustice in medical ML in terms of trust dysfunctions, by demonstrating the existence of a research gap that deserves timely attention and that I provide a suitable answer to.

Let me now turn to Chapter 3, which addresses the following research question:

RQ 3

Why is a separation of the epistemology and ethics of ML conducive to epistemic objectification?

In this chapter, I investigate the relationship between epistemology and ethics in explanatory machine learning (X-ML) in medicine, by focusing on how the relationship between these two dimensions is treated in the current debate. To this end, I make explicit the general tendency to treat epistemology and ethics as two separate dimensions, according to which the recognition of epistemological shortcomings (e.g., the opacity of the algorithm, and the difficulty in distinguishing between mere correlation and causation between datasets) *informs* the identification of ethical concerns (e.g., understood in terms of the autonomy of data subjects and the justification for being able to act upon the outcomes produced by ML systems). I label this approach, which tends to *compartmentalize* the epistemology and ethics of ML, as the *informativeness account*. By considering a concrete example of X-ML in healthcare, I examine a case that does not square well in the informativeness account because it paves the way to patients' epistemic injustice in the form of *epistemic objectification*. In turn, I maintain that a conception of the ethics and epistemology of ML that allows normative elements (such as patients' values) to affect central epistemic functions (such as an explanation in ML) is needed to prevent such objectification. Thus, this analysis represents a fundamental step in expanding the theoretical framework of epistemic injustice to account for the context of X-ML in healthcare. I dedicate the second part of this disser-

tation, starting with Chapter 4, to the further development of my framework for epistemic injustice in ML.

In Chapter 4, I consider the following research question:

RQ 4

What are the features of machine learning-induced testimonial injustice in medicine?

In this chapter, I analyze how *testimonial injustice* should be conceptualized within the context and novel challenges raised by medical ML systems. Drawing on the analysis worked out in the previous chapters, I consider the concrete and worrisome case of an ML system in medicine currently implemented throughout the USA to predict patients' likelihood of developing opioid addiction or misuse (PDMPs) (Oliva, 2022). I analyze the role of the system in mediating the patient-physician relationship by carefully examining its impact on physicians' assessments of patients' credibility and trustworthiness. Building on the conceptual analysis presented in Chapter 2, I argue that ML-induced testimonial injustice occurs in cases where these systems are treated as *markers of trustworthiness*. I also show how this contributes to propagating social inequalities at the expense of vulnerable social groups. In addition, I maintain that this leads to the silencing of patients and considerably constrains their testimonial contributions, thus revealing three main epistemological and ethical dysfunctions that constitute testimonial injustice specific to medical ML.

The first dysfunction is an automation bias that can lead to physicians attributing, by default, more credibility and trustworthiness to the ML system. The second is that the patient's right to convey information is strongly constrained. The third is that the scale of propagation of these forms of harm escapes the critical scrutiny of human physicians. It is, therefore, difficult to find a way to counteract these forms of injustice, at least at the level of what

individuals can do through virtuous behavior, as envisioned in Fricker's original framework (Fricker, 2007). In appendix A, I include the discussion of two main objections advanced to my account of testimonial injustice in medical ML. I refer to the replies to these objections as *a misguided equivalence* and *more than automation bias* (Pozzi, 2023c). This discussion aims to strengthen the arguments provided further and clarify the position taken in the course of the chapter. Overall, this chapter provides a novel conceptualization of testimonial injustice in relation to the opioid risk score case, by illuminating the features of testimonial injustice that emerge in medical contexts mediated by ML systems.

In Chapter 5, I dissect the same case (the opioid risk score case) to conceptualize hermeneutical injustice in medical ML. Therefore, these two chapters can be seen as complementary with the former conceptualizing testimonial injustice and the latter conceptualizing hermeneutical injustice in relation to ML. Against this background, I present the research question on which Chapter 5 is based:

RQ 5

What are the features of machine learning-induced hermeneutical injustice in medicine?

In order to answer this research question, I expand the other part of the original framework of epistemic injustice, i.e., *hermeneutical injustice*. I develop my framework by considering the position of a person who has been mistakenly ranked as being at a high risk of opioid misuse by the system previously described in Chapter 4. My framework for ML-induced hermeneutical injustice comprises two parts. The first part pertains to identifying ML-induced hermeneutical injustice occurring at the patients' expense. For this, I formulate three conditions, the fulfillment of which indicates the occurrence of an ML-induced hermeneutical injustice. The first condition states that an ML

system holds *unwarranted epistemic privilege*. In short, the epistemic privilege refers the fact that the system can reshape shared and medically grounded concepts (such as substance use disorder). For example, in the opioid case considered, I show that the risk scores for opioid addiction attributed to patients are produced by largely unclear metrics, and the efficiency of the system is based on the fact that physicians, following the systems' recommendations, issue fewer prescriptions. I argue that this leads to a definition of the risk of developing an opioid addiction that shifts away from the concept of addiction in its medically grounded and collectively shared meaning. This epistemic privilege is unwarranted because it eludes the critical scrutiny of human professionals. The second condition states that the way the system establishes conceptual resources hinders communication and understanding between patients and physicians. The third condition holds that the fulfillment of conditions one and two leads to considerable epistemic, moral, and practical disadvantages at the expense of hermeneutically disadvantaged patients (e.g., patients who are stigmatized due to the occurrence of substance use disorders). By applying this part of my framework to a concrete and socially relevant case, I illustrate the relevance of using my framework to tackle imperative injustices that were previously ignored.

The second part of my framework for ML-induced hermeneutical injustice is dedicated to the analysis of the epistemic role of physicians. Since they are epistemically privileged agents, standard debates in the epistemic injustice literature do not consider them as possible victims of epistemic injustice. However, as I show in this chapter, the use of ML systems in medicine also impacts epistemically well-positioned individuals to the extent that they, too, may experience an unjustified epistemic disadvantage. These considerations are novel in the analysis of epistemic injustice and form a constitutive part of my framework. To grasp how hermeneutical injustices that are ML-induced can impact physicians, I coin and advance the concept of *automated hermeneutical appropriation*. It is particularly worth noting that the epistemic

impairments experienced by physicians have an even more negative effect on patients. Physicians are the epistemic subjects that should, in principle, be able to recognize and counteract patients' hermeneutical injustice. However, since they are themselves impacted by it, they are constrained in their ability to support patients. This perspective of physicians as possible victims of epistemic injustice underscores the fact that ML-induced hermeneutical injustices are more wide-ranging than those occurring in human-human settings.

Chapter 4 and Chapter 5 together provide a novel framework of testimonial and hermeneutical injustice that emerges specifically *due to* the role played by ML systems in medical interactions.

In Chapter 6, I analyze the following research question:

RQ 6

How does participatory injustice in conversational AI emerge and how can it be mitigated?

In this chapter, I start with the consideration of a different case than the one scrutinized in the previous chapters: the case of a CAI system. In the face of the overall shortage of therapists to meet the psychological needs of vulnerable populations, AI-based technologies are often seen as a possible remedy. In particular, smartphone apps or chatbots are increasingly being used to offer mental health support, mostly through cognitive behavioral therapy. The assumption underlying the deployment of these systems is their ability to make mental health support accessible to generally underserved populations. However, considerations of the principle of justice in terms of its epistemic significance are still in their infancy in the debates on the ethics of mental health chatbots. This chapter aims to fill this research gap by focusing on a less familiar kind of harm that these systems can cause, namely, harm to users in their capacities as knowing subjects. More specifically, I address one form of epistemic injustice that arises through the use of these systems—*participatory*

injustice. This form of epistemic injustice is not part of Fricker's original framework but was proposed by Hookway (2010) in response to Fricker's monograph. In short, this concept aims to capture broader forms of epistemic harm that do not necessarily require testimonial exchanges between two or more interlocutors. Rather, it emerges due to problematic epistemic practices that unjustly constrain subjects in other central epistemic activities such as conjecturing, making hypotheses, advancing one's understanding of oneself and the world around us, refusing certain beliefs, and creating new ones, among many others.

To make sure my analysis is graspable and to demonstrate the urgency of this subject, I discuss the case of a mental health Chatbot, Karim, deployed to deliver mental health support to vulnerable populations such as Syrian refugees (Solon, 2016). This case substantiates the epistemological and ethical concerns arising from the use of mental health applications, specifically among vulnerable populations. Notably, this chapter offers a possible way to ameliorate the form of epistemic injustice discussed, that is, participatory injustice. This is done by arguing that epistemic participation can be conceptualized as a *capability* that can be accounted for within the framework of Capability Sensitive Design. The merit of this approach is that it provides designers and developers with the conceptual tools needed to critically question whether a certain CAI can bring about participatory injustice. This is conducive to identifying and anticipating previously undetected issues related to users' constraints in terms of epistemic participation.

In Chapter 7, I focus on another aspect that is closely related to the possibility of patients participating in the medical discourse, albeit through a different perspective. I show that recognizing the relevance of social aspects in disease causation is paramount to account for patients' social situatedness and the role the testimony of their lived experience should play in medical practices. I do so by analyzing the following research question:

RQ 7

What is the connection between the (un)recognition of the social dimension of disease causation and epistemic injustice in medical machine learning?

Building upon the conceptualization of the different forms of epistemic injustice discussed in the previous chapters, in this part of the thesis, I explore another aspect of epistemic injustice that is specifically related to patients' social situatedness.

The social aspects of causality in medicine and healthcare have been emphasized in recent debates in the philosophy of science as crucial factors that need to be considered to enable, among others, appropriate public health interventions (Russo, 2023). To this end, it seems central to recognize the bearing of social causes (broadly construed as, e.g., social inequalities and constrained access to health support) in bringing about certain concrete pathologies. Awareness of the relevance of social causes in medicine and healthcare is essential, considering the role that ML systems are increasingly playing in these high-stakes fields. This is because these systems could potentially fail to account for social aspects that are causally relevant. I illustrate that this is highly problematic because it paves the way to issues of epistemic injustice that need to be addressed.

This chapter discusses questions related to disease causation that are usually not considered in the social epistemology literature on epistemic injustice or in the standard philosophy of science discussions revolving around causality. I show that it is paramount to bring them together for, at least, two reasons. First, because the social factors that contribute to causing a higher incidence of certain pathologies in populations' subgroups carry a very strong epistemic component: identifying them as causes leads to understanding not only the pathology itself but also the social situatedness of underrepresented

social groups. In turn, clarifying the risks of ML systems contributing to overshadowing patients' social experience draws attention to their social positioning.

The second reason is that gearing efforts toward understanding the social situatedness of patients by highlighting the social causes of diseases helps to counteract forms of epistemic injustice (particularly testimonial injustice). In this chapter, I argue that the causes searched by medical professionals (i.e., either only biological or bio-social (Ghiara and Russo, 2019)) are connected with the sources of information and knowledge (such as patients' testimony) that are considered pertinent to inform the medical discourse. A central goal of this chapter is to show that accounts of disease causation that consider biological *and* social factors as causally relevant at the individual level are suitable for ensuring that patients' testimonial contributions play an important role in medical interactions.

I would also like to point out that a considerable part of this chapter is dedicated to making explicit the connection between social factors in disease causation and questions of epistemic injustice. Considerations related to how ML systems complicate this appear later in the chapter. Nevertheless, the discussion on social causes and epistemic injustice is paramount for demonstrating the relevance of including social factors in ML systems in medicine to avoid potential issues in terms of epistemic injustice (particularly testimonial injustice) from emerging. This chapter, thus, focuses on the need to include normatively relevant information pertaining to patients and their social situatedness, as discussed, albeit under a different light, also in Chapter 3.

Importantly, for working towards design solutions to facilitate epistemic justice, the analysis laid out in this chapter provides arguments for explicitly incorporating social factors related to disease in the development of ML systems in medicine and healthcare. In the course of this chapter, it will become clear that this is essential for epistemic justice. Since social factors often fail to be operationalized, the reasons put forth in this chapter depict the need to

make these issues central to the process of design and development of these systems.

In Chapter 8, I conclude this thesis with my final remarks. First, I summarize the key findings emerging from the previous chapters of the dissertation. This part of the chapter further highlights the novelty of the contribution provided by this thesis in addressing issues that have been so far neglected in the debate on the ethics and epistemology of ML in medicine. Second, I point out aspects of my analysis that require further attention. Finally, I indicate how my framework for ML-induced epistemic injustice can constitute the basis for further research.

I

Part 1: Theoretical Underpinning

2

What is Trustworthy AI?[1]

2.1. Introduction

Establishing AI systems' trustworthiness is increasingly considered a fundamental desideratum for their integration into society. This holds particularly true in human-sensitive domains such as medicine, healthcare, employment, government, energy, criminal justice, and security. The general principles for Trustworthy AI outlined by the EU Commission (European Commission, 2019), echoed throughout the specialized literature (e.g., Kaur et al., 2022; Li et al., 2023), advocate for caution and the pursuit of robust solutions. Many technical solutions are available today that aim to fortify our trust in AI and ensure their trustworthiness (e.g., Cho et al., 2019). But what makes an AI trustworthy? Why should we trust its output and behavior? Does it come down to merely scrutinizing the algorithm's patterns, or is there more to it?

To illustrate the interplay between trust and trustworthiness and set the

[1]This chapter is based on the following article:

Durán, J. M. & Pozzi, G. (under review). What is Trustworthy AI?

stage for the goals of this paper, consider a case of interpersonal trust that is easily relatable. We place our trust in physicians because they have undergone medical school, acquired the knowledge of medicine, and possess the ability to apply medical care in specific situations. Philosophically, this is referred to as the *reliance* on the trustee—the physician. Reliance, in this context, is an epistemic term, signifying a property that something or someone upholds for being *trustworthy*. We rely on the physician's competence based on having the right education. We rely on the bus because it is always on time. But is reliance alone sufficient for trust? Can we simply say we trust a physician because they went to medical school?

Our *trust* in a physician extends beyond the expectation that they will prescribe the right medicines and make accurate diagnoses. By trusting, we also hold a normative expectation that the physician will do the right thing. For instance, we expect them to act in our best interest, in accordance with the law, and follow the biomedical principles of beneficence and non-maleficence. In other words, trust places a moral demand on the physician to act in ways that surpass the mere value of their medical knowledge. It follows that reliance must be complemented with external considerations to constitute genuine trust. This is where we introduce the ambivalent concept of an "extra factor", often taking the form of responsibility, commitment, and goodwill.

While the example above may apply to interpersonal trust between two humans (a trustor and a trustee, a physician and a patient), its relevance to AI systems is less clear. Consider the "extra factor", for instance. Can we demand responsibility from an AI, and if so, what would that entail? More provocatively, can we expect an AI to have our best interests "at heart"? These questions form the foundation of considerations about *trustworthy AI*, as defined by the EU Guidelines. This article aims to analyze the complexity of this issue by first distinguishing *trust* from *trustworthiness*, and then discussing the former as a two-part concept: reliance and the "extra factor".

With these goals in mind, we divide this article into three main sections.

Section 2.2 briefly presents the philosophical literature on trust and trustworthiness. Two main takeaways stem from this section. First, trustworthy AI is a property of the algorithm—or its output—achieved by means of sanctioning its reliability. Second, we can analytically divide studies pertaining to questions of relations of trust in AI between reliance, focused on establishing the conditions for scientifically valid outputs, and an "extra factor", focused on identifying what motivates or drives trust in a morally loaded sense (Hawley, 2019).

Section 2.3 focuses on the first necessary component to establish relations of trust and the property of being trustworthy AI—that is, reliance. We treat reliance as the property of an algorithm—or its output—of producing scientifically valid outputs. Two main approaches emerge prominently: transparency, widely popular both in philosophical and data science circles and computational reliabilism, much less known but a major contender to transparency. In this section, we spell out the merits and shortcomings of both transparency and computational reliabilism (CR) as two viable approaches to ensure the reliability of AI systems.

In Section 2.4, we turn to the second part of the definition of trust, focusing on discussions revolving around the "extra factor". Here, we address the fragmented philosophical debates on trust in AI in an attempt to bring some order to the discourse. To this end, we subdivide the debates between those who believe that trusting AI is *not possible* (or even undesirable) and those who believe that trust in AI is *possible* and much needed (see Figure 2.1). We approach these debates critically, highlighting their merits and shortcomings. This should contribute to the analysis of different positions regarding the "extra factor" concerning both the conceptual and normative possibility of genuinely trusting (and not merely relying on) AI systems.

Finally, in Section 2.5, we provide a brief summary of the main findings of this article, and we sketch some suggestions for further research revolving around trust, trustworthiness, and AI systems. A summary of the key concepts

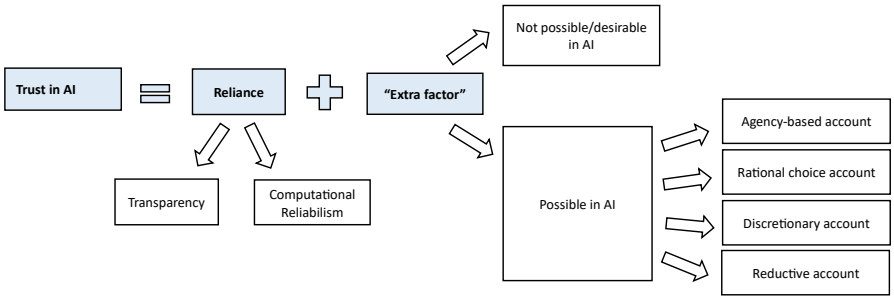

Figure 2.1: Definition of trust in AI as reliance plus some "extra factor".

used in this entry, along with definitions and proposed relevant literature, can be found in Table 2.1.

Table 2.1: Key concepts, definitions and proposed relevant literature

Key concepts	**Definition(s)**	**Relevant literature**
Trust	Attitude of a trustor toward a trustee (feelings of betrayal upon its breach)	(McLeod, 2021)
Trustworthiness	Set of properties of the trustee (that give the trustor reasons to trust)	(Hardin, 2002)
Reliance	Expectation that the trustee will continue to perform as expected (disappointment but not betrayal upon failed reliance); in AI, justification that an AI produces scientifically valid outputs; epistemic attitude of humans toward AI systems	(Durán & Formanek, 2018; Ryan, 2020)

Continued on next page

Table 2.1 – continued from previous page

Key concepts	Definition(s)	Relevant literature
Transparency	Revealing hidden workings, causal connections, low-level mechanistic relations, and interdependencies within algorithm to enhance understanding of its output	(Buijsman, 2022; Creel, 2020; Datta et al., 2016; Guidotti et al., 2019; D. S. Watson & Floridi, 2021)
Computational Reliabilism	Accepting AI's epistemic opacity, frequentist account assessing reliance through reliability indicators (system's technical robustness, compatibility with computer-based scientific practice, social construction of scientific beliefs)	(Durán & Formanek, 2018; Durán & Jongsma, 2021; Durán, manuscript; Ferrario et al., 2021)
"Extra factor"	Expectation that the trustee will be favorably moved of the trustee (Jones) Expectations + motivations/goodwill of the trustee (Hardin) The trustor relies on the trustee being willing to do what they are entrusted with (McLeod) Commitment of the trustee (Hawley)	(Hardin, 2002; Hawley, 2014, 2019; Jones, 1996; McLeod, 2021)
Agency-based account	AI systems have minimal agency; intentionality (minimal sense); AI systems are bearers of normative commitments	(Chen, 2021; Lewis & Marsh, 2022; Starke et al., 2022)
Rational choice account	Non normative, non-affective account; trust in AI amounts to relying on it without updating our beliefs (relying without monitoring)	(Ferrario et al., 2020, 2021)

Continued on next page

Table 2.1 – **continued from previous page**

Key concepts	Definition(s)	Relevant literature
Discretionary account	Trust manifests in the discretionary authority attributed to AI; normative account (AI object of normative expectations, designers are the bearers of it)	(Nickel, 2022)
Reductive account	Trust not in AI directly but in AI its human designers and developers	(Sutrop, 2019)

2.2. The multiple dimensions of trust and trustworthiness

The concepts of trust and trustworthiness are ubiquitous in our daily lives, forming the bedrock of interpersonal relationships and societal dynamics. Despite this fact, it is remarkably difficult to define them in satisfactory terms that encapsulate their complexity and elucidate their fundamental features. Typically, we use these concepts across various contexts to govern interpersonal relationships and articulate our expectations from others. We express our present and future trust in our physicians and friends because they have shown to be trustworthy. We might, however, be less inclined to extend the same level of trust to a politician who has exhibited signs of untrustworthiness in the past. Similarly, establishing clear conceptual distinctions between what it means to trust that someone will fulfill a promise and merely hoping they will is far from straightforward.

These intricate issues deepen when AI systems become integral to decision-making contexts where the judicious allocation of trust is particularly important. AI systems serve as mediators in trust relationships among different stakeholders, such as between doctors and patients, banks and loaners. They

are also positioned to be the direct recipients of our trust, as seen in recent cases in the judicial system (e.g., COMPAS, see Dieterich et al., 2016). Similar to interpersonal trust, the users' trust in AI systems is foundational for their acceptance and successful integration into relevant social practices (Choung et al., 2022).

Philosophical inquiries into trust and trustworthiness focus on understanding the practical commitment between a trustor and a trustee. Conventionally, this involves the expectation that the trustee will fulfill the commitments made to the trustor or undertake actions deemed appropriate based on their expertise, training, or responsibility (Hawley, 2019). For instance, a trustworthy physician delivers accurate diagnoses, while a trustworthy friend safeguards shared secrets. Now, our trust in physicians also extends beyond their medical training, and the trust in our friends is not solely based on the fulfillment of their promises. Our trust in physicians is rooted in their commitment to our well-being (beneficence) and the prevention of harm (non-maleficence), or in assuming an ethical and legal responsibility for their actions. Similarly, our trust in friends arises from their genuine affection and their willingness to refrain from deceiving us. This perspective underscores that trust involves the readiness of the trustor to put themselves in a situation of vulnerability, uncertainty, and risk (Lewis & Marsh, 2022).[2] It's worth noting that trust becomes relevant precisely in situations lacking full control, where delegation of a specific task to a trustee is necessary (McLeod, 2021).

Let us also note that we start with the premise that both the trustor and the trustee are individual human agents (e.g., ourselves, our physicians, our friends). Extending these results to a collective of agents (e.g., the Interna-

[2]Here, our analysis is limited to instances of trust where a trustor delegates a specific task to a trustee with the aim of achieving a particular goal (x trusts y to do z) (McLeod, 2021). Trust, by its nature, is predominantly contextual. For instance, one might trust their physician for medical prescriptions but not for car repairs. Concepts like "generalized trust" (Hardin, 2002), such as trustworthiness as a character disposition or virtue, and broader notions of trust that extend beyond specific tasks or actions, will not be addressed in this article.

tional Panel for Climate Change) or, more abstractly, to institutions (e.g., the WHO or the government of The Netherlands) should not prove overly challenging. It is, in principle, appropriate to assert that an institution like the WHO is trustworthy based on its prioritization of global public health in its decisions regarding COVID-19.

Problems arise, however, when the trustee is an instrument or, in the case of AI, a computational algorithm. To illustrate this, contrast our discussion with the feelings that arise when the practical commitment of trust is breached.[3] Instances of betrayal, deception, disappointment, and disgruntlement surface when we discover our physicians misdiagnosing us or our friend revealing a sworn secret. In cases similar to these, a breach of trust triggers genuine moral reactions due to the fact that the trustee fails to meet the normative or affective expectations we have of them. We expect from our physicians that they should act in our best interest and of our friends that they care so much about us to not reveal private information. In such instances, it is deemed appropriate to feel betrayed, and there is a rightful demand for an explanation regarding the failure to fulfill a specific commitment, or an apology is warranted. However, it seems unfitting to feel disgruntled when a light bulb is not working or consider that the car has betrayed us when the engine does not start. There is a figurative use of trust applicable to these inanimate objects; for example, we trust that the lamp will light up as long

[3]Let us point out that trust, distrust, and misplaced trust are conceptually distinguished attitudes. Distrust does not amount to a mere lack of trust since distrust entails a moral criticism that a lack of trust does not (Hawley, 2017). For instance, one may *distrust* a friend who has revealed a secret in the past. Here distrust is appropriate because the friend has demonstrably not respected the commitment they made (i.e., to keep my secret). Conversely, a *lack of trust* can occur in situations in which neither trust nor distrust would be appropriate. For example, I neither trust nor distrust my physician with the repair of my car simply because neither attitude pertains to the task domain of my physician. As for *misplaced trust*, Nickel considers situations in which physicians perform defensive medicine, e.g., by over-prescribing medication (say, antibiotics for the common cold), due to the fear of breaching the trust patients put in them (Nickel, 2009). However, expecting a physician to prescribe antibiotics for a simple cold is not a measure of the physician's trustworthiness, and thus, as Nickel argues, trust in this case is misplaced (see also Hawley, 2015). In this article, we focus exclusively on relations of trust.

as its basic functioning is unaltered, and we trust the car to start in the cold morning. Sometimes, we even say that the car is trustworthy or that such and such a company builds trustworthy cars. However, it needs to be clear that our relationship with these objects is one of reliance for specific purposes, not necessarily of a relation of trust. While it is appropriate to rely on the well-functioning of the lamp and the car, expecting loyalty from the former or demanding an apology from the latter would be inappropriate (Hawley, 2019, p. 2). A similar perspective seems to apply to computer-based algorithms.

The central issue here is that trust and trustworthiness have predominantly been conceptualized in *anthropomorphic* terms, characterized by their appeal to distinctly human and morally-laden emotions (such as betrayal, deception, intentionality, accountability, etc.). Approached in this manner, assertions about trust and trustworthy AI may seem inappropriate, unwanted, and potentially misleading. As we will delve into shortly (Section 2.4.1), this stance is firmly held by many philosophers working in this domain. The alternative involves formulating an account of trust and trustworthiness that explicitly incorporates AI as a significant component of its study. While current approaches exhibit notable shortcomings, there are compelling arguments that could guide us in gaining a more nuanced understanding of how to navigate these intricate issues (Section 2.4.2).

Let us now briefly but explicitly articulate the philosophical distinction between trust and trustworthiness. Trust is considered an *attitude* that reflects the trustor's inclination to place trust in the trustee. This is why we say we trust our physician, signifying an attitude of confidence, belief, or a similar sentiment towards the physician. On the other hand, trustworthiness is a *property* intrinsic to the trustee. It represents what makes a trustee "demonstrably worthy of trust." (Sutrop, 2019) In this context, we label the physician as trustworthy because they have demonstrated that they are deserving of our trust.

Thus understood, trust and trustworthiness are conceptually distinct, allowing for the possibility of trusting an untrustworthy person or entity and, vice versa, withholding trust even when the trustee is, in fact, trustworthy. Let us briefly consider both cases in turn.

Situations in which we trust the untrustworthy typically occur when we do not have much information about the trustee. For instance, consider the case of Zholia Alemi, who was found guilty of fraud for practicing as a psychiatrist for over 20 years without having acquired any medical qualifications (Bugel, 2023). Even though someone who falsifies a medical degree cannot be considered trustworthy, it is very likely that her patients trusted that she was a reliable and competent professional. The reason for this was a lack of information regarding the fact that she had not been to medical school. Knowing this information would have altered her patients' attitude of trust.

Instances in which trust is withheld, even though the trustee is trustworthy, often arise when the trustor holds biases that deflate the perceived trustworthiness of the trustee. Fricker's work on epistemic injustice underscores precisely this: individuals can fail to trust the trustworthy due to biases related to their interlocutor's social identity (e.g., biases related to gender, race, ability, socio-economic status) (Fricker, 2007). This phenomenon can occur in interpersonal relationships due to implicit biases a person can hold, but they can also be fueled by explicitly discriminatory and stigmatizing public attitudes and statements. Notable instances of this phenomenon emerged among Trump supporters after hearing his controversial remarks about immigrants during his presidential campaign announcement speech on June 16, 2015. In that speech, he stated, "(w)hen Mexico sends its people, they're not sending their best... They're bringing drugs. They're bringing crime. They're rapists. And some, I assume, are good people." (Phillips, 2017) This dubious and profoundly unfair statement can lead to failing to trust trustworthy individuals simply because they are the object of unfounded prejudices.

Now, while conceptually distinct, trust and trustworthiness are deeply interconnected in the sense that the presence of one entails the other. In other words, it is impossible to conceive (mis)trusting someone—or something—that lacks the property of being (un)trustworthy. This interconnection between trust and trustworthiness is pivotal in the debate over *trustworthy AI*—now understood in the general sense of the EU Guidelines.

Asserting that an AI system is trustworthy necessitates establishing the reliance of the system. It seems rather obvious that an AI with predictive accuracy for cancerous moles closer to 95% is deemed more trustworthy than one in the vicinity of 35%. High predictive accuracy, however, is not the sole criterion for entrenching reliance on a given AI system. One might argue that explainability is the key property of a trustworthy AI. Likewise, one might request that the AI system possesses specific scientific merits that make it trustworthy. In Section 2.3, we explore various options where the property of being trustworthy is elaborated and defended. Our focus centers on ongoing discussions on transparency and computational reliabilism, drawing insights from both philosophical and technical literature.

Trust, on the other hand, is a more complex issue for a comprehensive understanding of *trustworthy AI*. Recall that trust is an attitude pertaining to the trustor, to be inclined to trust the trustee. As such, it requires not only some degree of reliance on the trustee but also an "extra factor" (Hawley, 2014, p. 5). Take again the case of trusting a physician. It is not enough to deem them trustworthy just given the right credentials and certificates. Proper relations of trust only surface when the physician shows to be responsible for our well-being, is legally bound, or has our best interests at heart. This complexity of trust can also be illustrated with AI systems. Consider a Convolutional Neural Network (CNN) accurately predicting criminals based on facial traits (e.g., the curvature of the mouth, the distance between the eyes, etc). While this CNN can be deemed trustworthy due to highly accurate predictions—it has been reported an estimate of 95% accuracy (Wu & Zhang, 2016)—it cannot

be trusted in its outputs. If a judge were to sentence a person to prison based on the curvature of their nose, it would not only violate their rights and due process, neglecting the principles of fair and unbiased judgment, but also undermine justice, equality, human rights, and could lead to severe consequences such as wrongful imprisonment and perpetuation of systemic biases. At the same time, we cannot genuinely say that this DNN is responsible for its output, or it has the best intentions "at heart". Solving the issue of trust in AI is at the root of any comprehensive understanding of *trustworthy AI*. Philosophers recognize the difficulties of it, particularly pinning down the "extra factor".

What exactly is this "extra factor"? Opinions among philosophers are divided. Some interpret it as a positive view of the *motives* of the trusted person. For instance, one might trust a physician because they have the right motives to look after one's health. However, defining what constitutes the "right motive" requires further clarification. Is it because physicians are bound to the Hippocratic Oath, or is it due to legal accountability? On the other hand, some consider the "extra factor" as a *reasonable expectation* on the trusted. Jones, for example, defines it as "the expectation that the one trusted will be directly and favorably moved by the thought that we are counting on her." (Jones, 1996, p. 1) Yet others, such as Hardin, combine *expectations with motives*, stating that "the truster's expectations of the trusted's behavior depend on assessments of certain motivations of the trusted." (Hardin, 2002, p. xix) McLeod points out that the extra factor " generally concerns why the trustor (i.e., the one trusting) would rely on the trustee to be willing to do what they are trusted to do." (McLeod, 2021) This perspective thus puts the focus on the *willingness* of the trustee.[4]

Drawing from this literature, philosophers make efforts to accommodate AI. Within this context, two primary lines of argumentation surface. The first contends that the anthropomorphization of trust inherently rules out any possibility of trusting AI. That is, trust requires some form of responsibility,

[4]For a critical review, see also Goldberg (2020).

intentions, or normative commitment, none of which can be ascribed to algorithms. Section 2.4.1 presents and briefly discusses the main proponents of this view. The second line of argumentation posits that trust in AI is indeed feasible; we just need to accept some assumptions and conditions. Section 2.4.2 discusses this possibility. Let us finally mention that our treatment intentionally simplifies various issues, and for instance, we will not explore the role that the trustor's prior beliefs might, or might not play in establishing a relationship of trust with an AI system.

2.3. RELIANCE

It seems rather uncontroversial to say that we must secure an AI system's reliance before crediting our trust in it. After all, one would not trust a physician if they did not attend medical school. However, as uncontroversial as it might seem, it is far from clear how the reliance of an AI system can be established. In what follows, two theoretical frameworks are explored for establishing the general reliance of AI systems: *transparency* and *computational reliabilism*. Of specific interest is the application of AI in the scientific field.

Before we begin, two conceptual clarifications need our attention. First, reliance is not taken to be a property that AI systems have or fail to have. Rather, it comes in degrees. For instance, an AI system is reliable because it produces accurate results most of the time; or its output is reliable because we managed to get some degree of transparency. Second, we understand reliance as a method (or set of methods) by which we can justifiably state that an AI system is reliable or renders scientifically valid outputs.[5]

[5]We maintain a neutral stance regarding the precise definition of "a scientifically valid output". This concept can encompass various interpretations, such as being acceptable in terms of empirical predictions, formally correct, theoretically sound, and more. The specific criteria for scientific validity may vary depending on the context and the goals of the AI system.

2.3.1. Transparency

Transparency is undeniably one of the most highly regarded methods for establishing the reliability of AI systems.[6] The underlying sentiment is genuine: when we can clearly understand how a system operates, we have grounds to believe that its outputs hold scientific merit. As articulated by Guidotti et al., "[t]he availability of transparent machine-learning technologies would lead to a gain of trust and awareness on the fact that it is always possible to know the reasons for a decision or an event." (Guidotti et al., 2019, p. 93:2) In this respect, it is crucial to explore what transparency entails and how we can attain it.

The initial approach to defining transparency is to consider it as the opposite of opaque or "black box" algorithms, as suggested by Lipton (2018) and Creel (2020, p. 569). In simpler terms, a transparent AI system is one that is not opaque. Unfortunately, this interpretation is all but illuminating, raising questions about what constitutes an opaque system and what exactly is meant by the opposite of opacity. Furthermore, it fails to recognize that opacity can take on different forms, including epistemic, methodological, and semantic. Epistemic opacity, for instance, refers to the inherent cognitive limitations of humans to comprehensively understand and account for the state of a computer process, encompassing variables, system relations, and system status (Durán & Formanek, 2018; Humphreys, 2009, p. 618). Methodological opacity, on the other hand, concerns the coding practices and strategies used in the development of AI systems that are not always readily accessible to developers. These coding practices may involve complex algorithms or proprietary

[6] Transparency is a polysemous concept. For instance, transparency applies to the readiness of a company to share relevant information with their stakeholders (European Commission, 2019, p. 3), or of a government to disclose their plans. Thus understood, transparency amounts to a commitment to openly share information, processes, and decision-making with the public or its stakeholders. It involves clear communication, accessibility of relevant data, and a commitment to accountability. Transparency fosters trust by allowing external scrutiny, enabling informed decision-making, and demonstrating adherence to ethical and responsible practices. Here, we exclusively consider it in its epistemic sense of justifying our belief in the output of an AI system.

techniques that are not easily discernible (Burrell, 2016). Finally, semantic opacity relates to the difficulty in establishing a direct and meaningful representation between the AI system and real-world phenomena. This challenge arises from the abstract nature of AI algorithms, which might not always align perfectly with the complexities of the real world they seek to model or interact with (Humphreys, 2009, p. 619).

To illustrate the challenges in defining transparency in these terms, let us consider the opposite of epistemic opacity, which is *epistemic transparency*. In this context, transparency means the cognitive ability to comprehensively survey and account for variables, system relations, and other elements within the algorithm. To demonstrate this interpretation, we can examine any Deep Neural Network (DNN). In principle, it is impossible for any human agent or group of agents to halt a DNN at a specific time t and assert full knowledge of the DNN's general state at that moment (e.g., which values have been instantiated for various variables). Similarly, predicting the DNN's next step at time *t+1* (including computing the next step and determining which variables will be instantiated) or retroactively accounting for the past state of the DNN at *t-1* (e.g., identifying which variables were instantiated in the previous run) is exceptionally challenging. In summary, epistemic transparency implies having what could be presumed as complete cognitive access to the DNN at *t-1*, *t*, and *t+1*, as well as the ability to provide meaningful insights about the algorithm. However, it is a well-established fact that achieving such comprehensive access is not cognitively possible for human agents, especially when complex AI systems like DNNs are involved.

The problem here is that opacity tends to be seen in absolute terms: algorithms are either opaque or not, with many of them exhibiting opacity on one or more levels (epistemic, methodological, semantic). In contrast, transparency is a concept that exists along a continuum involving degrees of transparency. It is, therefore, quite difficult to define one in terms of the opposition.

There is a more nuanced interpretation of transparency that has been articulated by Creel, who identifies three distinct forms or levels of it (Creel, 2020, p. 569):

1. *Functional Transparency:* This refers to having knowledge of how the algorithm as a whole functions and operates.

2. *Structural Transparency:* This involves knowledge of how the algorithm is implemented in code, essentially the coding details that make it work.

3. *Run Transparency:* This is concerned with knowledge of how the program actually operates in a specific instance, including the hardware and input data used during execution.

While it is useful to distinguish between these different sources of transparency, Creel's framework does not explicitly address how effectively each form of transparency can be achieved. This leaves room for the possibility that there may be multiple existing methods to attain each individual form of transparency, diverse degrees of transparency, and incompatibilities among methods (e.g., different approaches may prioritize one aspect of transparency over others or employ different techniques and trade-offs). This underscores the complexity and multifaceted nature of transparency in the context of AI and computational systems.

Perhaps the most widely accepted interpretation of transparency involves making visible the low-level mechanistic relations that underlie how an algorithm operates. This interpretation places significant emphasis on revealing the inner workings, causal connections, and interdependencies within the algorithm to enhance our understanding of its functioning and the outputs it produces. Following the literature, let us call it "opening the black box".

Now, there are several ways to advocate for opening the black box. One is to consider uncovering the hidden causal structures within the algorithm. This entails revealing the cause-and-effect relationships that account for how the

algorithm generates its outputs, a pursuit that has roots in logic, philosophy of science, and computer science (Pearl, 2000; Spirtes et al., 2000). Another, not necessarily unrelated way, one opens the black box by explaining how a specific outcome of the algorithm is achieved. This explanation may require providing a clear description of the steps, processes, or mechanisms that lead to a particular result, as discussed in numerous works (Páez, 2019; D. S. Watson & Floridi, 2021). More generally, making low-level mechanistic relations visible can be understood as conveying "useful information of any kind" about how the algorithm behaves and its outputs rendered (Lipton, 2018). This encompasses a wide range of information that aids all stakeholders in comprehending the inner workings of the algorithm.

On more practical grounds, how can this latter form of transparency be achieved? To answer this question, we will refer to the classification provided by Guidotti et al. (2019, 93:15), which outlines four methods for opening the black box: (i) explaining the model, (ii) explaining the outcome, (iii) inspecting the black box internally, and (iv) providing a transparent solution. We have already covered methods (i) and (ii) in a previous discussion (see Durán (2021)), so we will exclude them here. Method (iv), on the other hand, is closely related to (i), as it involves directly providing a model that is either locally or globally interpretable. We will not delve into the details of either of these methods but instead defer to the authors for their explanation (see Guidotti et al. (2019, pp. 93:14-93:15)). We will, however, discuss method (iii) which focuses on inspecting the black box internally.

According to Guidotti et al., the process of inspecting a model involves providing a representation (which can be visual, textual, dynamic, static, etc.) that aids in our understanding of particular properties of the black box and leads to justification. For instance, sensitivity analysis plays a role in "observing the changes occurring in the predictions when varying the input [of the algorithm]." (Guidotti et al., 2019, p. 93:14) These changes can then be visualized, often through tools like partial dependence plots (Goldstein et al.,

2015) and variable effect characteristic curves (Cortez & Embrechts, 2013). The information extracted from various visualizations and plots contribute to the justification in the belief that the output has scientific value. Importantly, what distinguishes the process of inspecting a model is that sensitivity analysis focuses on analyzing specific properties of the black box without necessitating a comprehensive understanding of the entire system (Guidotti et al., 2019, 93:14).

A concrete example of inspecting a model is *Qualitative Input Influence* (QII). At its core, QII quantifies the joint influence that specific inputs have on the outputs of machine learning or Deep Neural Networks (DNNs). Datta, Sen, and Zick describe the fundamental principles of QII as follows: "A transparency query assesses the influence of an input on a quantity of interest, where the quantity of interest represents a system's behavior for a given input distribution." (Datta et al., 2016, p. 599) These assessments are later used to prepare *transparency reports* that accompany system decisions (e.g., explaining a specific credit decision) and for testing tools useful for internal and external oversight (e.g., to detect algorithmic discrimination).

To illustrate how a transparency report works, consider the case of Mr. X, a 23-year-old adult male from Vietnam with an 11th-grade education, never-married, with $14k in capital gains and $0k in capital loss (for a complete list of profile variables, see Figure 4a in Datta et al. (2016, p. 608)). According to QII, Mr. X is classified as a low-income individual, despite having high capital gains and low capital losses. This output is somehow shocking, as "only 2.1% of people with capital gains higher than $10k are reported as low-income." (Datta et al., 2016, p. 608) Given these unexpected results, there is a need to account for how this output is determined.

The transparency report can swiftly reveal which variables wield more influence over the output, thus justifying the belief that the result has scientific value. For instance, the report reveals that classifying Mr. X as a low-income individual is not due to his ethnicity or country of origin, as one might sus-

pect without inspecting the algorithm. Instead, it is primarily attributed to his marital status, relationship, and education. This crucial insight is easily gleaned by examining the transparency report, which typically consists of a bar graph indicating the measured quantity for each variable (see Figure 4b in (Datta et al., 2016, p. 611)).

Admittedly, our description of QII is a simplified overview. A more comprehensive, though still incomplete, analysis would involve discussing various metrics used to measure the correlation between variables, the strength of these correlations (e.g., Pearson correlation), the weighting of protected attributes (e.g., race, gender, drug history, arrests), the proportion of positive predictions (e.g., Disparate Impact Ratio), the assessment of dependence between random variables (e.g., Mutual Information), and considerations of group disparity (i.e., classifiers that do not use variables as inputs—such as, gender for a bank loan—that lead to group disparities tend to be fairer). Despite these simplifications, the essence of QII remains intact: the authors demonstrate how specific groups of variables (such as age, marital status, etc.) influence the machine learning model's output (e.g., Mr. X's classification as a low-income individual) through various measures made visible in the transparency report.

Let us close this section by noting that transparency encompasses a broader range of methods than our analysis of "opening the black box". We can attain forms of *functional transparency* without fully delving into the algorithm's inner workings or revealing its internal representations. This occurs when an output is explained in terms of the algorithm's high-level behavior. For instance, algorithms such as LIME can account for the predictions of any classifier by locally learning an interpretable model. In practice, if an ML system predicts that a patient has the flu, LIME can highlight the symptoms in the patient's history responsible for the prediction. 'Sneeze' and 'headache', for example, are key variables used by the algorithm. They are flagged as net contributors to the flu prediction. In contrast, 'no fatigue' is a variable used as evidence against the prediction (Ribeiro et al., 2016).

Objections to Transparency

In the pursuit of transparency, an array of resources has been dedicated to the cause. In this respect, it is imperative to recognize that transparency carries numerous, frequently overlooked, shortcomings. This section will briefly explore some of these issues and assess their potential impact on our confidence in the algorithm and on its outputs.

First, there is the issue of *algorithmic regress* or *transparency regress*, which becomes apparent when considering the fundamental goal of transparency, i.e., to unveil the internal mechanisms, causal connections, and interdependencies within an algorithm. In pursuit of this objective, researchers commonly employ an interpretable predictor (referred to as IP_1), designed to elucidate the generation of a specific output (Doshi-Velez & Kim, 2017). However, the challenge arises when we realize that, in principle, there is no inherent reason to believe that IP_1 accurately represents the algorithm's inner workings. It is conceivable that IP_1 may harbor biases, oversight of key internal mechanisms, or instances of manipulation such as reporting forms of "transparency" favoring specific groups interests. The algorithm COMPAS could be transparent in ways that align with Northpointe interests; QII could produce transparency reports that favor the bank instrests. To address this issue, we need to somehow ensure the transparency of IP_1. The best way we know to do so is by means of another interpretable predictor (IP_2). Yet, this only reintroduces the same concerns present earlier, perpetuating the cycle of transparency regress. In this context, there are no safeguards preventing us from suspecting that any interpretable predictor may possess faults or deficiencies.

Arguably, there are two ways to address transparency regression. Either we consider a primal interpretable predictor that is surveyable, contestable, auditable, and overall sanctionable by humans (designated as IP_n);[7] or we take

[7] We are not advocating for requiring all of those practices and properties of algorithms. However, it remains an open question which subset is sufficient for the purposes outlined here.

a leap of faith and accept any given interpretable predictor as reliable. In the former case, regressing down to a primal interpretable predictor is pragmatically undesirable since the accumulation of IPs makes the entire enterprise of opening the black box utterly useless. In the latter case, there is an epistemic pressure to provide reasons as to why an algorithm is reliable when our means of justification (i.e., transparency) are unconvincing.

Another challenge for advocates of transparency, related to transparency regress, is the need to demonstrate that the transparency of any IP_n entails the transparency of $\text{IP}_n - 1$, which in turn entails the transparency of $\text{IP}_n - 2$, and so forth. That is, we need to show that the succession $\text{IP}_n \rightarrow \text{IP}_{n-1} \rightarrow \text{IP}_{n-2} \rightarrow \ldots \rightarrow \text{IP}_2 \rightarrow \text{IP}_1$ effectively maintains transparency. In principle, transparency is possible, but in practice, it either involves a pragmatically undesirable transparency regress or a—possibly ungrounded—commitment to any given interpretable predictor.

The second objection is that reliance demands a sense of cognitive security that transparency might not be able to provide. The primary issue is that, for a transparent algorithm to be considered reliable, we must not only reveal the inner workings of the algorithm but also demonstrate a comprehensive understanding of them. For example, demonstrating that a mole is classified as melanoma based on specific conditions (e.g., size larger than 6mm, asymmetrical, etc.) does not guarantee that we understand why this classification occurs or even that it is the correct classification. To illustrate this point further, consider again the Convolutional Neural Network (CNN) that analyzes ID photos of individuals, identifies facial traits, and classifies each photo as either belonging to a 'criminal' or 'non-criminal' (Wu & Zhang, 2017). While we can show how an algorithm produces a given classification, it is an overestimation to claim that we have understood the sources of criminality. Transparency seem to be able to provide, at best, the former but not the latter.

2.3.2. Computational reliabilism

Transparency posits a perspective that relies on surveying the inner workings of an algorithm to justify its outputs. As mentioned, the merits of this viewpoint encounter difficulties under certain conditions. This is not to suggest, of course, that we should abandon the pursuit of transparency. The value of transparency as an ideal is not in question. However, we must be cautious not to conflate our pursued goals with the legitimate ends of inquiry. The search for transparency oftentimes blurs the line between the valued and the valuable, and what is effectively feasible.

The alternative to transparency that also fosters reliance on algorithms is to embrace their black-box nature. In other words, our justification in believing certain outputs no longer depends on opening the black box. What might seem like an acceptance of defeat is, in fact, a proposal for a new strategy for justifying our beliefs. Computational reliabilism (CR) was initially developed for computer simulations (Durán, 2018; Durán & Formanek, 2018; Durán, manuscript) and has recently been discussed in the context of medical AI (Durán & Jongsma, 2021). The concept behind CR is simple and appealing: beliefs formed by reliable computationally-related processes are better justified than those formed by unreliable ones. Advocates of CR argue that these beliefs do not necessarily arise from revealing the inner workings of the algorithm but from established practices, standards, methods, metrics, and a wealth of knowledge inherent in the design, development, use, and maintenance of algorithms. Importantly, none of these depend on employing a third-party algorithm (i.e., an interpretable predictor). Furthermore, CR operates under a frequentist theory that accepts occasional errors and misclassifications as long as, overall, the algorithm is reliable—that is, it produces outputs with scientific value. Formally, a reliable algorithm is defined as a belief-forming process that consistently renders outputs of scientific value more often than not. Under this heading, we must ask: what makes an artificial intelligence

system reliable? According to CR, three token *reliability indicators* can be identified:

1. RI_1 *Technical robustness*: it focuses on the design, coding, execution, and other technical aspects of artificial intelligence systems that make the system robust, including the collection, curation, storage, and analysis of data;

2. RI_2 *Computer-based scientific practice*: focuses on the practices incumbent to ML-based scientific research and which results from the implementation of scientific theories, principles, and hypotheses, as well as the interactions, debates, and other ways of engaging in standard scientific research; and finally,

3. RI_3 *Social construction of scientific beliefs*: focuses on the broader goals of accepting the AI and its outputs in diverse communities (e.g., scientific, academic, general public, etc.) through debates and other forms of intellectual exchange.

Let us now briefly consider each reliability indicator in turn. Take RI_1, where reliability primarily arises from enhancing the robustness, precision, and accuracy of AI, thereby reducing the error rate. Verification and validation methods, encompassing various sub-categories (see, for instance, Oberkampf and Roy (2010)), exemplify approaches aligned with this goal. Achieving high accuracy and minimizing errors indisputably enhances the reliability of algorithms. Of course, these methods vary among systems since validation methods for computer simulation are, in important ways, different from machine learning (Boge, 2022). Consequently, the quality of an algorithm's outputs is not solely contingent on its numerical proximity to a 'ground truth'. Outputs also hinge on the user's comprehension of their scope, its suitability for the intended purpose, embedded assumptions, trade-offs made for tractability, and the algorithm's representative performance. Thus understood, RI_1 shifts the

focus from the properties of algorithmic outputs (whether they are accurate or not) to the properties of the inquiry methods themselves (e.g., the appropriateness of verification and validation methods for specific goals). In this manner, high precision, accuracy, and a low error rate come with the same assumptions and considerations as the methods that bring them about.

RI_2, on the other hand, directs attention to how scientific theories, hypotheses, principles, and other propositions grounded in science are operationalized into the algorithm or the databases used. It is noteworthy that such embedding may not always occur explicitly and intentionally. Researchers might not consciously operationalize a specific set of scientific propositions into the algorithm. AI systems, particularly when applied in fields like medicine, have the ability to distill scientific knowledge from extensive literature reviews, scientific debates, and various sources. Notably, machine learning and deep neural networks in medical applications often leverage this principle. Given the impracticality or undesirability of explicitly implementing a medical theory into the algorithm, medical machine learning is often trained by selecting and cohesively assembling medical knowledge drawn from reputable journals. An illustrative example is Benevolent AI, a machine learning-based system in drug discovery that asserts its ability to "capture the interconnectivity of all relevant available data and scientific literature using their proprietary Knowledge Graph" (see https://www.benevolent.com/what-we-do).

RI_3 aims to capture the scientific debates conducted with AI methods, emphasizing active involvement rather than mere automation. In a typical scientific setting, algorithm outputs are subject to comprehensive scrutiny and testing within the relevant community before their acceptance. To illustrate this intricate process, consider discovering a new drug. Before it reaches the market for its intended purposes, it must traverse a series of rigorous stages, including clinical control testing, pilot studies, and scientific debates. This journey culminates in final approval for human use, requiring collaboration with other scientists. This collaboration involves engaging in debates on result

interpretation, scope, limitations, and, wherever possible, replication. Furthermore, approval of a new drug also requires independent testing by authorized institutions, such as the FDA in the US and the EMA in the EU. These components collectively contribute to the justification of the output, ensuring that they withstand collective scrutiny and meet the highest standards before integration into practical applications. In this respect, commitments to reliable AI extend to a comprehensive network of scientific methodologies, standards, results, and established traditions. As aptly noted by Elgin, this network enables scientists to build upon each other's work with confidence, ensuring that justified outputs align with the epistemic value prescribed by their respective disciplines (Elgin, 1996, p. 77). Naturally, within this network, disputes and disagreements are expected, encompassing conflicts related to (moral, scientific, political) values, methodological approaches, and the operationalization of varying concepts, theories, and other units of scientific analysis.

Earlier, we referenced BenevolentAI in the context of reliability indicators RI_2. The subsequent debate following BenevolentAI's output, particularly the revelation of baricitinib as a promising candidate to combat COVID-19 effects (Medeiros, 2021), serves as an illustration of how the justification of beliefs can be strengthened through scientific disputes and controversies. Favalli and colleagues, reporting on potential harms associated with baricitinib administration, notably an increase in herpes zoster and herpes simplex infection in specific patient groups (Favalli et al., 2020), prompted a reevaluation of the drug's target patients. The team implementing BenevolentAI, in agreement with Favalli's concerns, exercised caution in recommending the drug for those patients (Richardson et al., 2020). Notably, this debate played a pivotal role in determining the requirements for justifying AI outputs, establishing which errors and artifacts are tolerable, and validating the soundness of underlying assumptions. In essence, it showcases the dynamic and evolving nature of the discourse surrounding AI, emphasizing the importance of rigorous examination and collective consideration in shaping the future trajectory of this field.

Finally, it is important to highlight that CR represents a return to established scientific methodologies and practices, albeit with a unique twist. Now, researchers are compelled to integrate well-accepted principles of algorithmic design, utilization, and maintenance. According to CR, this integration enhances researchers' confidence in AI systems, justifying their belief in the scientific merit of the outputs, and ultimately fostering the reliability of AI. Remarkably, all of this is achieved without opening the black box.

Objections to Computational Reliabilism

Just as we observed with transparency, CR also has important challenges to overcome. A notable concern arises from the frequency at which beliefs are justified. While in many instances, the algorithm's output may indeed have scientific merit, leading researchers to deem the system reliable, there's a valid worry that the rare instances of system failure could have profound implications. To illustrate this, consider a medical AI providing various oncological diagnoses. Assume the system is generally deemed reliable because its outputs align with diagnoses made by human oncologists, demonstrating scientific merit. Users trust and treat the medical AI accordingly. Now, envision a scenario where the system misdiagnoses one single patient, inaccurately categorizing them as healthy instead of detecting a form of cancer. Under CR, even if this specific output lacks scientific merit, the medical AI as *a whole system* is still considered reliable. The critical question that arises is whether physicians are epistemically entitled to rely on the system after such a failure or if a significant reevaluation of the conditions under which the system operates is imperative. This example underscores the potential limitations and challenges associated with relying on CR in complex, high-stakes domains such as medical diagnosis.

A second limitation of CR is associated with the availability of reliability indicators. It is improbable that we are in possession of all the pertinent indi-

cators for a given AI system.[8] In such scenarios, researchers are tasked with evaluating the reliability of their system based on a limited set of indicators. Furthermore, the few available indicators may wield disproportionate influence over the attributed reliability of any AI system. To illustrate this counterfactually, our assessment of the reliability of a system would most likely differ had we had access to all the relevant indicators. We term this phenomenon the *tyranny of the few*, underscoring the importance of having available as many and as diverse reliability indicators as possible. Ultimately, it is still unclear how many reliability indicators are necessary to mitigate the tyranny of the few.

Despite these concerns, CR represents a significant advantage in evaluating the reliance on AI systems. One key aspect is the "decentralized" nature of the reliability indicators. This means that there are various sources of indicators available to us and that these sources operate independently from each other (e.g., validation is not contingent on scientific debates). Another crucial advantage of CR is that humans *are* in the loop in a meaningful way. This contrasts with transparency, where humans typically play a passive role in trying to understand an explanation or an interpretable predictor.

2.4. THE "EXTRA FACTOR"

Now we turn our attention to discussions revolving around the "extra factor". We present two main positions in the specialized literature in connection with the conceptual and normative possibility of trusting (or not trusting) AI systems. It is important to note that these positions are rather absolute in their views and in direct opposition to each other. Whereas one states that trusting AI is either not possible or undesirable, the other advances claims for its plausibility. In what follows, we discuss each one in turn.

[8] It is crucial to acknowledge that not all reliability indicators are universally applicable. For instance, while validation might be more pertinent for empirically-driven AI (e.g., climate change and mental mechanisms), it might hold less relevance for theoretically-driven AI (e.g., the origins of the universe and protein folding).

2.4.1. Trustworthy AI is neither possible nor desirable

In Section 2.2, we mentioned how interpersonal accounts of trust place humans at the center of their analysis. Drawing on similar philosophical ideas, adversaries of the possibility of trusting AI base their skepticism on the (rather obvious) differences between humans and machines. In this context, two main claims are set out. The first claim is that trust in AI is conceptually impossible because genuine trust in an inanimate entity (such as AI) is a category mistake. Scholars endorsing this claim typically argue that trust in AI would be incompatible with any philosophical account of interpersonal trust. The second claim is normative in nature and states that we *should not* place our trust in AI systems since this would lead to undesirable consequences. These amount, for instance, to the fact that a responsibility gap emerges given that trusting an AI system enables AI developers and designers to elude their (moral) responsibility by outsourcing it to AI systems (Starke et al., 2022). Of course, these two claims are not to be considered completely separated: usually, authors that deny the theoretical possibility of trusting AI also endorse the claim that AI systems are entities that *should not* be trusted. However, we keep these claims separate for analytic purposes. In what follows, we critically discuss accounts supporting these two positions and point out that they represent a considerable challenge to anyone defending the possibility and desirability of trust in AI.

We could identify three main positions pertaining to the *impossibility* or *undesirability* of trusting AI systems. In the following, we address each one in turn. The first account is known as the *affective account of trust* and consists of identifying the "extra factor" with the favorable disposition or goodwill of the trustee to fulfill the particular goal they have been entrusted with. This requires that the "trustee is favourably moved by the trust placed in them"

and that "the trustee has the trustor's interests at heart." (Ryan, 2020, p. 12) This account of trust emphasizes the value of the interpersonal aspects of the trust relationship, such as emotions, psychological states, and motivations (Ryan, 2020). For example, as Ryan points out, when we trust our friend, we assume—following the affective account—that she is willing to keep our secret because she does not want to wrong us and not because she would otherwise run into trouble. That is to say, the motivations behind her willingness to keep our secret come to the fore in the affective account of trust: our friend does not keep the secret merely for self-interest but rather because she cares about us.

The second account is known as *normative trust* because it refers to the normative expectations that the trustor has on the trustee. This account takes that the trustee *ought to* fulfill the commitments that emerge when the trustor decides to entrust her with a certain goal or task.[9] For instance, if a friend asks us to keep her secret, we *should* do so in virtue of the fact that she is entrusting us with a piece of information that we are not supposed to share. Clearly, this account requires the trustee to be the bearer of moral responsibility. In particular, in case a breach of trust occurs, the trustee needs to be a suitable receiver of blameworthiness. As Hatherley points out, "I rely on you when I predict that you will behave in a certain way, though I trust you when I judge that you ought to behave in a certain way." (Hatherley, 2020, p. 3) It is obvious that both the affective and normative accounts require that the trustee is aware of the fact that the trustor has placed trust in them.

The third account, known as the *rational choice account*, sees the trustor as making a rational evaluation when deciding to trust the trustee based on the likelihood that the trustee will behave as intended towards the fulfillment of a certain goal. This does not entail any kind of demand (normative or otherwise) on the trustee for the trustor to engage in a trust relationship.

[9] As Ryan rightly points out, this does not mean that the trustee will have to fulfill every task she has been entrusted with. The moral acceptability of the particular task in question needs to be secured before saying that the trust relationship entails normative expectations.

It also does not require the trustee to be moved by the "right reasons" to act as the affective account postulates. It only requires that, based on a regular frequency, the trustee behaves as intended. So, contrary to the other two accounts of trust, motivations and normative expectations do not play a central role in the rational choice account. As such, rational trust "is reliant on specific features of a situation, rather than the relationship between the trustor and the trustee." (Ryan, 2020, p. 11)

Quite intuitively, affective and normative trust set a standard for the extra factor that cannot be fulfilled by AI systems qua inanimate entities without attributing to them genuinely human traits (e.g., agency, emotions, motives, etc.). In fact, if we were to consider affective trust for an AI system, we would need to ascribe to it some forms of human agency and emotional states (awareness, empathy, compassion) to be able to say that the system is "willing" to live up the demands of a trust relationship. However, attributing these genuinely human traits to AI systems seems to be unwarranted. Moreover, and as mentioned before, it seems inappropriate to have sentiments of betrayal and deception—that usually would be in place when affective trust is breached—towards inanimate entities.

Something similar can be said for normative expectations that are central to the normative account. Since AI systems are unaware of any form of trust that we may pose in their functioning, normative expectations on their performance (i.e., that they should work as we trust them to do) would be utterly misplaced. Therefore, due to the impossibility of AI systems to be the appropriate bearers of normative demands, according to Hatherley, "the pursuit of trustworthy AI represents a notable conceptual misunderstanding." (Hatherley, 2020, p. 3) In the face of what has been said so far, the attribution of trust to AI systems would require us to anthropomorphize AI systems by attributing to them relevant human traits (such as some forms of agency or consider them receivers of moral responsibility, for instance).

These considerations seem to suggest that only a rational choice account

of trust is attributable to AI systems. This is the case if we want to avoid the anthropomorphization entailed by the affective and normative accounts. However, contrary to the other two accounts of trust considered, the rational choice account does not require any "extra factor" to be in place. Indeed, this account does not demand attention on the motivations of the trustee, but rather on "a rational calculation of whether the trustee is someone that will uphold the trust placed in them." (Ryan, 2020, p. 4) As such, the rational choice account does not require us to make a conceptual distinction between morally loaded reactions of betrayal or being disappointed when the trustee fails to meet the trustor's expectations. Instead, we are solely concerned with the frequency with which the trustee upholds the trust placed in them.

Now, it is crucial to conceptually distinguish the "extra factor" from mere reliance. Therefore, Ryan's considerations[10] work in support of the claim that trustworthy AI is a conceptual mistake and "one needs to either change 'trustworthy AI' to 'reliable AI' or remove it altogether." (Ryan, 2020, p. 17) Thus, authors who hold a skeptical position regarding trustworthy AI conclude that, since it is impossible to trust AI without anthropomorphizing it, AI systems cannot be seen as genuinely trustworthy. Therefore, trusting an AI system amounts to misplaced trust (Ryan, 2020, p. 4). Scholars defending the conceptual impossibility of trust in AI deny the possibility of it fulfilling the conditions needed to account for the "extra factor", regardless of its nature.

Let us now turn to the second claim, that is, that trusting AI systems is *undesirable*. Starting from the assumption that trust needs to be a relation between peers in which beliefs and promises are made, Bryson (2018) defends both claims, i.e., that trust in AI cannot occur and should not be pursued. However, in her critique, she particularly emphasizes the danger of ascribing to AI human-like features such as trust and trustworthiness. The danger lies, according to Bryson, in the fact that developers and companies owning AI sys-

[10] That a rational choice account of trust is nothing more than mere reliance has been pointed out also by Nickel et al. (Nickel et al., 2010).

tems could use this to outsource their responsibility to these systems and evade moral blameworthiness when something goes awry. In Bryson's words: "malicious actors will attempt to evade liability for the software systems they create by blaming the system's characteristics, such as autonomy or consciousness." (Bryson, 2018) In the face of these considerations, she concludes by stating that "AI is not a thing to be trusted. It is a set of software development techniques by which we should be increasing the trustworthiness of our institutions and ourselves." (Bryson, 2018) Thus understood, the undesirability of attributing trust to AI systems amounts to the fact that, among others, a responsibility gap would emerge. Moreover, it would confer to AI systems capabilities that need to remain in the domain of human expertise, creating unrealistic expectations of what AI systems can effectively achieve. A similar critical position is also shared by Tallant (2019), who states that efforts pushing forward trustworthy driverless cars, for example, are nothing else than a marketing move (Tallant, 2019, p. 116).

Along similar lines but focusing on the nature of trust in medical contexts, Decamp and Tilburt (2019) advances the claim that talking about trust in AI could lead to a decrease of trust in medical practitioners since they could, on occasions, not achieve the level of accuracy secured by some AI systems. However, mistakenly confounding accuracy and reliance with proper trust can lead to devaluing the physicians' abilities and expertise. As Decamp and Tilburt point out "(p)romulgating trust in AI could erode a deeper, moral sense of trust." And continues: "(t)rust properly understood involves human thoughts, motives, and actions that lie beyond technical, mechanical characteristics. To sacrifice these elements of trust corrupts our thinking and values." According to this author, therefore, we should not put our trust in AI systems if we want to preserve the importance of the morally loaded form of trust we are ready to put into human physicians (Decamp & Tilburt, 2019).

To sum up, the main reasons advanced by authors critiquing the possibility and desirability of trusting AI can be boiled down to the following points.

First, trust in AI would lead to the danger of impoverishing the notion of interpersonal trust in its morally loaded sense (Ryan, 2020), reducing it to not much more than mere reliance (see the critique of the rational choice account). This would blur the line between two clearly distinguished concepts, i.e., reliance and trust. Moreover, it would lead to the impossibility of having a discourse about (genuine) trust in AI without falling into the trap of its unwarranted anthropomorphization (Ryan, 2020). In other words, neither the requirements for the normative nor the affective account can be fulfilled without attributing human traits to AI systems. Second, trust in AI is undesirable because it would lead to the unjustified attribution of responsibility to computational systems, representing a possibility for designers, developers, and companies to evade (moral) duties intrinsic to their professional role. This seems to be particularly unsatisfactory in situations in which the allocation of responsibility and blameworthiness plays a particularly salient role.

In the face of these substantial criticisms regarding the very conceptual possibility and normative acceptability of trusting AI systems, several efforts have been made to respond to these critiques. In the next section, we analyze different positions of scholars attempting to conceptualize trust (and trustworthiness) so that it can be meaningfully used in AI-mediated contexts. We will present the most prominent positions and critically analyze their merits and shortcomings in view of what has been discussed so far.

2.4.2. Trustworthy AI is possible and desirable

Even though arguments advanced in defense of the claim that genuine trust in AI is not possible have merit, it can still seem unsatisfactory to exclude the possibility of trusting AI systems. In fact, given the ubiquitous nature of AI-based technologies, they play a relevant role in mediating interpersonal relationships, and their influence is increasingly interwoven in our social structures (von Eschenbach, 2021). Moreover, the crucial role of trust in accommodating

complexity (Lee & See, 2004) and the fact that we are increasingly vulnerable to AI technologies (Chen, 2021) have motivated scholars to continue pursuing a suitable conceptualization of trust in AI. However, as Nickel, Franssen, and Kroes point out, "(a)ny applicable notion of trustworthy technology would have to depart significantly from the full-blown notion of trustworthiness associated with interpersonal trust." (Nickel et al., 2010, p. 429) In the remainder of this section, we sketch out some of the most prominent accounts in favor of trusting AI.

We could recognize four overarching approaches according to how the challenges raised in the previous section are faced:[11] 1. approaches that admit some form of (minimal) agency in AI (Chen, 2021; Lewis & Marsh, 2022; Starke et al., 2022)—*agency-based approaches to trust*; 2. approaches that deny the normativity and affective dimensions of trust in AI, thus taking a non-normative and non-affective position (Ferrario et al., 2020, 2021)—*the rational choice account of trust*; 3. approaches that take a normative stance but without making AI systems the bearer of moral obligations (Nickel, 2022)—*the discretionary view on trust* ; and, finally, 4. *reductive accounts of trust*[12] that take AI systems to be the indirect recipients of our trust (Sutrop, 2019).

To start with the first approach, i.e., the *agency-based approach*, Starke and colleagues build their "argument on the rather strong assumption that one can reasonably attribute agency to AIs." (Starke et al., 2022, p. 157) They do not ascribe full agency to AI systems (in the sense of mental states such as beliefs and desires) but rather a form of minimal agency or agency in a weak sense. Such a minimal agency stems from the embedding of AI systems in socio-technical contexts along with their per-designed ability to adapt, evolve, and influence it. To make their case, the authors consider Latour's case of the Berlin key that cannot be removed from the lock without locking the door (Latour, 2000). In this example, the key plays, by design, a role towards a

[11] Of course, we do not have the pretense to be exhaustive in this regard.
[12] We adopt Nickel's label for this form of trust (Nickel, 2022, p. 5).

certain goal, namely, making sure that the door is locked from the inside. In this context, the authors take that "by playing its part in a complex network of actors that would not be feasible without its material manifestation, the key contributes to the disciplinary relationship itself." (Starke et al., 2022, p. 157) For this reason, the key is not to be seen as a mere object but rather as an *agent* in a specifically described environment that contributes to the intended purpose. So, by analogy, if a key can be considered an agent in this minimal sense, these authors do not take it to be far-fetched to attribute agency to AI systems as well. Drawing on this assumption, Starke and colleagues take that trusting AI systems is possible if considered across three different dimensions. Those are intentionality, reliability, and competence. While reliability amounts to the avoidance of malfunctions and competence to validity and accuracy of predictions, the intentionality of a system can be perceived, again, along the lines of a weak sense of agency. Therefore, so goes the argument, if an AI system brings about discriminatory effects, one has less of a reason to trust its intentions (in the weak sense of the term). However, if, on the other hand, a system has a high level of interpretability, one has good reasons to trust the system's intentions to bring about a certain goal (Starke et al., 2022, p. 159).

A similar position that assumes some form of agency and intentionality in AI systems is also taken by Chen (2021) and Lewis and Marsh (2022). Chen assumes a form of "derived intentionality" in AI systems that stems from the ability to display what can be considered intelligent behavior, such as playing chess or performing natural language processing (Chen, 2021, p. 1435). Let us note that AI systems' intentionality is, also according to Chen, not to be understood in a strong sense. The author rather states that "(a)s products of human intentional action, they have a *prima facie* claim to some form of derived intentionality." (Chen, 2021, p. 1436) So, supporting a middle-way position between defenders of trust in AI and accounts that state only the occurrence of reliance without trust in AI, Chen sees what he calls *trust-responsiveness* as the most suitable alternative: "a disposition to prove reliable

under the trust of others." (Chen, 2021, p. 1441) In order to achieve this, engineers need to put efforts into making sure that AI systems are reliable and robust so that we are justified to trust them (see Section 2.3).

In a similar vein, Lewis and Marsh take that a functionalist view on intentionality and agency entails a *functionalist view* on trust (Lewis & Marsh, 2022). According to these authors, excluding the possibility that agency and intentionality can be meaningfully attributed to AI systems would be an unjustified instance of human exceptionalism (Lewis & Marsh, 2022, p. 47). On the contrary, from a functionalist angle, the ability of AI systems to betray our trust is deeply connected with their purpose and the possibility of deception. As such, considerations regarding the transparency of AI systems' goals, for instance, come to the fore when questions of whether we are justified in trusting them need to be considered (Lewis & Marsh, 2022, p. 45).

Whereas these arguments support some form of agency in AI systems for the attribution of trust, one could still be skeptical that this is not the right kind of agency for genuinely trusting AI. In fact, one could make the case that this form of minimal agency is not enough to consider AI systems to be able to live up to the normative expectations that characterize trust relationships. For instance, questions about the attribution of responsibility and accountability to these systems arise. How minimal is this "minimal agency"? Are AI systems to be considered as equally responsible and accountable as human agents? In sum, requirements in terms of minimal agency still raise concerns about the actual normative expectations of AI systems.

This brings us to the second main position on trust in AI, which does not require any form of agency for AI systems but rather focuses on advancing a non-normative and non-affective account. This position—i.e., the *rational choice account* of trust—is advanced by Ferrario and colleagues (Ferrario et al., 2020, 2021). Key to these authors' account is that trust in AI comes in degrees, and we do not need to consider AI systems as suitable bearers of affective or normative expectations in order to meaningfully say that we trust them. Let

us, in particular, consider two of the three forms of trust conceptualized in their incremental model of trust. According to these authors, a minimal form of trust (they call it *simple trust*) is secured if we rely on a system without constantly updating our beliefs regarding its reliance. In their words, "trust involves economising on monitoring." (Ferrario et al., 2021, p. 437) That is to say, the readiness of the trustor to rely on the AI without control is the step needed to move beyond mere reliance and trusting an AI system. Consider, for example, a medical AI system that provides physicians with treatment recommendations for their patients. One first phase of reliance *only* is in place when the physician interacts with the system and forms beliefs regarding its performance, accuracy etc. In this phase, the physician engages in the evaluation of the system's performance to assess its reliability (as discussed in section 2.3, this can be done in different ways). After a certain amount of positive interactions with the system, the physician will likely consider it reliable. At this point, she can start to rely on it without seeking further evidence supporting the fact that her reliance is indeed justified. In other words, the cognitively burdensome activity of monitoring the AI system's ability to deliver a suitable treatment recommendation for her patients and updating her beliefs moves into the background. As soon as the need to monitor the system disappears altogether and the physician is ready to rely on the AI without control, an instance of simple trust occurs (Ferrario et al., 2021). Simple trust is thus a property of the trustor (the physician) and not of the trustee (the AI system providing treatment recommendations) (Ferrario et al., 2020). Therefore, it is important to consider that for simple trust to be in place, we are not required to deem the system trustworthy overall. On the other hand, a situation in which we are ready to rely on an AI system without monitoring it and, on top of this, consider the AI to be trustworthy is, according to Ferrario and colleagues, the most complete form of trust in AI. They call this form of trust *paradigmatic trust*. The authors emphasize that the latter is what is usually referred to in the literature revolving around trustworthy AI, even though

fulfilling the conditions needed to attribute simple trust would be enough to meaningfully talk about trust in AI. Not considering the affective and normative dimensions of trust in AI in their account of simple trust, these authors aim to maintain a conceptual distinction between trust and reliance without running the risk of anthropomorphization addressed in the previous section.

An objection to the notion of simple trust could still be advanced by questioning whether it substantially differs from mere reliance, as the authors claim. In fact, backed into their concept is the assumption that assessing the system's reliability requires constant monitoring that is no longer needed once its reliability is effectively confirmed. From that point on, we have simple trust in the system. However, it remains unclear why, after securing the system's reliability and giving up our critical monitoring, we go a step beyond relying on it. In other words, why does reliance need constant monitoring while trust does not? It seems plausible to think of situations in which relying on the fact that something will be the case does not require a constant update of our beliefs. Once again, questions emerge when normative (and/or affective) considerations remain unconsidered, and we still want to maintain a conceptual distinction between reliance and trust.

These considerations bring us to the next position on trust in AI. In fact, for some authors, we cannot simply disregard the normative dimension of trust, excluding it from the picture. The third position we consider is advanced by Nickel (2022), who develops a *discretionary account of trust* in which the normativity of trust comes to the fore.[13] According to this view, trust manifests in the discretionary authority that, for example, physicians decide to attribute to a medical AI involved in medical decision-making. Discretion is understood as a "circumscribed authority accorded to another entity" (Nickel, 2022, p.

[13] While the normative dimension of trust is central to the discretionary account, it does not encompass an understanding of trust in affective terms. So, while it takes a very different stand regarding the normativity of trust in AI compared to the rational choice account, it shares with the latter a lack of emphasis on motives and desires. We thank an anonymous reviewer for encouraging us to clarify this point.

7) and, according to Nickel, "(t)ransferring discretionary authority to another entity carries distinctive moral weight." (Nickel, 2022, p. 4) According to the authors' view, discretionary authority amounts to trust only if predictive and normative expectations on the trustee (i.e., in our case of interest, on AI systems) are in place. For example, consider a physician who decides to attribute discretionary authority to an AI system that estimates patients' likelihood of being admitted to the intensive care unit after surgery. In attributing discretionary authority to this system, the physician holds normative expectations on it as she expects the system to function as intended, i.e., to function as it *ought* to. The normative dimension goes, thus, hand in hand with the purpose and goal of the system, that is to say, with what the system has been designed and implemented for. As Nickel points out, "(w)hen such function-based expectations are relevant to the needs and goals of clinicians, they provide the basis for giving some of the clinician's own discretionary authority to the AI application, allowing it to (help) answer questions that previously went unanswered, or that were previously answered using other means." (Nickel, 2022, p. 7)

In view of this, how can this normative dimension be accounted for without falling into the unjustified anthropomorphization of AI we discussed in the previous session? According to Nickel, when discretionary authority is attributed to an AI system, the AI is the *object* of a moral obligation but not its bearer. This means that physicians do not trust the AI system directly, they rather trust AI designers, developers etc. "*through* the AI application." (Nickel, 2022, p. 6) So, AI practitioners have the (moral) obligation to ensure that the AI system that has been granted discretionary authority functions as intended in respect of shared values such as fairness and efficiency, for instance (Nickel, 2022, p. 4). The discretionary account allows thus to preserve the normative dimension of trust without having to take a stand regarding the thorny issue of having to attribute some form of responsibility directly to AI systems. In fact, AI practitioners are taken to be responsible for guaranteeing

that a certain AI system is up to the expectations of physicians who are ready to confer discretionary authority to it.

According to what has been said so far, a possible limitation of Nickel's account is a lack of clarity about the actual locus of our trust. In fact, it can be objected to Nickel's view on trust that it effectively amounts to trust in AI practitioners. While Nickel leaves this question open in his conceptualization of trust in AI, there are authors who clearly defend the position that the only possible form of trust in AI is in the human beings involved in the development of AI systems and not the systems directly.

Relatedly, the fourth and last position that we consider concerning trust in AI—the *reductive account*—has been defended, among others, by Sutrop (Sutrop, 2019). She argues that "when we speak about trust in AI, in reality, we are speaking about trust or distrust of individuals and institutions who are responsible for developing, deploying and using AI." (Sutrop, 2019, p. 512) This position thus takes that while AI systems can be meaningfully relied upon, the object of our trust can only be the humans involved in the development of AI systems (e.g., designers, engineers, computer scientists etc.). So, according to this account, we indirectly trust an AI system because we trust the human beings behind its development, and they have the moral obligation to make sure that AI systems meet the expectations we pose in their functioning. However appealing, this position does not come without problems. For instance, the self-learning and adaptive abilities of most AI systems are an indication that it is not always clear to what extent computer scientists and engineers can foresee a problematic behavior of the system that is possibly perceived as a breach of trust by the end-user (say, a medical AI that leads to a misdiagnosis). Therefore, it is not a straightforward solution to consider trust in AI as amounting to trust in the humans behind the development of the system instead of the system itself.

Considering what has been said so far, it becomes clear that views regarding what trust in AI amounts to strongly diverge. The responses to the critiques

advanced by scholars who are skeptical regarding its conceptual possibility (and normative stand) are formulated in very different ways—all with their weaknesses and strengths. What is common to all the positions defending the possibility of trust in AI, however, is that some form of system reliability must be accepted.

2.5. FINAL REMARKS

In this article, we put forward an analysis of trust and trustworthy AI aiming at dissecting its main components, possibilities, and limitations. To this purpose, we started by making an analytic distinction between reliance and an "extra factor", both requirements present in standard accounts of interpersonal trust in the philosophical literature. With this distinction in mind, we considered two opposing views on how to secure the reliance of AI algorithms, namely *transparency* and *computational reliabilism*. We showed that even though some form of reliance is often taken for granted in the literature on trustworthy AI, it is not a trivial matter to find a way to successfully account for this necessary desideratum. We argued that transparency, understood as methods aiming at opening the black box, are typically taken to be the gold standard to assess the scientific merit of an AI system's output. Even though the search for transparency can be seen as intrinsically valuable, we argued that it suffers from considerable shortcomings worth debating. In this respect, two chief problems were presented and briefly discussed. One that shows that transparency might imply some form of *transparency regress*, in which case the justificatory status of the algorithm is pragmatically and epistemically compromised. The second issue is that transparency demands for a kind of cognitive security difficult to obtain, thus casting doubts on the kind of understanding that it is able to offer. Despite these, transparency is a very valuable goal to pursue.

We also discussed computational reliabilism as the chief contender to transparency. Contrary to the latter, computational reliabilism does not require "opening" black box algorithms. Instead, justification comes from *reliability indicators*, many of which are quite familiar to us as they draw from standard scientific practice. As we discussed, computational reliabilism is also limited in important ways. We mentioned the frequency by which beliefs are justified, and the tyranny of the few. Despite these, computational reliabilism still proves to be a suitable method to account for the reliance condition needed to secure trust in AI systems.

With these results in place, we moved on to the analysis of the "extra factor", understood as the second component in the definition of trust in AI. Here, we showed that scholars holding a skeptical view regarding the very possibility and desirability of trust in AI systems advance convincing arguments that need to be accounted for. In particular, we discussed the unwarranted anthropomorphization of AI systems and possibly undesirable consequences in terms of responsibility gaps. As we further considered, the accounts of authors trying to respond to the criticisms advanced are many and contrasting. Among others, we sketched some approaches that try to exclude the normative dimension of trust from the picture, while others attribute to AI systems either some form of minimal agency or require to trace trust back to the human beings behind its development. Even though these efforts have merit, we pointed out that some issues remain unresolved. For example, it is unclear whether the agency attributed to AI systems accounts for the "extra factor" or whether it is legitimate to exclude the normativity of trust and thus blur the line between (mere) reliance and genuine (morally robust) trust. As we have shown, the debate is vast and opposing views are defended. This article reconstructs key fundamental aspects of the debate in an attempt to bring clarity and order to an otherwise fragmented debate.

Admittedly, the way in which we structured the debate necessarily leaves out important considerations about trust and trustworthy AI that deserve fur-

ther attention. One of particular importance is the right level of stakeholders to consider. In this article, we narrowed down the scope of our analysis to individual interactions with machine learning systems. For instance, we referred to the trust (or distrust) that a physician can have towards an AI system providing treatment recommendations for their patients. However, this is not the only dimension across which trust can be established. As one can distinguish different levels on which information is created and transmitted,[14] along similar lines, one could say that trust relations develop across three different dimensions: interpersonal (or individual), collective, and institutional.

Another issue stemming from focusing exclusively on an individual form of trust in AI is that trust can exceed the restricted dynamics of interactions we considered here. For example, what about the trustworthiness of a hospital in providing physicians with the assistance of AI systems for their decision-making? How can its trustworthiness (or lack thereof) be secured? An adequate answer to these questions would require going beyond the considerations of the reliability and the "extra factor" pertaining to a particular AI system. We would need to account for a whole other range of considerations aiming at evaluating the credibility and trustworthiness of the hospital as an institution in relation to AI applications used for medical decision-making. For example, which design choices were made? Which values have been designed for? To what extent has stakeholder engagement taken place at the different stages of the design, development, and deployment of a certain AI system? To this end, it is also necessary to critically consider whether the analytic distinction between reliance and the "extra factor" we made throughout this article would still apply. For example, what would be the nature of the "extra factor" if we want to analyze trust on this more high-level dimension? Or should we leave this analytic distinction aside and work towards a completely different

[14] Goldman (2019) pointed out that the creation and transmission of knowledge can occur throughout three different dimensions: interpersonal, collective, and institutional. We think that this consideration can be transferred also on how trust is established. For more on trust in institutions, see Alfano and Huijts (2019).

way to assess trust and trustworthiness? In fact, it is not straightforward to know what kind of normative commitments or affective attitudes are appropriate when we trust a group of people or certain institutions. These are all extremely relevant and timely questions that would have exceeded the scope of this article but that need nevertheless further research and critical scrutiny.

3

Informativeness and Epistemic Injustice in Explanatory Medical Machine Learning[1]

3.1. Introduction

Artificial intelligence (AI) plays an increasingly morally relevant role throughout various domains, bearing the potential to significantly influence crucial decision-making processes that are usually reserved for human expertise. There are studies showing that AI-based methodologies, in particular machine-learning (ML) techniques, are paving the way for promising developments in high-stakes fields, such as medicine and healthcare (e.g., Esteva et al., 2019; Topol, 2019).

Unfortunately, the excitement associated with these developments is not always justified. Epistemic limitations in connection with the way in which

[1]This chapter is based on the following article:

Pozzi, G. & Durán, J. M. (2024) From ethics to epistemology and back again: informativeness and epistemic injustice in explanatory medical machine learning. *AI & Society*. https://doi.org/10.1007/s00146-024-01875-6

these systems operate give rise to serious ethical concerns. Central to the success of ML systems is their capacity to reconstruct sets of rules from large datasets, which in turn can reveal new patterns in the data (Alpaydin, 2014). Due to the large amount of data processed by these systems and the complexity of the calculations, they become epistemically opaque to human enquirers (Beisbart, 2021; Durán & Formanek, 2018; Humphreys, 2009).[2]

The consideration of how the epistemic limitations of ML systems lead to ethical issues has, justifiably, gained a central stage in current debates and has given rise to a wealth of literature on the topic. For example, scholars have pointed out that the epistemic opacity of ML algorithms is connected to ethically relevant problems that range from fairness-based concerns (Zarsky, 2016) to questions of accountability (e.g., de Laat, 2018) and the trust we are justified in attributing to these systems' outputs (e.g., Hatherley, 2020). Other authors connect the epistemology and ethics of AI even more explicitly. For instance, Grote and Berens (2020) identify the epistemological pitfalls of ML systems implemented in medicine (e.g., issues of peer disagreement and epistemic uncertainty) as directly conducive to crucial ethical implications (e.g., problems of paternalism, patients' informed consent, and defensive medicine). Similarly, Bjerring and Busch (2021) recognize in the black-box nature of AI systems the concrete possibility that it undermines the ethical ideal of patient-centered medicine. Relatedly, Babushkina and Votsis (2022) consider primarily how epistemological constraints of ML systems in the context of medical diagnoses lead to ethical considerations in terms of epistemic responsibility.

Thus, the relevance of showing the bearings of epistemological issues on ethical concerns has been recognized and extensively analyzed. However, an analysis of the extent to which ethical considerations influence the epistemology of ML is still lacking, as Russo et al. (2023) point out in a recent paper. In addressing this research gap, these authors take an approach that aims to *explicitly* point out the interconnected nature of the ethics and epistemology

[2]In a less nuanced and more metaphorical way, these algorithms are known as "black-boxes".

of AI. They do so without presupposing that epistemological considerations are prior to ethical considerations, as is often assumed in the debate.[3] This is a position that we endorse in this paper. We share these authors' approach in seeing "the equal importance of the two fields [i.e., of ethics and epistemology] and their intertwinement." (Russo et al., 2023, p. 2) We are, in fact, committed to the same goal of overcoming a division in the ethics and epistemology of AI that is not tenable if we want to understand the impact of these technologies on society. However, our approach differs from theirs in at least three respects.

First, Russo et al. take a holistic approach that accounts for "the process of design, implementations and assessment of AI that simultaneously considers ethics and epistemology, and the expertise of the actors that inquire into these two." (Russo et al., 2023, p. 2) We take a more fine-grained level of analysis than this high-level approach. In fact, we analyze the role that the intertwined dimensions of epistemology and the ethics of AI should play in a very concrete setting, that is, one in which the medical decision-making of an ML system is analyzed regarding its displacement of physicians from their epistemically authoritative position.

Second, by analyzing a concrete case, we provide more reasons as to why the available approach in the literature is limited. In Section 3.2, we dissect it in its parts, analyze its underlying logic, and show its shortcomings. In this sense, our analysis considerably expands on one of the fundamental premises made by Russo and colleagues—that is, the ethics and epistemology of AI are largely disconnected in the current debate.

Finally, in our analysis, we consider the epistemology of AI as a genuinely and inherently normative dimension and place a strong emphasis on this point throughout the entire article. This comes particularly to light in our concrete case analysis in Section 3.3, which underscores that what a physician *should*

[3] As we point out later in the paper, this is one of the limitations we recognize in the approach available in the literature.

believe and the explanation she *should* accept is partially determined by an ethical feature of the particular situation in focus.[4] In making a distinction between epistemic (e.g., explainability) and normative (e.g., fairness) aspects, Russo et al. (2023, p. 10) do not seem to embrace this aspect, which is, however, central to our analysis.

We are convinced that the role that ethical features play in the epistemology of AI needs special attention, and it is the overall aim of this paper to lay the groundwork for a more explicit discussion of this important aspect. To effectively show the relevance of our argumentative goals, some considerations are in order, starting from the kind of ML systems that are the object of our analysis.

In this contribution, we focus on ML systems that displace or risk displacing physicians from the center of knowledge production. Here, their courses of action are dependent upon those indicated by the ML system involved in the decision-making process. Even though this scenario is surely undesirable, as we would expect these systems to remain under the ultimate control of experienced professionals—and particularly for ML systems implemented in medicine—it is, unfortunately, not too far-fetched. As we will show, some currently deployed ML systems dramatically disappoint this expectation. For example, algorithmic Prediction Drug Monitoring Programs (PDMPs) used to predict patients' likelihood of opioid misuse and currently implemented in the USA to inform clinician's decisions on a daily basis have been shown to be *de facto* replacing—instead of merely supporting—medical decision-making (cf. Oliva, 2022; Szalavitz, 2021). Furthermore, these ML platforms are opaque to their end users (i.e., physicians) in that they lack insight into how the algorithms classify patients as being at a high risk of opioid abuse (Szalavitz, 2021). Lastly, the proxies used to determine patients' risk scores are not necessarily indicative of opioid misuse and can result in misleading ML outputs that do not represent a patient's actual drug consumption (Oliva, 2022). Given that these

[4] We clarify the nature of the moral and epistemic "should" and their relation in Section 3.3.

are "law enforcement-developed digital surveillance systems" (Oliva, 2022, p. 51), physicians are expected to act upon the outcomes generated, even though they lack any kind of understanding regarding how the system's results are obtained. In fact, due to these constraints, physicians are in no position to determine whether a patient is justifiably considered at risk of drug misuse or whether the systems establish disparate correlations that are not reliably connected to a person's drug consumption (Oliva, 2022; Pozzi, 2023a, 2023b). Thus, these systems are incontestable and are clearly displacing physicians from their epistemic and moral authority, creating undesirable effects that have led to patient abandonment and denial of medication (Szalavitz, 2021).

In the face of the harm that epistemically authoritative systems similar to ML-based PDMPs can generate, which theoretical approach can be functional in effectively addressing the epistemic and moral issues they bring about? This question motivates our analysis of the relationship between the epistemology and ethics of ML. We label approaches that consider the bearing of epistemological issues on ethical concerns but neglect the impact of ethical elements on epistemic features of situations involving ML as the *informativeness account*. We elaborate on the assumptions built into this account and analyze an example in the field of explanatory ML in healthcare that does not square well into it. We argue that in cases similar to the one under scrutiny, it is paramount to consider the role that ethical properties play in influencing and regulating epistemologically relevant aspects of ML (e.g., explanatory ML). We dedicate the main part of this contribution to the effort to make explicit the compelling nature of this claim.

With these considerations, we gain a purchase on how certain epistemic practices with ML in medicine (such as the ones illustrated in our case in Section 3.3) expose patients to diverse forms of epistemic injustice. We are particularly interested in showing how, following the logic of the informativeness account, ML algorithms *epistemically objectify* patients. The section on epistemic injustice aims to further substantiate the claim that we need an

approach in the ethics and epistemology of ML that considers the impact of ethics on epistemology. Although these considerations strongly suggest the need to expand the informativeness approach, it is beyond the scope of this article to show how this is effectively done.

The remainder of this article proceeds as follows. We provide a description of what we define as the informativeness account (Section 3.2). We then substantiate our case through an example of explanatory ML in medicine that cannot be adequately accepted when considered within the framework of the informativeness account (Section 3.3). Finally, we consider how the situation experienced by the patient in our example leads to a case of epistemic injustice understood in terms of epistemic objectification (Section 3.4).

3.2. Defining informativeness

To advance claims regarding the suitability of a merely informative approach, we deem it useful to zoom out from the analysis of specific issues and consider the logic underlying the general relationship between the epistemology and ethics of ML, as it has been treated so far. To achieve this goal, we consider an often-cited overview of the debates revolving around the epistemology and ethics of ML, an article published a few years ago by Mittelstadt and colleagues: "The ethics of algorithms: Mapping the debate" (Mittelstadt et al., 2016).[5] This article has justifiably served as the basis for much good research on the epistemology and ethics of ML, providing a systematic organization of an otherwise fragmented debate. Although, on the one hand, we acknowledge

[5]The approach taken by these authors has been restated and substantiated through more updated literature in a recent publication by Tsamados et al. (2021). In the latter article, the methodology adopted by Mittelstadt and colleagues in analyzing the relationship between epistemology and ethics in ML has been kept unchanged (Tsamados et al., 2021, p. 2). Moreover, Morley et al. (2020) recently provided a mapping review of the ethics of ML in healthcare, also adopting the methodology developed by Mittelstadt et al. (2016). Since we are interested in discussing and building upon the approach considered by these authors in accounting for ethical and epistemological issues, we mostly keep referring to Mittelstadt's contribution throughout this paper.

the value of the contribution provided by these authors, on the other hand, we want to complement this general approach by taking into consideration specific aspects pertinent to the debate that have not been considered by the authors. We scrutinize this review because we see it as particularly clearly illustrating a general approach taken in the literature that is characterized by considering the epistemology of ML as serving an *informative* role in the ethics of ML.[6] To substantiate this claim, we dive deeper into Mittelstadt et al. (2016)'s contribution and analyze its underlying logic.

In their mapping review, Mittelstadt and colleagues provide a conceptual map that allows for the identification of ethical challenges related to the use of decision-making algorithms whose inner logic is cognitively inaccessible to humans. They "are interested in algorithms whose actions are difficult for humans to predict or whose decision-making logic is difficult to explain after the fact." (Mittelstadt et al., 2016, p. 3) To this category belong, among others, clinical decision support systems (CDSS) that recommend diagnoses and treatments to physicians in the field of healthcare (Morley et al., 2020). Following Mittelstadt et al.'s conceptual map, the authors identify six different types of ethical and epistemological concerns raised by algorithmic mediation in decision-making processes. Three are classified as epistemic (i.e., inconclusive evidence, inscrutable evidence, and misguided evidence), two as normative (i.e., unfair outcomes and transformative effects), and a sixth (i.e., traceability) is understood as an overarching concern that, it is argued, can neither be considered entirely epistemic nor entirely normative (Mittelstadt et al., 2016,

[6] As previously pointed out in the first part of this introduction, examples of approaches in the literature that recognize the bearing of epistemological issues on ethical concerns (but not the other way around, i.e., the bearing of ethical properties on epistemic matters) abound. We decide to consider, specifically, the approach advanced by Mittelstadt and colleagues because, from our perspective, it illustrates at best the dichotomy existing between epistemological and ethical aspects of ML in the general debate. This makes it more immediate to effectively show the extent to which approaches that investigate only the bearing of epistemological features on ethical concerns are limited in important ways. This does not imply that this paper's approach is the only one that can be labeled *informative* according to our definition.

pp. 4-5).[7] We will now show that their analysis of the general relationship between epistemology and ethics develops exclusively on the information-serving level.

There are two main dimensions that, as we see it, characterize what we define as the informativeness account, and that can be recognized in the approach underlying the authors' methodology in mapping the debate. That is, epistemological claims about algorithms are (1) *instrumental* to and (2) *autonomous* of ethical considerations.[8] Let us discuss each one in turn.

To see what we mean by *instrumentality*, consider the analysis of the first three kinds of epistemic concerns advanced by Mittelstadt and colleagues, which predicate on the quality of the output (O) produced by ML algorithms. These are inconclusive evidence, inscrutable evidence, and misguided evidence (Mittelstadt et al., 2016, p. 4). The authors' analysis includes showing how these epistemological shortcomings, such as a lack of certainty in O, lead to ethical concerns related to O (Mittelstadt et al., 2016, p. 4). For example, epistemic limitations due to the difficulty of knowing whether connections within datasets are causal (or merely correlational) and the inaccessibility to

[7] Since we want to explicitly address the relationship between ethics and epistemology as considered in the approach taken by Mittelstadt and colleagues, the consideration of traceability as an overarching ethical concern exceeds the purpose of our analysis. In fact, even though questions regarding responsibility attribution of actions in response to ML systems' outputs are of great importance, it is not our aim to discuss this problem in this contribution. Rather, we focus on the two parts of Mittelstadt's conceptual map (Mittelstadt et al., 2016, p. 4) in which both epistemic and normative concerns are explicitly addressed, since there we can most effectively show the informative nature of their general approach. It is, however, true that traceability could also be understood as an epistemic issue leading to an ethical concern. That is to say, the difficulty of accessing the inner workings of ML algorithms (an epistemological issue) constrains the possibilities of responsibility attribution (an ethical concern). Nevertheless, we limit our analysis to the parts of Mittelstadt and colleagues' map that they explicitly recognize as being either ethical or epistemological in nature. We thank an anonymous reviewer for suggesting this possible reading of Mittelstadt et al.'s traceability problem that further supports our interpretation in terms of an informative relation.

[8] Let us note that these two dimensions are not mutually exclusive. In fact, we take that instrumentality applies exclusively to the analysis advanced by the authors in the first part of their conceptual map (Mittelstadt et al., 2016, p. 4) (i.e., the one addressing *epistemic concerns*), while autonomy applies exclusively to the second part of the same map (i.e., the one addressing *normative concerns*). Whereas we see instrumentality as unproblematic, we consider autonomy to be the aspect of their approach that needs to be abandoned to enable a regulatory approach. We will make a case for this claim in Section 3.3.

the connection between the processed data and the conclusion reached by the algorithm lead to concerns about the (lack of) moral justification of actions taken in response to possibly inconclusive outcomes (Mittelstadt et al., 2016, p. 5).

A similar approach is taken in considering how identifying epistemic limitations, understood as the inscrutability of the evidence produced by ML algorithms, leads to ethical problems. The latter are related to, for instance, meaningful consent to data processing and how algorithmic opacity affects the autonomy of data subjects (Mittelstadt et al., 2016, pp. 6-7). The authors also point out that a lack of transparency in how these algorithms operate can lead to a loss of trust from the side of lay data subjects in ML systems and in data controllers (Mittelstadt et al., 2016, p. 7). The same method of analysis also applies to the consideration of what they define as misguided evidence, i.e., the fact that due to technical constraints or flaws in the data that are unintentionally taken up by the algorithm. That is, biased outcomes can be traced back to epistemic limitations that characterize how ML algorithms operate.

From the reconstruction of the first part of Mittelstadt et al.'s conceptual map (Mittelstadt et al., 2016, p. 4), it becomes clear that the analysis and assessment of the epistemology are understood as being prior to claims regarding the ethical acceptability of the outputs of ML systems. In fact, epistemic limitations understood in terms of inconclusive, inscrutable, and misguided evidence not only temporally precede the recognition of ethical issues but are also taken as the very source of these concerns and as instrumental to their identification.

Thus understood, in Mittelstadt et al.'s analysis of epistemic concerns, the ethical assessment of ML is strongly related to and dependent upon its epistemological merits. As previously pointed out, this is a legitimate assumption underlying Mittelstadt et al.'s analysis. Unfortunately, the same dependence cannot be recognized in their assessment of the epistemology of ML, which remains decoupled from ethical considerations in Mittelstadt et al.'s approach.

Relatedly, the second dimension that characterizes Mittelstadt et al.'s analysis is the visible degree of *autonomy* of the epistemological treatment of ML with respect to ethical assessments. The dimension of autonomy comes to light in the second part of their conceptual map, that is, the one related to normative concerns and "based on how algorithms process data to produce evidence and motivate action." (Mittelstadt et al., 2016, p. 4)

In particular, in their assessment of unfair outcomes, the authors leave out the consideration of epistemological factors altogether (implicitly assuming the suitability of the epistemology), stating that the "ethical evaluation of algorithms can also focus solely on the *action* itself." (Mittelstadt et al., 2016, p. 5) Here, the epistemology of ML no longer fulfills an instrumental role; rather, it is completely left unconsidered and disconnected from the ethical analysis. The same applies to their analysis of transformative effects, in which the authors investigate the impact of algorithmic decision-making in terms of how they affect the autonomy of data subjects and the changes they cause to our understanding of privacy and to the concept of personal identity (Mittelstadt et al., 2016, pp. 9-10). For example, Tsamados et al. (2021) point out that the increasing use of profiling algorithms substantially limits the control that data subjects have over their own information. The fact that users are unaware of how their data are processed can contribute to a decreasing level of personal autonomy (Tsamados et al., 2021, p. 9). This analysis is highly relevant in pointing out non-obvious ethical concerns related to how ML algorithms reshape our self-understanding and the way we perceive and interact with the world.

However, zooming out from the relevance of the particular issues addressed, it can be said that in analyzing the general relationship between the ethics and epistemology of ML, ethical considerations are treated as partly disconnected from epistemological issues since the former cannot influence the epistemic features of a given ML. Indeed, at no point in Mittelstadt et al.'s analysis of transformative effects do the authors refer back to the epistemology of ML sys-

tems, nor do they advance claims regarding the role that ethical considerations should play in regulating it to avoid the ethical issues they discuss.

Drawing on the consideration of these two dimensions, the epistemological treatment of ML emerges as either instrumental to its ethical assessment or autonomous from it. Thus, the epistemological assessment of O fulfills the informative role of identifying the scope and merits of different ethical concerns. The contrary—that is, including ethical considerations into epistemological assessments of ML—seems to be missing from the framework of the relevant literature they analyze in their mapping review. We will make such a view a centerpiece in this paper, showing, in the next section, the limitations of the informativeness account and the need for an approach capable of accounting for the conflating nature of the epistemology and ethics of ML. To make these considerations more graspable, in the next section, we zoom into specific issues that derive from the application of the logic underlying the informativeness approach to a concrete case.

Although the informativeness approach is correct in many respects and, as we pointed out in the previous section, is indeed the approach that has been mostly endorsed in the literature, the epistemological and the ethical assessment of ML emerges as partly decoupled. Instead, we aim to show that this way of seeing the relationship between the epistemology and the ethics of ML sidelines two central aspects. First, the fact that ethical features also exert influences on the epistemological counterpart. Second, and perhaps more importantly, this influence is not merely informative but regulatory of the epistemology of ML to the extent that an ethical feature of the situation should lead to, on occasion, the re-evaluation of central epistemic functions such as explanation.[9] With this analysis, we aim to point out some difficulties that emerge in connection with the general approach to the epistemology and ethics of ML. In particular, we intend to show the necessity of seeing the epistemology and ethics of ML as substantially intertwined, thereby reaffirming

[9]We make clear what we mean by a regulatory relation in the next section of the paper.

their mutually regulatory role.

Now that we have provided a brief characterization of the informativeness account by analyzing the logic of the general approach taken by Mittelstadt and colleagues according to the two dimensions identified, in the next section, we take into consideration an example of explanatory medical ML that does not square well into this general approach. This should be functional to show in a more tangible and compelling way the need to expand and build upon the informativeness approach, accounting for the fact that the ethics and epistemology of ML are not to be considered compartmentalized dimensions.

3.3. Beyond informativeness: a case for explanatory medical ML

In what follows, we focus our efforts on showing the shortcomings of the logic underlying the informativeness account as we reconstructed it in the previous section. This allows us to argue for the need to consider the mutually regulatory role of epistemological and ethical features of situations mediated by a medical ML.

When confronted with the output of an ML system, the human inquirer is prompted to form beliefs about the empirical world.[10] These beliefs are intended to be associated with and populate our system of knowledge and understanding of the world, broadly conceived. To see how these beliefs are formed, consider *med+ML*, a cancer-detection system that renders as output the following explanation: "the chances for melanoma for patient X are 89% given the analysis of the following characteristics: the image shows a mole that is 98% asymmetrical; the image shows a mole that is 8mm long ($<$ 6mm

[10] An anonymous reviewer rightly pointed out that the human inquirer could also simply suspend their judgment. Even though we acknowledge this possibility, we consider the more relevant case in which a decision needs to be made following (or not) an ML output. This entails that beliefs need to be formed connecting the ML output with the empirical world to render the former actionable.

considered no melanoma), etc." (Esteva et al., 2019 present such a system). Should the physician believe this output, then *med+ML* has induced a specific belief about the patient's mole, namely that it is carcinogenic. Let us note that this belief is based on specific biological markers about the patient that *med+ML* detects and analyses. These markers, along with any explanation of how they are obtained, populate the physician's body of knowledge about the patient's medical condition, potential treatments, and prognosis. Let us also note that the output of *med+ML* might also induce moral beliefs about the most suitable medical action,[11] the general principles that the physician must follow, and the like. In fact, based on the biological markers measured, the physician forms a moral judgment that will guide her actions: the best treatment for this patient is surgery, chemotherapy, or something else. Naturally, these decisions are not exclusively made by physicians but also depend on the values upheld by the medical department, the hospital, and the national health service.[12]

3.3.1. On the mutual dependence of the epistemology and ethics of ML

To illustrate the limitations of the informativeness account, consider the following situation for med+ML. After analyzing large amounts of data pertaining to a given patient p, along with other relevant medical information, theories, and data, *med+ML* classifies p's image of a mole as melanoma. Suppose now that *med+ML* suggests chemotherapy as the most promising treatment for p's melanoma. Consider further that *med+ML* also offers a *bona fide* explanation for this output (cf. Durán, 2021). That is, the explanation is well-structured,

[11] By "most suitable", we mean a medical course of action that takes into account the biological metrics of a patient as well as their personal, moral, and other values to inform their decision-making.

[12] In this respect, we follow philosophers of science, moral philosophers, and sociologists of science who have long debated about epistemic and non-epistemic values and their crossovers (see, e.g., Douglas, 2009 and Longino, 2004).

answers why-seeking questions—as opposed to merely classifying the output—and delivers epistemic goods, such as understanding the output and coherence with a larger body of medical beliefs. For the sake of the argument, let us say that *med+ML* offers a reliable diagnosis and an accurate treatment.

Thus, the output of *med+ML* plays a critical role in forming the physician's epistemic attitude: the physician believes to possess medically relevant knowledge about p having melanoma, and that the best treatment is chemotherapy. Furthermore, having an explanation of the output also fosters a moral belief in the physician, one in which she is justified in administering chemotherapy to p. We frame it this way because, *ex hypothesi*, the physician is in no epistemic, cognitive, or moral position to confirm, contend, or opt out from believing the output of *med+ML*. As presented, the physician is epistemically justified in believing the output of *med+ML* and morally justified in subjecting p to chemotherapy treatment (Durán & Jongsma, 2021).[13]

In light of this example, the physician is convinced that she holds a piece of knowledge about p and that she is compelled to accept the treatment suggested by the *med+ML* as likely the most suitable for p. In terms of the informativeness account, the physician is then morally justified in preparing and subjecting the patient to chemotherapy, as per the epistemically grounded recommendation of *med+ML*.

Consider two further developments. First, chemotherapy induces anemia as a consequence of blood loss, bone marrow infiltration with disruption of erythropoiesis, and functional iron deficiency as a consequence of inflammation. This is a frequent and unfortunate consequence that many patients must face during chemotherapy (Bryer & Henry, 2018). For a number of reasons, depending on the medical and genetic conditions of p, anemia can be treated with a blood transfusion. Second, p's personal values dictate that receiving

[13] If this scenario sounds unlikely to happen, consider, again, physicians who are compelled to act upon the recommendations produced by algorithmic Prescription Drug Monitoring Platforms (PDMPs) (see Section 3.1). We consider systems such as PDMPs in assuming the epistemic and moral dependence of the physician on the ML system.

a blood transfusion is an unacceptable form of treatment, and it must be unequivocally rejected.

In light of the new information, one could argue that the physician can reject the output of *med+ML* and thus avoid any conflict with p's values. However, without further consideration, we see this move as problematic. First, we cannot assume the physician to be the absolute knowledge-generating entity capable of epistemically overriding *med+ML*. In fact, medical ML cannot be taken as yet another medical instrument for decision-making (such as MRI or blood count analysis) since it effectively displaces physicians from their epistemic role. In our case, this means that p's treatment is, at best, on standby, awaiting the physician's decision on a course of action. In more complex cases, this might not even be possible. Let us also notice that the introduction of moral values in the epistemic assessment of ML might require, as it does in our case, a new treatment recommendation. In such a case, the physician has one of two options: either disregard the ML altogether, effectively neutralizing its use, or "factor" the moral values into the system. Our argument is that the informativeness account fails to consider the latter case.

Second, and more importantly, the suggestion to reject the output of *med +ML* cannot eschew a case of value conflict. It presupposes that refusing to treat p with chemotherapy follows from the principle of non-maleficence. However, this decision also clashes with the principle of doing no harm insofar as, without treatment, p's biomedical well-being will be neglected. For the reasons given above regarding physicians' epistemic displacement by ML, we cannot take for granted that the physician will have a further course of action clearly figured out once it becomes clear that p is against blood transfusion.[14] It can be challenging—if not entirely impossible—to find an alternative treat-

[14] A related point concerns the extent to which physicians would consider personal values as relevant for diagnosis and treatment and thus as morally problematic. According to diverse approaches in medical ethics, the physical well-being of a person supersedes personal values (Richman, 2004). Although we cannot elaborate on these considerations, they seem relevant to the issue at hand.

ment compatible with p's values, and using an ML system as an epistemically powerful entity could be of great support. For this purpose, we need an epistemological framework that allows relevant ethical information (in the case considered, pertaining to p's values) to have a direct bearing on crucial epistemic functions (such as explanation).

Now, by construction, the informativeness account is not sensitive to how a new piece of ethically relevant information (e.g., p's personal values) should be included in the evaluation of central epistemic features of the systems (i.e., crucial epistemic functions, such as explanations). Concretely, this means that in view of the informativeness account, the physician remains epistemically and morally justified in subjecting p to chemotherapy. This is the case because, within the theoretical framework of the informativeness account, how epistemic functions should be adapted in order to include relevant ethical considerations remains unconsidered. To admit the possibility that ethical properties have a regulatory influence on epistemological functions of ML means accepting that the epistemology and ethics of ML must constantly be re-evaluated. However, as pointed out in Section 3.2, the informativeness account remains silent on the possibility that ethical considerations—in our case p's values—can have a bearing on the epistemological assessment of *med+ML*.[15] For these reasons, the informativeness account does not fit situations similar to the one under scrutiny.

To render the output of *med+ML* based on p's values actionable for the physician, the epistemic assessment of *med+ML* must be reconsidered, including relevant ethical information pertaining to p's situation. To our mind, the fact that an ethical property of the situation (i.e., p being against blood trans-

[15]The problem presented here is different from assuming that the epistemology of ML is empty of values (moral, cultural, economic, political, etc.). Mittelstadt et al. would admit that an explanation rendered by *med+ML* depends on the kind of question we want to answer, the information provided, etc. In summary, the epistemology of ML is not value-free. The crucial difference is that Mittelstadt et al. consider that once the epistemology of ML is settled, it is informative. Our contention is that there is a "loop back" from the ethics to the epistemology, a loopback that is unaccounted for by the informativeness account.

fusion) should lead to the reassessment of an epistemic function highlights how epistemological and ethical considerations of ML are closely intertwined and mutually regulatory instead of compartmentalized, as the informativeness account takes them to be.

Against this background, how can we evaluate the recommendation produced by *med+ML* for further medical action? On the one hand, *med+ML*'s recommendation is based on standard medical and biological theories, evidence, and the general body of knowledge on diverse types of cancer. In this respect, we can say that the physician is *epistemically justified* (Durán & Formanek, 2018) in believing that the recommendation is pertinent since the epistemic state induced by *med+ML* supports such a recommendation (i.e., the output is based on an accurate analysis of the biological state of p). On the other hand, the physician is not *morally justified* in following through with the chemotherapy treatment because this conflicts with p's values. The solution to this problem is to factor p's values into the system to render a new treatment (e.g., the new best treatment given p's values is surgery).

Let us also clarify the nature of the epistemic and moral normativity in place in the case under scrutiny. While, as already pointed out, the physician is justified to believe that chemotherapy is the best treatment biologically speaking, she is not justified to believe that it is the *overall* best treatment for p. This is because the moral claim entailed in p's rejection of blood transfusion has a direct bearing on what the physician should believe is the most suitable course of action all things considered. Under a definition of health that exceeds the evaluation of biological parameters and also includes moral, social, and otherwise relevant considerations, the physician is epistemically justified to believe that chemotherapy is no longer the best treatment for p.[16]

If the above considerations are correct, then a more overarching view of the

[16]This interpretation of health as more encompassing than biological health is aligned with the WHO definition of this concept: https://www.who.int/about/governance/constitution For further debates on different conceptions of health, see also Richman (2004). Unfortunately, we cannot expand on these issues in this paper.

epistemology and ethics of ML emerges. Whereas Mittelstadt and colleagues rightly emphasize the informative value of the epistemology of ML on moral actions, we complement the missing parts of the framework by showing the merits of an epistemology regulated by the ethics of ML. We believe that cases similar to the one considered here are better analyzed through the lenses of a different approach, one that, as we argue, takes epistemic and moral features of medical ML as substantially regulatory—rather than informational—of each other. In a regulation-based framework, we submit that p's personal values become a substantial part of the epistemology, regulating the epistemological assessment of the system. The regulatory role of ethical features in the epistemology of ML comes to light in its considerable impact on the physician's beliefs. Even if, all things being equal, she would be justified in believing and acting upon the explanation provided by *med+ML*, this is no longer the case as soon as a relevant ethical property of the situation comes to the foreground. Only within a regulation-based framework do induced epistemic attitudes of the physician elicit a clear stand in either being (or not being) morally justified in proceeding with a given course of action.

Drawing on the previous discussion and on the example under scrutiny, we can now consider how, following the logic underlying the informativeness account, situations in which p is the victim of an epistemic injustice do not find treatment. We turn to this analysis in the next section.

3.4. Epistemic injustice

We argued in Section 3.3 that the informativeness approach does not account for information regarding the patient's (p's) values that become relevant after the ML has outputted a treatment recommendation. How does this affect the practice of healthcare with ML, above and beyond the fact that the system's recommendations are unsatisfactory in cases such as ours? We submit that

there is a wrong done to p, understood in terms of Miranda Fricker's account of *epistemic objectification*, which falls within her analysis of *epistemic injustice* (Fricker, 2007).

In its broadest sense, epistemic injustice designates flawed practices in meaning-making and knowledge-creating processes, leading to marginalization, unfair distrust, silencing, and exclusion (among others) (Pohlhaus, 2017). As such, epistemic injustice is a wrong done to epistemic subjects in their capacities as knowers, that is, as recipients and conveyors of knowledge. Issues of epistemic injustice have mainly been addressed in terms of a credibility deficit attributed to individuals belonging to vulnerable societal groups,[17] precisely due to their perceived social identity from the side of their interlocutor(s) (*testimonial injustice*) or to an inability to comprehend and make sense of their own social experience due to a lack of or access to shared hermeneutical resources (*hermeneutical injustice*) (Fricker, 2007).

This multi-faceted phenomenon has been receiving increasing attention in the philosophical debate at the intersection between social epistemology and ethics in recent years (e.g., Byskov, 2021; Carel and Kidd, 2017; Chung, 2021; Moes et al., 2020; Thomas et al., 2020; Wardrope, 2015). Since ML systems are epistemically authoritative and increasingly involved in decision-making procedures that strongly impact patients' lives, it is of paramount importance to ensure that they do not undermine epistemic subjects in their capacities as knowers. As such, epistemic injustice in ML-mediated contexts requires particular attention (Pozzi, 2023a, 2023b; Symons & Alvarado, 2022). Issues of epistemic injustice emerge, generally, if patients are excluded from influencing decision-making processes and if their lived experiences, testimony, and personal values (epistemic, moral, societal, etc.) are not acknowledged as legitimate sources of knowledge, among many other factors that still need to be explicitly addressed and investigated in depth.[18] The analysis of our example

[17] Vulnerable epistemic subjects can be considered such due to their gender and race but also because they find themselves in precarious health conditions.

[18] The analysis of all the aspects mentioned above would go well beyond the scope of this

in light of the phenomenon of epistemic injustice should point to the importance of working toward the development and deployment of ML systems that do not represent an obstacle to the active participation of relevant stakeholders in shared decision-making. Operationalizing systems that do not impair the process of understanding and forming beliefs regarding our lived experiences is, in fact, essential to avoid genuinely epistemic forms of injustice that can otherwise emerge.

The following analysis allows us to shift the focus of the debate from a conception of epistemic injustice, which has been mostly considered in a human-centered fashion, to its application to cases in which epistemic subjects interact with ML systems. It is our aim to show that Fricker's concept of epistemic objectification can be successfully applied to our ML case to capture the moral wrong suffered by p.

Let us now turn to the reconstruction of Fricker's account of epistemic objectification so that we can, in a second step, show that it can capture at best the moral wrong inflicted on p in the case addressed in Section 3.3.

3.4.1. Epistemic objectification

According to Fricker, a subject is epistemically objectified in situations in which she is, due to prejudices from the side of the hearer(s), completely deprived of her active role as an informant and is, as such, reduced to a *mere* source of information.[19] Drawing on Craig's account of the State of Nature (Craig, 1990), Fricker argues that the distinction between "informant" and "source of information" in the process of conveying knowledge is an epistemological aspect that entails relevant ethical meanings. Informants are to be

paper. In this contribution, our investigation is limited to pointing out how the example under scrutiny can be interpreted through the lens of Fricker's conception of epistemic objectification.

[19] The analysis excludes cases in which the hearer judges their interlocutor as epistemically untrustworthy and, for this reason, and not due to some forms of prejudices, she does not acknowledge them in their role of informants (Fricker, 2007, p. 136).

considered epistemic agents who are able to convey information actively and share knowledge with their interlocutors (e.g., by communicating information). In the field of healthcare, a patient can be considered an active informant in that she can communicate relevant information regarding her physical and mental state to her physician, thus participating and playing a role in informing medical decisions.

Differently, sources of information are states of affairs from which an inquirer can deduce information. Therefore, as Fricker points out, individuals can be both informants, being able to actively express themselves and convey knowledge, and sources of information, in that the inquirer can derive information about their current state, for instance, through observational evidence of their behavior (Fricker, 2007, p. 132). For a human being to be a source of information could be no reason for concern from an ethical point of view; this is the case, for example, in a situation in which a physician concludes that a patient suffers from a particular pathology due to the analysis of medical tests conducted on the patient in question. That is, the physician comes to a conclusion regarding the current state of the patient without the patient actively communicating it.[20] Against this background, it is undisputed that the dimensions of being an active informant and a source of information coexist in human beings as epistemic subjects.

By contrast, treating someone as a *mere* source of information implies an instrumentalization of the subject, depriving them completely of their role as active informants. One does not need to adopt Kantian principles of morality to acknowledge that the instrumentalization of subjects is universally wrong from a moral point of view. In the context of medicine and healthcare, being treated as a *mere* source of information would mean that the patient is expected to provide basic information regarding her current state but is deprived—due to, for example, prejudices that physicians or other healthcare

[20] A more straightforward example from everyday life could be a case in which someone blushes, and from this behavioral feature, we derive that he or she is embarrassed.

professionals have related to their social identity—from the possibility of participating and contributing in a substantial way to the collective epistemic activity of sharing their lived experience. However, this is arguably key to making sense of their health situation and actively participating in shared medical decision-making processes (Carel & Kidd, 2017).

Fricker takes the phenomenon of epistemic objectification as reconstructed as a particularly harmful form of silencing and the central wrong derived from epistemic injustice (Fricker, 2007, p. 6). Indeed, the fact that a subject's active contributions are limited or impaired altogether represents a considerable restriction to their agential role as rational individuals and strongly constrains their participation in the production and exchange of knowledge. Thus, this can be considered the primary wrong that epistemic injustice understood in terms of epistemic objectification perpetrates on its victims since, in these cases, the knower is deprived of her active agential role and, as such, "wronged in a capacity essential to human value." (Fricker, 2007, p. 44) A secondary kind of wrong can manifest in more practical—but not less detrimental—terms, also creating a clear disadvantage for the subjects involved. In the context of healthcare, for example, the risk of attributing to patients a deflated level of credibility on the basis of prejudices connected to their status as ill persons could lead to being misdiagnosed.[21]

A growing body of literature addresses the fact that ill persons can be considered a particularly vulnerable category inclined to suffer epistemic injustice in Fricker's sense (Carel & Kidd, 2014). Kidd and Carel (2017) argue that judgments about the epistemic credibility of ill persons are often prejudicial, being produced and sustained by both negative stereotypes and the structural characteristics of healthcare practices (Kidd & Carel, 2017, p. 175).

[21] As a matter of fact, in a case in which "the style of interaction between clinician and patient is one that closes down communication, such that important information is potentially lost" (Carel & Kidd, 2014, p. 531), it is not too far-fetched to think of the possibility of misdiagnosing a patient as a legitimate practical concern deriving from an instance of testimonial injustice.

In particular, they point out that ill persons are vulnerable to epistemic injustice through the supposed attribution of characteristics such as cognitive unreliability and emotional instability that deflate their testimony's credibility. However, there are also structural features of healthcare systems that can be regarded as the causes of hermeneutical forms of injustice (rather than the intentions of individuals). For example, considerable time limitations and the use of standardized protocols contribute to the marginal role assumed by personal needs and values (Kidd & Carel, 2017, p. 176). Further, the difficulty of articulating particular aspects related to a patient's illness is an aspect that can be considered challenging from a hermeneutical point of view (Kidd & Carel, 2017). Overall, being in a physically and mentally precarious condition puts the patient in a situation of vulnerability and dependence, which undermines her own epistemic confidence (Kidd & Carel, 2017, p. 176).

Drawing on what has been said so far, we can conclude that in standard, that is, non-ML-mediated practices in healthcare, there are factors such as the ones previously mentioned that put the patient p into a position of epistemic vulnerability. We submit that the situation becomes even more pressing when an additional epistemically authoritative entity, such as *med+ML*, becomes involved in this relationship.

3.4.2. Informativeness and epistemic objectification in machine learning

We now consider how *med+ML*, without allowing for the possibility of integrating p's values into its epistemology, brings about a case of epistemic objectification at p's expense in the example under scrutiny. This analysis aims to further show the need to implement ML systems that allow ethical features (say, a patient's values) to regulate epistemologically relevant aspects (e.g., an explanation provided by the system). However, before turning to this

analysis, some considerations are in order. Whereas Fricker sees epistemic objectification as the most direct expression of instances of testimonial injustice, we need to detach ourselves from her human-centric approach to make our case for epistemic objectification brought about by *med+ML* at p's expense. As previously mentioned, the wrong that she aims to capture is caused by unjustly deflated credibility judgments that a subject receives from her interlocutor due to prejudices related to her social identity. Since, in our case, the physician does not play an active role in mediating between *med+ML* and p, prejudicial judgments that could be detrimental to p's epistemic positions are out of place.[22] Even less plausible would be the assumption that *med+ML* holds prejudices that deflate p's credibility. In fact, it goes without arguing that attributing these genuinely human traits to an ML system would be a category mistake. Thus, the epistemic objectification we aim to capture is one that emerges because *med+ML* cannot pick up on p's values, and therefore, p's agential contribution to the decision-making process cannot be successfully considered. This is, we claim, due to how *med+ML* operates, as elaborated in detail below.

To convincingly argue that p suffers a form of epistemic objectification brought about by *med+ML*, we need to account for the fact that p's knowledge (e.g., in the form of her personal epistemic values) is wholly excluded from the decision-making process leading to the output. Relatedly, having shown this will allow us to argue that p is treated as a mere source of information and not as an active informant. It follows that p is epistemically objectified. We show that p is utterly excluded from the decision-making process by making explicit a vicious circularity in how *med+ML* produces its output, which is unsolvable following the informativeness account. We take the epistemic objectification of p as a direct consequence of the vicious circularity to which our discussion now turns.

[22] Note that, as previously mentioned, we exclude the possibility of the physician intervening independently from the *med+ML*.

As previously argued, by construction, the informativeness account puts forward an investigation of how the epistemology of ML informs the ethics of ML. We showed that ethical elements substantially affecting the epistemological counterpart are left unaddressed. Thus, the informativeness account adopts a unidirectionality that goes from the epistemology to the ethics of ML but not the other way around. The example in Section 3.3 illustrates this unidirectionally.

As indicated, the informativeness account considers, *ex hypothesi*, the epistemological assessment of *med+ML* to be "fixed" and therefore unmodifiable by new incoming information that may be relevant to p's medical condition. From this perspective, *med+ML* induces in the physician the belief that the explanation is suitable, along with the moral justification for acting upon it. This informs, in turn, the physician's actions. It follows that at the moment in which p is confronted with the output brought about by *med+ML*'s suggestion of chemotherapy and, consequently, blood transfusion as the most suitable treatment, p will have to refuse the suggested therapy, restating that it goes against her personal values. At this point, the vicious circularity becomes obvious: since *med+ML* is unable to factor this relevant piece of information into central epistemological functions (such as in the explanation of the output), p can only be confronted anew with the same outcome produced by the medical ML system, an unsuitable treatment recommendation for her set of values. Our claim is that the reiteration of this hypothetical yet logically consistent scenario exerts distinctive negative influences on p's epistemic confidence and, more importantly, strongly limits her agential role. Indeed, from the moment in which the output of *med+ML* is created, p is not in a position to actively influence the decision-making process and is, consequently, completely left out of it. This leads us to the second claim, that is, *med+ML* leads p to be treated as a *mere* source of information.

As previously pointed out, treating p as a source of information is generally unproblematic and thus also in cases in which a medical interaction is

mediated by an ML such as *med+ML*. This is the case since *med+ML* can elaborate information regarding p's physical state that she might not have directly provided but that has been acquired through different processes. Indeed, the system can effectively elaborate information regarding p's physical condition from indirect sources, such as laboratory tests or any kind of medical examination she has undergone. The epistemically and ethically relevant problem in terms of epistemic objectification arises as soon as p is in possession of a relevant piece of information (i.e., the fact that p is against blood transfusion) that cannot, however, be accounted for by *med+ML*. That is, at the point where p should actively convey new relevant information, she is prevented from doing so due to the role played by *med+ML*. As a consequence, p cannot receive an appropriate medical treatment compatible with her personal values.

Drawing on the discussion so far, it can be argued that p is treated as a mere source of information since her agential contribution is left unconsidered by the explanation provided by *med+ML*. This constitutes, as such, a case of epistemic objectification. Very crucially, the unidirectionality previously pointed out leads to a unidirectional exchange of knowledge: the end users of *med+ML* are merely recipients of knowledge but are not able to actively influence the knowledge-producing process itself. This outcome leads to undesirable consequences for p: either an endless circle in which no suitable alternative to the output produced by *med+ML* is found or a medical procedure that infringes on her personal—moral and epistemic—values.

Bottom line, the informativeness account does not provide the theoretical backdrop needed to tackle moral and epistemic issues in terms of patient objectification as we have been discussing them. These considerations reinforce the need for a flexible epistemology capable of incorporating new relevant information as it is acquired to overcome moral and epistemic concerns in connection with the epistemic objectification of the relevant stakeholders involved in medical decision-making processes.

3.5. Final remarks

This contribution aims to point out the limitations of an approach in the epistemology and ethics of ML that sees these two dimensions as compartmentalized. In particular, we analyzed considerations of the general relationship between the epistemology and ethics underlying the approach taken by Mittelstadt et al. (2016). We reconstructed their methodology in terms of an information-serving relationship between the epistemology and ethics of ML according to two dimensions that, to our mind, characterize their analysis (i.e., instrumentality and autonomy). In Section 3.3, we substantiated our claims by considering a case of explanatory medical ML that cannot be appropriately solved following the logic of the informativeness account. We analyzed the ethical consequences of this situation for the patient involved in the example considered in terms of Fricker's concept of epistemic objectification (see Section 3.4).

Our main criticism toward the informativeness account is that it is not designed to address cases that require the analysis of how an ethical property (such as patients' values as discussed in the case in Section 3.3) should lead to the re-evaluation of central epistemic functions in situations mediated by a medical ML. The informativeness account remains silent on the possibility that p's values motivate the rejection of an otherwise well-constructed explanation (thus requiring a new one). Hence, morally problematic situations in which a physician is no longer justified to act upon the ML output produced do not find treatment within this theoretical framework.

Whereas, in line with the relevant literature, Mittelstadt and colleagues rightly emphasize the informative value of the epistemology of ML on moral actions, our aim was to make clear the need to complement their framework by showing the merits of an epistemology regulated by the ethics for ML. The regulatory role that ethical features have on epistemic functions has been made

explicit in the example analyzed in Section 3.3: in that case, the physician is, in principle, not justified in acting upon the explanation provided by the ML system in the face of p's values as a relevant ethical feature of the situation. In turn, this means that an otherwise sound explanation needs to be reformulated, including the consideration of p' values. It is in this sense that we take that an ethical property *regulates* what counts as a morally acceptable explanation and what does not.

Admittedly, there is a need for more consideration of how a flexible epistemology can be formalized. However, with our work, we hope to have contributed to the debate, showing the importance of further pursuing this direction in future research to avoid epistemic and ethical issues such as those highlighted in this contribution.

II

Part 2: Forms of Epistemic Injustice in Machine Learning

4

Testimonial Injustice in Medical Machine Learning[1]

4.1. Introduction

Machine learning (ML) systems are increasingly being introduced in high-stakes fields such as medicine and healthcare. On the one hand, it has been shown that these systems hold great potential in ameliorating medical delivery, reaching high levels of accuracy and precision (Topol, 2019). On the other hand, it is also widely acknowledged that they are the source of salient ethical questions regarding, for example, patients' autonomy, responsibility allocation and trust (Durán & Jongsma, 2021; Grote & Berens, 2020; Morley et al., 2020). This paper focuses on a less discussed but not less relevant epistemo-ethical issue: the role of ML systems in medicine in causing *testimonial injustice*, that is, a form of epistemic injustice (Fricker, 2007; Symons & Alvarado, 2022).

[1]This chapter is based on the following article:

Pozzi, G. (2023). Testimonial injustice in medical machine learning. *Journal of Medical Ethics*, 49:536-540. https://doi.org/10.1136/jme-2022-108630

To make my case for testimonial injustice arising in connection with the use of ML in medical contexts, I consider ML systems deployed in the USA to predict patients' likelihood of opioid addiction or misuse, that is, automated Prediction Drug Monitoring Programs (PDMPs) (Oliva, 2022; Szalavitz, 2021). I show that these systems' role in medical decision-making could deflate patients' credibility on epistemically invalid grounds, reducing the overall epistemic relevance of their testimonies and thus harming them in morally significant ways. Thus, I aim to show that patients are wronged as epistemic subjects, crucially due to how these systems mediate patient-physician interactions.

This paper is structured as follows. In the next Section 4.2, I point out the problematic nature of ML-based PDMP risk scores and briefly introduce the problem of testimonial injustice in medicine. In Section 4.3, I address the question of how PDMPs deflate patients' credibility, and I argue that the main reason can be traced to the fact that PDMPs are treated as markers of trustworthiness. That is to say, I show that the risk scores generated by these systems are treated as crucial indicators on which assessments regarding patient credibility are formed. In Section 4.4, I focus on the epistemic and ethical concerns that arise from treating these systems as markers of trustworthiness, and I argue that this practice is both morally and epistemically unjustified.

4.2. Testimonial injustice in medicine

"I don't think you are aware of how high some risk scores are in your chart." (Szalavitz, 2021) With these words, a woman named Kathryn was discharged from a hospital in July 2020 while still in a precarious health condition. It turns out that an algorithmic system (NarxCare algorithms, see Szalavitz (2021)) that is supposed to deliver an accurate estimation of her likelihood of opioid misuse was decisive for this to happen. In fact, these "law enforcement-developed digital surveillance systems" (Oliva, 2022, p. 51) play an increasingly

relevant role in physicians' decision-making.[2]

In the aforementioned case, the risk score assigned to Kathryn led her physician to discharge her and the pharmacies to deny service to her. What made her situation even worse was that she could not overturn that unfavorable outcome with her *knowledge* of her own physical and mental condition. She knew that she was not addicted to opioids and had never misused drugs. However, the authority taken up by the automated system generating her risk score overrode the legitimacy of her testimony and constrained her possibility of contradicting an inaccurate assessment of her drug consumption.

Even if PDMP scores are ideally supposed to be used by physicians as a starting point to engage in a conversation with their patients, PDMPs in their current deployment *de facto* hinder rather than facilitate fruitful exchanges. Their black-box nature and the fact that legal actions can be taken on physicians labelled as overprescribers are factors that lead healthcare professionals to over-rely on these systems (Oliva, 2022).[3] The bottom line is that Kathryn

[2]The extent to which automated PDMPs influence medical decision-making varies from state to state, according to different provisions (e.g., 13 states mandate healthcare workers to consult PDMP records, and in other states, physicians check PDMPs only in cases that are deemed suspicious (Bulloch, 2018; Leichtling et al., 2017).) Even if these systems might not be the overall predominant method of practice (yet), empirical studies have shown that concern that these systems are leading to a worrisome shift in medical practice is legitimate (Haines, Lam, et al., 2022). As Oliva points out, "due to state PDMP use mandates and law enforcement surveillance, clinicians (...) increasingly rely on PDMP risk scores to diagnose and treat patients. And there is little doubt that such clinical reliance will become even more pervasive." (Oliva, 2022, p. 109) I thank an anonymous reviewer for encouraging me to clarify the extent to which automated PDMPs are currently used in medical practice.

[3]As I elaborate in more detail in Section 4.4, one can argue that the objectivity and neutrality wrongfully attributed to these systems bear the potential to deflate the value of patients' testimonies in medical decisions. Haines, Savic, et al. (2022) pointed out that healthcare providers are "more likely to accept the default settings of automated systems at the expense of other relevant emotional and psychological patient information." They continue stating that this "can result in errors of commission, where the value of information attributed to the automated tool overrides the value of clinical expertise, even where the automated information contradicts clinical training and evidence." (Haines, Savic, et al., 2022, p. 2) This corroborates the claim that even if automated PDMPs are intended as decision support systems, they considerably influence physicians' judgments in crucial decision-making practices so that physicians often end up following these systems' recommendations (particularly in combination with other features of medical practice, such as time limitations). Under this heading, the value of patients' testimony in shared decision-making is likely to be obfuscated by the scores attributed to them. I thank an anonymous reviewer for encouraging me to clarify this important point.

was excluded from a decision-making process in which, strikingly, she was intended as the sole beneficiary.

This briefly depicted scenario clearly illustrates that the injustice suffered by Kathryn is palpable. However, going beyond the deep sense of injustice that our moral intuitions can capture, how can the nature of the wrong she experienced be conceptualized? It is noticeable that Kathryn was not only denied access to her fair share of medical assistance but was also undermined in her role as a *knower*. This comes to light considering that she was wrongfully disadvantaged in her possibility to communicate relevant information about her health condition and having this information recognized and acted on by medical professionals. I argue that the framework of epistemic injustice, particularly Fricker's conceptualization of testimonial injustice, can capture the moral and epistemic wrong that Kathryn suffers from. In fact, the injustice she experienced cannot be understood exclusively in distributive terms.

As Fricker (2007) points out and Symons and Alvarado (2022) extensively elaborate, epistemic injustice is not necessarily connected with the unfair distribution of goods (e.g., access to information or, in the context of medicine and healthcare, medical professionals' advice, and medical support and care). Epistemic injustice is to be considered a discriminatory injustice in which a person's epistemic status is unjustifiably diminished for epistemically invalid reasons (on which I elaborate below). Thus, the framework of epistemic injustice points out a more subtle form of harm that can easily go unnoticed. More precisely, it sheds light on the mechanisms underlying epistemically illegitimate reductions in a person's credibility. Even if these instances can often be connected to other inequalities, they deserve to be considered in their own right (Symons & Alvarado, 2022). In this contribution, I consider the role played by ML-based PDMPs in causing this genuinely epistemic form of moral wrong. Before turning to this analysis, I reconstruct in more detail the main features of testimonial injustice.

Fricker's analysis of testimonial injustice relies on the observation of dis-

criminatory practices that question the epistemic status of individuals belonging to disadvantaged categories based on unfounded prejudices. That is to say, an individual suffers testimonial injustice if she receives, as a consequence of prejudices that her interlocutor holds related to her social identity, less credibility than she would have received in the case that prejudicial judgements were not in place (Fricker, 2007).

A sadly common case of testimonial injustice is when a woman is attributed less credibility because of her gender. For example, in pain medicine, it has been shown that women's pain is often underestimated, a phenomenon not encountered at the same frequency by men (Lloyd et al., 2020). This occurrence of unfair treatment is due to an inappropriate withdrawal of credibility, often rooted in stereotypes that deflate women's credibility levels. For instance, women are often perceived as more emotional and apt to complain than men and are granted, for these reasons, less credibility overall (Carel & Kidd, 2014, 2017; Kidd & Carel, 2017). Also, racial biases can be the root of illegitimate credibility deficits, leading to unjust pain management. A study by Trawalter et al. (2012) shows that black patients are often undertreated due to misguided beliefs regarding their ability to endure pain. The consequences of an unjustified lack of credibility can have a detrimental impact on patients' general well-being, above and beyond the fact that they are wronged in their capacity as knowing subjects: the information they seek to convey is not taken seriously and does not get to inform the decision-making processes they are directly affected by. As Fricker points out, certain stereotypes and prejudices related to one's social identity are so entrenched in our social structures that they are not easily detected, let alone amended (Fricker, 2007).

The PDMP case previously described indicates that ML systems implemented in medicine and healthcare can create further imbalances in physicians' assessments of their patients' credibility. In the next section, I analyze how this happens and for what reasons, identifying the main features of testimonial injustice when it is induced by ML-based PDMPs.

4.3. Credibility in ML-mediated medical decision-making

A reciprocal relationship of trust between patients and their physicians is grounded in respect for epistemic duties and rights, among other aspects. From a patient's perspective, the latter amounts to the right to receive relevant information regarding one's health condition (e.g., the results from tests the patient underwent in understandable terms and free from complex technicalities), to convey knowledge about one's mental and physical state, and to have information shared with physicians taken into account, among others. Respectively, patients' epistemic rights translate into epistemic duties for physicians. The latter are categorized by Watson as basic epistemic duties and comprehend the duties to "seek, receive, and impart information", (L. Watson, 2021, p. 36) along with their negative counterparts (e.g., avoiding seeking irrelevant information about a patient that exceeds the purpose of a diagnostic procedure). A successful patient-physician relationship holds as long as patients can trust physicians to respect their epistemic rights and physicians can trust their patients with the duty to be sincere when, for instance, patients report their symptoms.

Most situations in which we trust someone are situations of vulnerability for the trustor. In such situations, we defer to the trustee (i.e., the person we (need to) trust), confident that they will fulfill the expectations that are implicitly or explicitly intrinsic to the trust relationship itself (Frost-Arnold, 2020).[4] While patients need to trust medical professionals' expertise and beneficence, physicians also have to trust that the information patients provide about themselves (e.g., their symptoms, medicament used) is not deceptive (Rogers, 2002).

[4] In a patient-physician relationship, a physician's epistemic duties are rendered explicit by institutionalized practices (e.g., the Hippocratic Oath), codes of conduct, and the four fundamental biomedical principles (i.e., beneficence, non-maleficence, justice, and non-discrimination).

Therefore, it can be said that trust is closely related to credibility. In assessing a patient's testimony, the unjustified withdrawal of credibility can be disadvantageous, potentially leading to injustice.[5]

Credibility assessments are particularly prone to be distorted by biases and stereotypes connected to a person's social identity because to form these, we usually rely on so-called *markers of trustworthiness* (Fricker, 2007). In fact, when deciding whether to rely on someone's testimony, we need to find a way to assess their epistemic trustworthiness. After all, we want to accept testimonial exchanges that are, most probably, truth-conducive so that relying on a person's testimony will lead a subject to form true beliefs about the world (Lehrer, 2006). However, identifying markers of trustworthiness that can successfully fulfill this epistemic task also has a considerable moral dimension. Testimonial injustice is often in place if what leads to the acceptance of certain markers of trustworthiness are prejudicial assessments connected to one's interlocutor's social identity (i.e., age, gender, ethnicity, social status).[6]

In medical encounters, markers of trustworthiness play a more or less important role depending on the situation. Typically, two different scenarios can be distinguished. On the one hand, on occasions in which a patient's reported symptoms are objectively connected to a visible and easily quantifiable cause—say, a patient reports pain and an X-ray shows a broken bone—credibility attribution happens in a quite straightforward manner, and the need to recognize markers of trustworthiness to assess the credibility of a patient's testimony moves into the background. On the other hand, particularly epistemically and

[5]Let me clarify that avoiding epistemic injustice does not require physicians to take patients' testimonies at face value. If a physician has valid reasons to deem a patient epistemically untrustworthy, she is, of course, entitled to disregard her testimony (without infringing her epistemic rights). Crucially, testimonial injustice occurs if the reasons a patient's testimony is disregarded are epistemically invalid, such as in a case of unfounded prejudice related to a patient's social identity. I thank an anonymous reviewer for encouraging me to clarify this relevant point.

[6]For example, Fricker refers to the characteristic of being a gentleman in seventeen-century England as a marker of trustworthiness. Conversely, the lack of this characteristic in women, for example, was taken to be the opposite, that is a marker of untrustworthiness. This is an example of a non-epistemically grounded marker of (un)trustworthiness (Fricker, 2007, p. 119).

morally salient are cases in which patients' reports of their symptoms are the main or only way in which physicians can have epistemic access to their medical condition, while lacking quantifiable and objectively recognizable physiological manifestations that could explain a patient's complaints. This is often the case in chronic pain patients, patients who suffer from psychosomatic diseases and more generally, in the clinical assessment of pain (Buchman et al., 2017). In the latter cases, individuating suitable markers of trustworthiness is crucial to formulating appropriate credibility judgements that inform medical decisions, for instance, regarding whether to prescribe opioid medication.

Patients in need of opioid medications often belong to the categories mentioned. Moreover, the possible stigma of drug addiction or misuse adds a further layer of complexity and inclination toward prejudicial judgment, which can easily further deflate the credibility of a patient's testimony. Given these difficulties, the flair of objectivity and neutrality often (mistakenly) attributed to ML systems could seem like a viable solution. This is precisely the idea behind the implementation of systems such as NarxCare algorithms to manage the opioid epidemic pervading the USA. However, as widely agreed on in the AI ethics literature, ML systems are never value-neutral (see, e.g., Mittelstadt et al., 2016), and their perceived objectivity can be highly misleading. I argue that automated PDMP systems exacerbate and reinforce—rather than mitigate—the occurrence of testimonial injustice in the particular case of interest. The reason is that they are crucially considered markers of trustworthiness on which credibility assessments are often formed. If patients are not granted credibility and their testimonies are discredited *because* PDMPs are treated as markers of trustworthiness, then a form of testimonial injustice is *induced* by ML systems.

To substantiate the claim that PDMPs are treated as markers of trustworthiness, briefly consider how they operate. As I have extensively discussed elsewhere (Pozzi, 2023a), these ML systems are not only epistemically opaque in the technical sense of the term: the unwillingness of the company owning

NarxCare algorithms to reveal information regarding the weight and nature of the proxies that inform patients' risk scores makes a critical assessment of the results produced pretty much impossible (Oliva, 2022).

Despite this fact, automated PDMPs considerably influence medical decision-making. An empirical study by Leichtling et al. shows that "(i)n response to worrisome PDMP profiles with new patients, participants [i.e., clinicians using PDMPs] reported declining to prescribe, except in the case of acute, verifiable, conditions." (Leichtling et al., 2017, p. 1063) This means that information produced by these systems about patients can easily override the epistemic value of their testimonies. Against this background, PDMP scores could, on occasion, be treated as if they were able to say everything that needed to be said about a patient's drug consumption level and eligibility for opioid prescriptions.

Hildebran et al. (2014) and Hildebran et al. (2016) show how communication practices between patients and physicians tend to cut off testimonial exchange, creating an atmosphere of distrust that leads to medical decisions that can hardly be challenged by a patient needing medical attention (Pozzi, 2023a). This further supports the claim that credibility assessments often happen by relying heavily on the scores provided by ML-based systems.

In the next section, I analyze why basing credibility assessments on PDMP scores is unjustified from both epistemic and moral viewpoints.

4.4. PDMP RISK SCORES AS MARKERS OF TRUSTWORTHINESS: EPISTEMIC AND MORAL CONCERNS

What has been said thus far points to the moral harm that misguided PDMP scores can cause. Treating these systems' scores as markers of trustworthiness is epistemically and ethically unjustified for multiple reasons. First, it points to an overestimation of what ML systems can achieve, indicating automation bias (Logg et al., 2019).

However, it is essential to recognize these systems' limitations. ML systems are statistical systems that make predictions based on correlations established at the population level. The latter are generated by connecting people who share certain salient attributes and are, as such, categorized as being part of the same reference class (e.g., turning to a certain number of physicians for medical care, having experienced certain traumatizing events, criminal history (Oliva, 2022)) with the target class of interest (in this particular case, people who are likely to develop opioid addiction or misuse) (Greene, 2019). Nevertheless, it is not an epistemically legitimate step to switch from the population level to the individual level without further consideration and without taking into account the particular situations and values of patients in their singularity.

This means that these predictions can be highly misleading, connecting attributes pertinent to a certain category of patients but not necessarily connected to a person's drug consumption or possible tendency to misuse opioids (Pozzi, 2023a). Moreover, using, for example, criminal history as a proxy that informs patients' final risk score disadvantages racial minorities, who tend to be underinsured and overpoliced in the US compared with white people (Oliva, 2022). Consequently, these systems tend to misclassify already disadvantaged social groups, playing a crucial role in further exacerbating their inability to counteract the testimonial injustice connected to wrongful credibility assessments made considering their risk scores.

A second epistemically and ethically relevant issue that the framework of testimonial injustice successfully captures is that these systems can result in patients being silenced and deprived of a major epistemic right: to actively convey information. Strikingly, these systems displace physicians from their authoritative position (Oliva, 2022; Szalavitz, 2021), and at the same time, they lead to a shift back to a more paternalistic approach in medicine, from a patient's perspective. Therefore, physicians are less empowered because, as previously pointed out, their medical decisions can be considerably influenced by ML outcomes, and patients are less involved in decision-making since the

credibility of their testimonies is strongly deflated by the risk scores assigned to them. Thus, they end up being merely recipients of medical decisions that are in alignment with their risk scores. The bottom line is that these systems constrain epistemic participation, possibly undermining the widely accepted principle of shared decision-making in medicine (McDougall, 2019).[7]

Furthermore, when PDMPs are treated as markers of trustworthiness, the testimonial injustice they perpetrate cannot be considered interpersonal anymore (i.e., happening between two or more interlocutors) but rather assumes a structural character: it is not in the hand of a single physician to look beyond a risk score generated by the system, as a physician is limited in their epistemic and moral agency by these systems (Pozzi, 2023a). Expecting physicians to make medical decisions upon mostly unchangeable and flawed risk scores thus becomes an institutionalized procedure. It follows that the injustices perpetrated go beyond the episodic instances of testimonial injustice that can occur in human-human interactions. As Anderson points out, "(t)estimonial exclusion becomes structural when institutions are set up to exclude people without anyone having to decide to do so." (Anderson, 2012, p. 166) As a direct consequence, a "contestability vacuum" emerges: the impossibility of achieving recourse in the occurrence of an inaccurate risk score and the scale of propagation of harm caused by these systems is what makes ML-induced testimonial injustice particularly harmful (Symons & Alvarado, 2022). This is not to say that human credibility assessments are always flawless. They can be, of course, just as biased. However, in a non-ML-mediated scenario, the biased decision of a single doctor with prejudices that is not prone to provide a patient with pain medication for epistemically invalid reasons (e.g., a patient's gender) can be, at least in principle, spotted and amended by non-biased physicians who are

[7]There is research showing that in some cases, patients were not even informed that automated systems were involved in the decision-making and medication denial was based on their risk scores (Oliva, 2022; Szalavitz, 2021). This can be seen as contrary to practices of shared decision-making in medicine. More needs to be said about how ML-based systems impact shared decision-making. However, due to the limited scope of this paper, I cannot pursue this issue further.

involved in a medical decision-making process. In contrast, ML systems, such as PDMPs, can systematise inequality so that contestability escapes the possibilities of single individuals. The shift from an interpersonal to a structural dimension thus bears a significant moral component.

4.5. Final remarks

In this paper, I considered how the use of ML systems, such as automated PDMPs, as markers of trustworthiness that inform physicians' judgements of patients' credibility, brings about instances of testimonial injustice in ML-mediated medical practices. These considerations advanced the question of whether using ML systems to identify patients at risk of drug abuse is epistemically and morally legitimate.

I maintain that this is not the case because striking the right balance in credibility assessments is paramount in this kind of practice. Using these systems can shift the balance in patients' disfavour, particularly for those belonging to already disadvantaged social categories. If I was successful in showing the epistemic and ethical limitations of these systems in terms of testimonial injustice, it is clear that trying to ameliorate the opioid crisis by outsourcing delicate decisions to ML systems fails its purpose.

While this paper shows the limitations of systems such as automated PDMPs, it does not provide possible solutions. However, having a clearer grasp of the problem is a step needed to move forward in striving for the development and deployment of ML systems that consider the fundamental principle of justice, not only in its ethical but also in its epistemic significance.

5

Machine Learning-Induced Hermeneutical Injustice in Healthcare[1]

5.1. Introduction

The implementation of AI-based methodologies—particularly machine learning (ML) techniques—in healthcare has the potential to provide improved diagnostic accuracy that could by far exceed physicians' expertise. Particularly impressive results have been achieved, for example, in the field of radiology, pathology, and ophthalmology (Bejnordi et al., 2017; Golden, 2017; Rampasek & Goldenberg, 2018; Singh et al., 2018). According to the current stand, machine vision can interpret specific medical images as accurately as—or even more accurately—than humans (Topol, 2019, p. 47).

[1]This chapter is based on the following article:

Pozzi, G. (2023). Automated opioid risk scores: a case for machine learning-induced epistemic injustice in healthcare. *Ethics and Information Technology, 25*(1), 3. https://doi.org/10.1007/s10676-023-09676-z

Moreover, AI systems implemented in healthcare play a considerable role in managing the challenges raised by the current COVID-19 pandemic (Lim et al., 2022). For example, the MIT Technology Review reports of an AI device used in hospitals in the UK to perform initial readings of a patient's chest X-rays to be able to recognize the features of COVID-induced pneumonia in the fastest way. As such, delicate and decisive decisions regarding patient triage have been happening according to the recommendation of AI systems, contributing to managing the vast patient loads during the current pandemic (Hao, 2020).

These considerations suggest that we have a *prima facie* moral reason to make use of these systems since they are supposed to provide healthcare professionals with the most suitable and advanced techniques to improve healthcare provision in fundamental medical practices such as diagnostic, treatment recommendations, and prognosis, among others.[2] However, epistemic limitations connected to how these systems operate are a reason of great concern in the current scientific debate, particularly when assessing the results produced by algorithmic systems implemented in critical areas such as healthcare. In fact, the decision-making logic of ML systems is often epistemically inaccessible to the human knower, constituting the widely discussed problem of epistemic opacity. Indeed, said opacity leads to the concern that high-impact procedures are shifting away from human control. Throughout this article, I take the notion of epistemic opacity as defined by Humphreys (2009) and formally advanced by Durán and Formanek (2018). Drawing on Humphrey's formulation of the problem (Humphreys, 2004, 2009), these authors particularly focus on the justificatory aspects of epistemic opacity, defining it in terms of accessibility and surveilability conditions on justification. According to their definition of epistemic opacity, a human agent, due to her limited cognitive resources, not only fails to access every relevant step in the justificatory chain

[2]So understood, the use of computationally powerful AI systems is very much aligned with the fundamental medical principle of beneficence (Lawrence, 2007). See also Van den Hoven (1998, p. 100).

but, should accessibility ever be possible, she would not be able to check every relevant passage (Durán & Formanek, 2018, p. 650).

Most machine learning and deep learning algorithms implemented in healthcare are epistemically opaque.[3] The peculiarity of these algorithms is their ability to change their decision-making rules autonomously as more information is introduced into the system (Alpaydin, 2014). Therefore, these systems can reach a level of complexity that, combined with the processing of huge amounts of data, is not graspable by human cognitive abilities. So understood, the problem of epistemic opacity translates into ethically relevant questions that concern, among others, the degree of trust we are justified to attribute to the outputs produced by these systems and under which circumstances we are epistemically and morally justified in acting upon them (Mittelstadt et al., 2016). For these reasons, debates about the epistemic opacity of ML systems as well as their implementation in highly sensitive fields such as healthcare have, legitimately, gained increasing attention in the academic debate in recent years (Esteva et al., 2019; Grote and Berens, 2020; London, 2019). Nevertheless, the question of how genuinely epistemic forms of injustice emerge due to the role played by ML systems has not acquired a central stage in the current debate yet.[4] This paper aims to address the issue of ML- induced epistemic injustice, specifically related to the use of ML systems in the context of medicine and healthcare.

Very broadly, epistemic injustice is lurking and needs to be explicitly addressed as soon as epistemic limitations of ML systems represent an obstacle to the meaningful participation of relevant stakeholders (e.g., physicians and patients) in medical decision-making processes. This happens, I submit, if

[3] Of course, not all AI implementations are epistemically opaque. Methods like simple naive Bayes classifiers, decision trees, linear regressions, and rule lists are usually interpretable (Lipton, 2018; Páez, 2019). Nevertheless, since most of the AI systems currently deployed in healthcare are either machine learning or deep learning systems, I am interested in the analysis of implementations that are epistemically opaque to human agents.

[4] Of course, there are some exceptions. For example, the recently published and insightful paper by Symons and Alvarado (2022) initiates a discussion on epistemic injustice in data science technologies.

the system is incontestable from the side of human experts and if it establishes knowledge-creating processes[5] that systematically exclude patients' lived experiences as legitimated sources of knowledge that are qualified to inform decision-making. I consider the conceptual framework of epistemic injustice as developed by Miranda Fricker in the field of social epistemology (Fricker, 2007) and its application in the context of ML implementations in healthcare. I am convinced that Fricker's pioneering work can be pivotal in unveiling subtle forms of injustice that are potentially going unnoticed in the current debate revolving around the epistemology and ethics of ML in the field of healthcare (Section 5.2).

Against this background, the main goal of this paper is to address issues of hermeneutical injustice, i.e., a form of epistemic injustice, starting from the consideration of an ML-based tool currently deployed in the USA to predict patients' risk of opioid misuse, i.e., Prediction Drug Monitoring Programs (PDMPs, see Section 5.3). Through careful consideration of the current flaws identified in this system's functioning, I show that it fulfills three fundamental conditions to identify forms of hermeneutical injustice suffered by patients affected by PDMPs' decision-making (Section 5.4). I further argue that the hermeneutical injustice these systems bring about is mainly due to the fact that an automated hermeneutical appropriation (Section 5.5) has occurred.

The overall aim of this paper is to show that ML-induced epistemic injustice is a present and pressing concern. In doing so, I also aim to enrich the current landscape of ethical and epistemological issues connected to opaque ML systems implemented in healthcare. This should hopefully motivate further research aiming at providing adequate answers to the concerns raised.

[5] An AI system's knowledge-creating process can be, for instance, how a scoring system attributing to people risk scores of being at drug abuse elaborate information regarding a particular subject leading to a precise output. The piece of knowledge produced by the system would be, in this case, the score attributed to said individual. I extensively discuss this case in the sections below.

5.2. Epistemic injustice

Drawing on Fricker's groundbreaking book (Fricker, 2007), a large body of literature has arisen, seeking to expand her theoretical framework and trying to find new critical areas in which the analysis of this concept can be helpful to uncover dormant practices that undermine epistemic subjects as knowers. Kidd et al. (2017) point out that "[i]n the era of information and communication, issues of misinformation and miscommunication are more pressing than ever. Who has voice and who doesn't? Are voices interacting with equal agency and power? In whose terms are they communicating? Who is being understood and who isn't (and at what cost)? Who is being believed? And who is even being acknowledged and engaged with?" (Kidd et al., 2017, p. 1) These questions assume particular relevance if we consider that ML systems are powerful epistemic entities increasingly involved in decision-making processes that strongly impact the life of knowing subjects in a morally salient sense. For these reasons, the phenomenon of epistemic injustice (and the relevant questions it brings about) in the field of ML deserves particular attention.

Fricker (2007) distinguishes between two forms of epistemic injustice: *testimonial* and *hermeneutical injustice*. Even though the analysis of ML-induced testimonial injustice is surely worth pursuing,[6] this paper's focus is on hermeneutical injustice. Fricker conceptualizes hermeneutical injustice as stemming "from a gap in collective hermeneutical resources—a gap, that is, in our shared tools of social interpretation." (Fricker, 2007, p. 147) Hermeneutical resources can be defined as cognitive and linguistic resources, i.e., concepts and words we deploy to understand and communicate about the world. They

[6] Fricker defines testimonial injustice as occurring when a speaker is downgraded in her role as a knower due to prejudices a hearer holds related to her social identity (Fricker, 2007). Most notably, testimonial injustice emerges when a person is granted less credibility for epistemically invalid reasons such as her race, gender, or social status. In this contribution, I focus exclusively on how hermeneutical injustice can be machine learning-induced in the particular case of interest. However, also a testimonial injustice is in place. I discuss this form of injustice in detail elsewhere (Pozzi, 2023b).

are collectively shared to the extent that they are widely comprehensible to society at large (Mason, 2021, p. 248).

Hermeneutical injustice emerges in two problematic cases that concern either the absence of or a failure to apply shared hermeneutical resources. First of all, issues in terms of hermeneutical injustice arise when a gap exists in the shared pool of said resources. That is to say, a subject S is experiencing something that she cannot make sense of and, consequently, cannot express to others because what she is experiencing is not part of the shared meanings created and accepted by society. For example, consider the time before postpartum depression (PPD) was recognized as a medical condition (Fricker, 2007, p. 149). At that time, neither the word *PPD* nor the concept **PPD** was available. Hence, a gap was present in the pool of shared hermeneutical resources, leading to a lack of understanding that can amount to a proper injustice. In fact, this disorder was not recognized for a long time as a medical condition requiring due medical consideration because of women's hermeneutical marginalization in biomedical research. Healthcare professionals did not sufficiently understand the symptoms, and, as a consequence, women clinically diagnosed with PPD often felt ashamed because of the feeling of not being able to meet the standards for motherhood imposed by society. This was due to a lack of conceptual and linguistic tools needed to comprehend one's own experience and render it intelligible to others (Chung, 2021). In such cases, the injustice amounts to the fact that S is taken to be disappointing duties connected to motherhood instead of being recognized as experiencing a particular psycho-physical medical condition for which she cannot be considered morally blameworthy. The impossibility of "giving a name" to this experience obscures understanding and hinders communication practices, leaving women alone with what they are experiencing.

Second, hermeneutical injustice can emerge in different terms, i.e., in the case that there are hermeneutical resources available and widely accepted within a society to conceptually grasp and linguistically express a certain expe-

rience e. However, subject S does not feel represented by these. For instance, this would be the case if her lived experience falls outside the scope of or is not aligned with the socially accepted definition of e. Fricker shows this by referring to the definition of homosexuality imposed by society in the 1950s and the protagonist's experience in Edward White's autobiographical novel "A Boy's Own Story" (Fricker, 2007, pp. 163-168). Here, the injustice amounts to the fact that the definition imposed on S shadows his own identity and self-understanding. In this case, hermeneutical injustice is not due to a lacuna in collective hermeneutical resources but rather emerges from a failure to apply to one's own lived experience the concept as it is rendered collectively available (Mason, 2021).[7]

In the face of these considerations, it can be more generally said that in cases of hermeneutical injustice, the subject cannot properly comprehend her own experience and, consequently, cannot render it communicatively intelligible to others. What is important to highlight in this respect is that no agent perpetrates hermeneutical injustice—it is a purely structural notion, according to Fricker. For this reason, in the absence of a clearly identifiable perpetrator, this form of epistemic injustice is particularly difficult to overcome.

The analysis of aspects related to the health condition of patients that put them in a position of epistemic vulnerability that can compromise their epistemic confidence has attracted much attention in the literature (Chung, 2021; Kidd & Carel, 2017; Kidd et al., 2017). That is to say, standard, non-AI-mediated epistemic practices between patients and physicians are already prone to put the patient in a position of epistemic weakness (Blease et al., 2017; Carel & Kidd, 2014).

[7]Fricker does not explicitly recognize hermeneutical injustice as a failure in the application of available conceptual and linguistic resources. However, I follow here Mason (2021)'s interpretation as I think it captures at best this second kind of issue that is not related to a lack of hermeneutical resources but rather to a misalignment between the definition available and the subject's experience.

Against the theoretical background provided by Fricker, I aim at analyzing the role that AI systems play as a further authoritative epistemic entity in medical decision-making. In fact, when an AI system offering recommendations and diagnoses[8] enters this fragile ground, it is paramount to ensure that this does not further weaken the patient's epistemic position. More concretely, should this be the case, it would mean, in its broadest sense, that patients are being attributed a deflated level of credibility. Moreover, it would imply that they are excluded from shared decision-making and are limited in fundamental cognitive activities such as understanding *due to* the role played by the ML system involved in the decision-making process. Considering that ML systems are epistemically authoritative, hardly contestable, and a constitutive part of decision-making procedures that directly impact patients' lives, we need to make sure that they do not undermine them in their capacities as knowers.

To avoid abstract considerations, the theoretical framework underpinning the identification of forms of epistemic injustice needs to be applied to the careful analysis of concrete cases. Hence, in the following sections, I analyze instances of epistemic injustice arising in connection with how ML-based Prescription Drug Monitoring Programs (PDMPs) are currently deployed throughout the US to determine opioid users' likelihood of overdose and drug misuse.

[8] Here, a clarification is in order. Throughout this paper, with phrasings such as, for instance, "an ML system *provides* recommendations", I do not mean to explicitly attribute any form of agency to ML systems. The issue of whether or not ML systems as technical artifacts can be attributed some degree of agency is a thorny and highly debated issue, towards which this paper remains neutral. I thank an anonymous reviewer for encouraging me to clarify this point.

5.3. Injustice in the production of knowledge: the case of PDMPs

To face the challenges raised by the ongoing opioid crisis pervading the USA (cf., e.g., Vadivelu et al. (2018)), municipalities throughout the country have adopted automated PDMPs to control opioid prescriptions (Oliva, 2022). PDMPs have already been introduced as electronic databases in the 1990s to monitor the prescription of controlled substances to sink the risk of misuse, addiction, and overdose (Haines, Savic, et al., 2022). The rash advancement of AI and ML in the last decade has led to the implementation of advanced algorithm-based PDMPs (Oliva, 2022). The main goal of these systems is to contribute to containing the opioid epidemic by attributing to patients a risk score to determine their likelihood of developing opioid misuse. The score provided to each patient is supposed to inform medical decision-making regarding which therapies to subject patients to, which medicines to prescribe, and whether a pharmacist should grant patients access to opioids (Oliva, 2022).

The dominant algorithmic platform used to determine patients' risk scores is called NarxCare and is produced by the company Appriss (recently renamed Bamboo Health) (Szalavitz, 2021). The fact that a proprietary algorithm is deployed to determine patients' likelihood of drug misuse constitutes a black-box since information regarding the data sources used to produce the results are not made publicly available. As such, the latter cannot be reproduced or externally validated (Szalavitz, 2021). Therefore, the users of these systems are not provided with any explanation regarding how the system came to generate a particular output in classifying an individual as having, say, a high risk of drug abuse. Particularly worrisome is also the fact that the proxies adopted to determine risk scores are questionable in themselves (Oliva, 2022). In fact, to estimate patients' risk scores, factors such as the following are considered: the distance a patient travels to reach a physician/pharmacy, the number of

specialists consulted within a specific time frame, payment method used to purchase medicines, number of prescribers, whether the patient in question has a history of sexual abuse or other similar traumatizing events, criminal history, among others (Oliva, 2022, p. 97).

These risk indicators augment stigmatization and discrimination of minorities, disadvantaged socio-economic groups, and patients with a complex medical history, not to mention that they can produce genuinely misleading results. For example, someone living in a rural area is more likely to travel a long distance to purchase her medication. Furthermore, patients with particularly serious pathologies are more likely to consult more than one medical professional to receive the suitable amount of care that their particular condition requires. These, taken for themselves, innocuous facts automatically raise a patient's risk score because traveling a long distance to purchase medicament and having multiple doctors are indicators of so-called "doctor shopping" behavior, which is taken to be strongly connected to a high risk of drug misuse (Oliva, 2022, p. 97). Other plausible reasons why a patient needs to travel a long distance to receive medical assistance are not considered in the risk score estimation. For these reasons, deploying these risk-scoring systems as a basis to inform medical decision-making can have dramatic consequences for patients, particularly those belonging to minorities and/or socially disadvantaged groups.

As I further consider below, even though Bamboo Health states that "NarxCare scores and reports are intended to aid, not replace, medical decision making" and continues affirming that "[n]one of the information presented should be used as sole justification for providing or refusing to provide medications" (see https://bamboohealth.com/narxcare-and-patients/), the reality of daily medical practice strongly contradicts these claims. In fact, even if Bamboo Health insists that these tools are not conceived of as substituting medical decisions, and the last decision is always in the hands of experienced professionals, physicians are expected to make use of these systems and consider

their outputs. Healthcare providers can prescribe opioids to red-flagged patients at their own peril. The risk of being labeled an overprescriber could have extreme consequences for them and even lead to losing their practicing license (Oliva, 2022, p. 103). As a matter of fact, a study conducted by Picco et al. (2021, p. 8) reports that the number of patients being refused medical assistance and medication supply and of patients being discharged from practice has considerably increased due to the role taken up by PDMP platforms.

Sadly in line with what has been said so far, Szalavitz (2021) reports the story of Kathryn, a young woman suffering from endometriosis, a pathology that causes her severe abdominal pain that she could mostly get under control by being administered oral opioids. On one occasion, when she went to the hospital with severe pain, she was administered opioids to alleviate it. However, a couple of days later, she was dismissed from the hospital with no explanation and still in a precarious health condition. It turns out that, to her unknown, her PDMP risk score resulted in being very high, and on this basis, her gynecologist abandoned her interrupting their relationship (Szalavitz, 2021). Her situation is deeply problematic for multiple reasons. First, she was not informed that an automated system was involved in such a delicate decision-making process in the first place (Szalavitz, 2021), and when she was dismissed from the hospital, she lacked any understanding of what was happening to her. Moreover, it is unclear from which sources NarxCare gathers data to develop risk assessment scores, raising important concerns regarding patients' right to privacy and informed consent. Even though these are all ethically problematic aspects that require due attention, the focus of my analysis will be limited to whether patients experiencing situations similar to Kathryn's can be considered a victim of hermeneutical injustice and, if so, how the particular forms of ML-induced epistemic injustice they are suffering can be conceptualized. Frame the issue in question and spelling out its problematic characteristics is the first step to raising awareness of the problem and starting to work toward possible solutions.

5.4. Defining hermeneutical injustice in ML

Let me continue with analyzing the epistemic position of a patient whose risk score has been defined by the PDMP algorithmic system. For the sake of my argument, I consider the situation of a patient who has been, like Kathryn, mistakenly flagged as being at high risk of opioid misuse.

To substantiate the claim that the epistemic authority assumed by opaque ML systems such as PDMPs in decision-making processes in healthcare can lead to a form of hermeneutical injustice at patients' expenses requires to show that:[9]

1. the ML system involved in a decision-making process (diagnoses, treatment recommendation, among others) holds an *unwarranted epistemic privilege* (it is an epistemic privilege because it can establish meanings and plays a decisive role in shaping shared hermeneutical resources. It is unwarranted because it eludes human intervention);

2. this unwarranted epistemic privilege and the way in which hermeneutical resources are established hinder *understanding* and *communication* among relevant stakeholders involved in decision-making processes (physicians, patients). This renders significant aspects of social experience not intelligible to them.[10] If this is the case, the combination of these factors points to the fact that an *automated hermeneutical appropriation*[11] has occurred;

[9] The following conditions are a revised version of the ones formulated by Wardrope (2015, p. 344) along Fricker's lines. I revise them to capture the role played by the ML system in question in being conducive to hermeneutical injustice. These three conditions are closely related to each other and I make a clear distinction between them for analytic purposes.

[10] In which ways this happens is, of course, context-dependent. In the following, I discuss how this is happening in the case under scrutiny.

[11] I discuss what automated hermeneutical appropriation amounts to in more detail in section 5.5.

3. the interplay of 1. and 2. leads to considerable disadvantages, particularly at the expense of hermeneutically marginalized groups (e.g., patients, minorities, already stigmatized groups due to substance use disorders, etc.).

I argue that these conditions are all met in the case considered, and we are therefore dealing with a clear instance of hermeneutical injustice. I address these points in turn.

5.4.1. Condition 1: PDMPs and unwarranted epistemic privilege

Let me start with condition 1. In order to be able to assess whether the ML system in question holds an *unwarranted* epistemic privilege, it is essential to have a clearer view of when such a privilege can be considered as being indeed *unwarranted* and when not. For example, it is highly plausible to take that a physician holds a *warranted* epistemic privilege when interpreting, say, the meaning of a patient's CT scan and acting upon said interpretation (Carel & Kidd, 2014, p. 536). In this case, the physician's epistemic privilege is justified in her capacities and expertise that render her epistemically well-positioned in offering a grounded interpretation of a CT scan. That is to say, her epistemic privilege is warranted by a combination of long medical training, experience with different forms of patients' diseases, the information provided by scientific literature, and similar other sources, which contribute to rendering her an overall reliable and trustworthy professional. On the other hand, an example of an unwarranted epistemic privilege could be the role that standardized protocols play in certain healthcare practices. One could argue that standardized protocols constrain patients' testimony of lived experiences in rigid schemes that cannot account for their more subjective experiences (Moes et al., 2020). If, due to the rigidity of protocols, the subjective experience of patients is

not considered a legitimated source of knowledge (because of, for example, the difficulty or impossibility of expressing it in quantifiable terms) and, as such, is *excluded altogether* from informing decision-making, one can conclude that standardization holds at least the possibility of taking up an unwarranted epistemic privilege.[12] More precisely, this would be the case if standardization had a considerable bearing on which forms of knowledge are recognized as such and can inform decision-making and in the case that the form in which the evidence is presented is decisive with respect to epistemic participation in medical decision-making.

An unwarranted epistemic privilege can come to light also in connection to how hermeneutical resources, i.e., concepts and meaning, are established and shared. As such, considerations regarding the existence of an unwarranted epistemic privilege seem to apply to how the PDMP system under scrutiny establishes meanings connected to a patient's drug misuse. In the following, I argue that these systems enforce their meaning of what opioid addiction amounts to on both patients and physicians, exceeding the decision power of a system that should be functional to improving and supporting human decision-making and not replacing or impairing it, paving the way to cases of hermeneutical injustice.

In the case considered, a problem connected to hermeneutical resources is not to be traced back to a lack thereof. In fact, we certainly have appropriate linguistic and conceptual tools, such as, for instance, the word *addiction* that is suitable to articulate the concept of **addiction**. We also know that the latter amounts to clearly definable criteria above and beyond the fact that the concept has a grounded medical definition (I take addiction, but it could

[12] Of course, whether this amounts to an unwarranted epistemic privilege or not in a particular case is context-dependent. It surely strongly depends on how healthcare professionals deal with the information gathered in protocols and with other forms of knowledge that exceed them (say, whether they take the time to engage with the patient in question in order to account for the knowledge they possess and could not be captured through the rigidity of protocols. In the latter case, the potential unwarranted privilege of these would be compensated through appropriate conduct from the side of the physician).

also be the words and concept related to what is medically understood under substance use disorder (SUD), for example). The problem amounts to the fact that the ML system defines the concept of **addiction** according to not shared metrics so that the *meaning* attributed by the system to a red flag can considerably shift away from the widely shared, accepted, and medically grounded meaning of the same concept, without the possibility for stakeholders affected to amend this shift.

This comes to light in the face of several considerations. First of all, in defining parameters for determining, for example, the threshold that allows the system to differentiate between concerning and not concerning cases in connection with the risk of opioid misuse, value-laden choices are necessarily made. These range from the choice of the proxies used to determine the risk scores (along with each proxy's weight in determining the final score) to the model's design and the definition of the goal that the system should fulfill. How this is done by developers of the system, the company owning it, or the system itself due to its self-learning and adaptive mechanisms has the effect that this is not rendered explicit and transparent to the stakeholders directly involved in and affected by these systems' decision-making. It follows that this entails failing to ensure that the system's representations indeed mirror the accepted definition of the concept.

Furthermore, Oliva (2022) points out that the end goal of these systems is to reduce opioid prescriptions, regardless of the consequences that this has for the patients affected by the risk scores produced. This means that as long as it can be shown that due to the use of PDMPs, physicians issue fewer opioid prescriptions, the system is deemed *efficient*. However, this is also a value-based choice that does not show that patients flagged as being at a high risk of opioid misuse indeed are running this risk (according to the collectively shared meaning of concepts such as **addiction** or **SUD**). NarxCare does not assess whether clinical prescription decisions improve or exacerbate patients' pain, mental health, or overall quality of life (Oliva, 2022, p. 88). That is to say, the

measured effectiveness of the system is limited to the number of prescriptions issued, without further investigating whether a particular patient actually benefits from the interruption of opioid medicament and she was indeed at risk of opioid misuse, or she was red-flagged due to correlations established by the system that are not indicative of opioid addiction or misuse. It follows a shift in the meaning of addiction or SUD as it is established by the system toward its end goal of reducing prescriptions. However, patients are confronted with a risk score that is treated as grounded knowledge that reliably mirrors their drug consumption levels.

Consequently, a misalignment can emerge between the knowledge that a patient (like Kathryn) has about herself and a risk score taken to indicate drug misuse and addiction, according to the definition established by the PDMP system. It follows that ambiguous and heterogeneous information that can be tangential and not representative of establishing a risk assessment is treated and elaborated as a form of *knowledge* suitable to guide medical decisions. This is, to my mind, the extent to which algorithmic PDMPs such as NarxCare hold an unwarranted epistemic privilege.

The latter is considerably reinforced by the fact that PDMP predictive platforms are "the only law enforcement-developed digital surveillance systems that health care providers have ever utilized to diagnose and treat patients." (Oliva, 2022, p. 51) Thus, even though, in theory, they should serve as tools to support physicians' decision-making, in practice, they strongly limit patients' and physicians' possibilities to critically question the risk score they assign. Bottom line, they take up a decision power that sidelines the weight that patients' knowledge—in the form of their testimony, personal (moral and epistemic) values, and lived experiences—can have in the process of medical decision-making.

Furthermore, the concern emerges that how the ML systems under scrutiny produce what is considered legitimate knowledge able to inform medical decision-making is unidirectional: physicians are provided with a score that

cannot be revised according to relevant information that could potentially overturn it. That is to say, users are not able to feed back into the system valuable information that should be taken into consideration, making the user just a recipient of knowledge coming from unknown sources, but they do not get to actively influence the knowledge-producing process itself (Pozzi & Durán, 2024).

In line with these considerations, it can be stated that the ML system involved in the case under consideration holds a clear unwarranted epistemic privilege: it establishes what it means to be at a high risk of drug misuse by attributing to each patient a risk score related to their likelihood of misusing opioids based on questionable proxies and without allowing for contestability of the results. The fact that the consideration of these systems' outputs is legally enforced on physicians exacerbates the weight of their authority even further. Condition 1 is thereby fulfilled.

5.4.2. Condition 2: Understanding and communication impairments

Let us move forward with the consideration of condition 2. As already mentioned, being NarxCare a proprietary algorithm, the possibilities to get meaningful insight into how it works and which criteria were decisive in estimating a patient's high-risk score are very much constrained. That is to say, in this case, technical opacity—due to the black-box nature of ML algorithms used in generating the risk scores—is reinforced by the unwillingness of the company that owns these systems to disclose information regarding their functioning (Oliva, 2022). Regardless of the sources of the opacity of these systems, the result is that users involved in and affected by the system's decision-making processes lack the explanatory resources needed to have a proper understanding of how its outputs are created and, as such, are not able to assess whether

these are justified or not. Thus, these considerations imply that they cannot show that the systems' results have been produced, in a particular case, by the interplay of factors not indicative of drug misuse. This leaves patients and physicians in the dark regarding what actions can be legitimately taken upon the generated output and how patients can question a risk score they do not identify with.

The general lack of understanding leads to the fact that communication between patients and healthcare providers is strongly constrained. There is relevant literature pointing at communication difficulties emerging from the ubiquitous use of PDMP systems throughout medical practice, which are particularly concerning and indicative of patients' potential to suffer epistemic injustice. After having conducted interviews with medical professionals using PDMPs as a basis for decision-making, Hildebran et al. (2016) individuate detrimental communication styles that are, to my mind, representative of the disruptive role taken up by these systems in mediating interactions between patients and physicians. The quotations below are particularly concerning expressions of how these systems' mediation in medical practice can be detrimental to a "good" patient-physician relationship.[13] As pointed out by these authors, they amount to avoidance and a confrontational communication style (cf. also Picco et al., 2021, p. 14), both of which are particularly concerning and indicative of hermeneutical injustice. In the context of the interviews conducted by Hildebran et al., 2016 and Hildebran et al., 2014, physicians stated that:

[13]In the last decades, a paternalistic approach in medicine has been slowly replaced by a more active role taken up by patients towards shared decision-making. The latter is particularly ethically relevant as it enables patients' autonomy and self-determination (McDougall, 2019). The communication styles ensuing from the increasing use of PDMPs for medical decision-making seem to shift patient-physician interactions back to more paternalistic models. While I cannot pursue this issue further here, I think it is important to consider it.

> "It's a cat and mouse thing. So I keep it [the PDMP] secret as much as possible." (Hildebran et al., 2016, p. 2063) (*avoidance*)

> "I will leave the room to get something and pull it [PDMP report]. I may confront them if it seems like they're lying and say, 'Well here's what I got here. It seems like you haven't been very honest with me and so I'm not going to provide you with prescriptions.'" (Hildebran et al., 2014, p. 1183) (*confrontational*)

The first communication style, avoidance, completely closes down communication between patient and physician. The patient in question is not informed of the existence of an epistemic authority, i.e., the PDMP producing the scores, that has a strong or even *the* final say on crucial matters regarding their mental and physical well-being. On top of this, the risk score is not critically scrutinized by the physician. The latter could, in fact, gain fruitful information regarding the patient's condition and drug consumption engaging in an open discussion with her, showing empathy and understanding for a possible problematic situation in which she is in, being this, arguably, an essential aspect of the patient-physician relationship.[14] In such cases, the patient is not even given the possibility to share and communicate the knowledge she possesses about her current state, which could show her miscategorization as a red-flagged patient. This kind of communication style shows how the role assumed by PDMPs encourages physicians' disengagement with their patients, with detrimental consequences for the latter. In fact, it is practically impossible for the patient to make sense of why she is being denied medication or medical assistance altogether, being wholly excluded from the communication practice. It is evident that this can create a gap in understanding and sense-making of one's own situation.

[14] This seems to be the case if we take a deliberative approach on shared decision-making as the paradigm underpinning the patient-physician relationship (Emanuel & Emanuel, 1992).

While avoidance suffocates any possibility for the patient to react to a potentially inaccurate risk score, a confrontational communication style, along the lines of approaches similar to the one mentioned in the passage reported, seems to be strongly conducive to distrust and credibility deficits on the side of the patient (Pozzi, 2023b). This is the case even if, at least in principle, she is provided with the possibility to argue against a possible risk score with which she does not identify. In the face of a red flag that categorizes a patient as being at a high risk of drug misuse, it will be particularly challenging for her to argue otherwise in the case that the risk score attributed to her does not depict her actual drug consumption, particularly if the physician attributes *by default* more credibility to the PDMP, as it seems to be the case in the passage previously quoted. In this sense, we could say that PDMPs directly influence the credibility judgments of healthcare professionals and could exacerbate already existing prejudices. Concretely, this means for patients that the humane, empathetic, communicative experience between them and their physicians that is paramount to coping with their mental and physical condition of being ill is strongly impaired.

Drawing on these considerations, the risk of epistemic injustice is evident in the case that a stigmatizing authority (i.e., PDMPs) plays a role in mediating decision-making processes in medical care. In fact, above and beyond possible (subconscious) prejudices of healthcare professionals towards socially disadvantaged groups due to their perceived social identity (Kidd & Carel, 2017), a risk score is attributed to them, and, as previously mentioned, this tends to disadvantage members of said groups. It seems indisputable that the risk score further deflates their credibility and weakens their epistemic position.

These considerations indicate communication difficulties emerging from the role that PDMPs play and show that condition 2 for hermeneutical injustice is also fulfilled.

5.4.3. Condition 3: Hermeneutical disadvantage

Condition 3 aims at capturing how the interplay between the considerations expressed in the analysis of conditions 1 and 2 constitutes a disadvantage, particularly for members of minorities and social groups vulnerable to discriminatory practices. From what has been previously pointed out, it seems plausible to take that the decision power of these ML systems, combined with their incontestability, provides them with an epistemic authority that imposes a form of unchangeable knowledge. Consequently, patients are excluded from the process of informing medical decision-making and are *hermeneutically marginalized*. Kathryn was hermeneutically marginalized at the moment in which, without receiving any explanation whatsoever, was sent home from the hospital. Her marginalization continues today since even though she could grasp why the system was erroneously attributing her a high score (she has sick pets that need strong medicament),[15] she could not find a way to clear her record. Moreover, her knowledge continues to be ignored every time she unsuccessfully seeks the support of a physician ready to prescribe her the medicines she needs to cope with the pain caused by her complex medical condition (Szalavitz, 2021). Kathryn's situation represents a palpable instance of hermeneutical injustice.

In the face of what has been said so far, the nature of the disadvantage can be captured in epistemic, moral, and practical terms. As already mentioned, an epistemic disadvantage is at play since patients' knowledge is not considered in informing the PDMP's risk score. It follows that patients' testimony, personal values, and lived experiences are not accounted for as a legitimate source of knowledge that is able to contest an unjust risk score.

[15] A brief side note is due here. Not every patient is in the condition of fighting against a risk score that is not perceived as depicting one's own actual opioid consumption as Kathryn did. This could be the case due to different factors depending on one's own situation, such as, for instance, physical and mental condition or technical literacy. Therefore, it is paramount to give voice to injustices that can arise from the fact that said systems make it so difficult for stakeholders affected by a wrong risk score to show otherwise.

From a moral point of view, these systems fuel stigmatization, discrimination, and unfair treatment, particularly for already disadvantaged societal groups. For instance, women are particularly exposed to miscategorizations by the PDMP due to the proxies used. The fact that sexual trauma is considered an indicator that raises the likelihood of drug abuse disadvantages women. As a matter of fact, they are, on average, more likely to report sexual abuse and seek psychological support compared to men (Oliva, 2022). This means that they will, by default, have a higher risk score than men, a fact that reinforces existing inequalities, increasing the probability that they are thereby denied medication on illegitimate grounds. Gender-related stereotypes connected to women's perceived emotional instability or their alleged tendency to "exaggerate pain" have led to considerable and persisting disparities in pain management already in not ML-mediated processes (cf. Lloyd et al., 2020).

Moreover, it seems clear that the consequences of clinical dependence on models that generate high false positive rates by mischaracterizing patients with, for instance, low-risk complex chronic pain as high-risk opioid use disorder are particularly severe (Oliva, 2022, p. 105). Indeed, this further contributes to stigmatizing and hermeneutical marginalizing already fragile categories of epistemic subjects,[16] strongly constraining their ability to grasp and make sense of what is happening to them. As such, PDMPs reinforce prejudices and discrimination already rooted in our social practices, leading to their systematization. In fact, since PDMP systems are used by physicians almost on a daily basis and are treated as evidence that directly influences medical decision-making, these kinds of biases are likely to propagate considerably, escaping the critical scrutiny of physicians who are possibly aware of these issues and could actively engage in preventing their occurrence. Of course, the same applies to ethnic inequalities in pain care (Mossey, 2011) that see, for example, patients of color face considerable obstacles in receiving access to

[16] On how the risk of epistemic injustice is particularly present for chronic pain patients see Buchman et al. (2017).

pain medication and a lower quality of care more generally (Oliva, 2022, pp. 94-95).

Finally, as is evident in Kathryn's case, patients' disadvantage can express itself in very practical terms: in the denial of medical delivery, in patients' abandonment, and in the condemnation to live with unbearable pain that could be otherwise alleviated. This can lead to very damaging consequences, such as increased suffering, the feeling of not being heard and understood, and being the victim of a system in which one does not get to play a role as an active epistemic subject but is rather the object of decisions that affect their lives to a great extent.

The fulfillment of these conditions shows that an ML-induced hermeneutical injustice is in place in the case under scrutiny.[17]

5.5. Automated hermeneutical appropriation

What has been said in the previous sections points to the fact that an automated hermeneutical appropriation occurs in the case considered. Under this umbrella term, I understand the interplay of factors that are directly conducive to hermeneutical injustice, specifically *due to* the ML system's role in taking over aspects of the patient-physician interaction that should, arguably, remain under human control.

This relates to what has been said in sections 5.4.1 and 5.4.2 regarding the unwarranted hermeneutical privilege taken up by PDMP systems and the communication difficulties that emerge between patients and healthcare professionals due to the role they play in mediating healthcare encounters. Both the idea of causing a misalignment between collectively shared meanings and

[17] Of course, in the case considered, the issues recognized in terms of hermeneutical injustice are not to be restricted to the flaws inherent to the ML system considered in isolation. In fact, these technologies are sociotechnical systems in which technical and institutional design must play a decisive role in mitigating the issues pointed out, along with suitable regulations and correct deployment by relevant institutions (Van de Poel, 2020).

the meaning established by the PDMP and constituting an obstacle in the process of understanding point at a role taken up by the system that, very much intuitively, *exceeds* its allegedly intended purpose of supporting medical decisions.[18]

However, ML-induced epistemic injustice cuts particularly deep because it affects not only members of disadvantaged social categories (such as, for instance, patients with substance use disorder). It also negatively impacts favorably positioned epistemic agents such as physicians that otherwise could, through virtuous behavior (Fricker, 2007, p. 169), mitigate the injustice experienced by patients as vulnerable epistemic subjects. This last point further indicates of an automated hermeneutical appropriation and requires further elucidation.

In the attempt to clarify under which conditions an epistemic injustice can be considered indeed an *injustice* in a proper sense, Byskov (2021) formulates five conditions, one of which is, to my mind, particularly suitable to capture the wrong experienced by physicians as epistemic subjects, i.e., what the author calls the *stakeholder condition*. Byskov defines it as follows: "In order for someone to be unjustifiably discriminated against as a knower, they must be somehow affected by the decisions that they are excluded from influencing." (Byskov, 2021, p. 8)

Physicians' stakeholder rights should, intuitively, encompass the fact that they are entitled to actually influence medical decision-making if they are to be considered epistemic and morally responsible for their decisions. In the case discussed, physicians are not excluded from straightforwardly influencing decisions since they have, at least formally, the last word on whether or not a patient will be granted opioid prescriptions. However, they are expected to make decisions regarding a patient's prescription without being able to do so in a system-independent way.

Kathryn's doctor does not have meaningful insights into the reasons as

[18]See, again, NarxCare's website: https://bamboohealth.com/narxcare-and-patients/

to why the PDMP system attributed to her a high-risk score for drug misuse. Nevertheless, she is supposed to act upon the outcome produced by the system (to her unknown of the relevant factors that led to it). Even if she makes the final decision regarding Kathryn's treatment, she does not get to influence the decision-making process itself in an active way since she is compelled to act according to the risk score provided (Szalavitz, 2021). Hence, physicians themselves can also be considered victims of this system since "(d)epending on the State specific legal requirements of the PDMP, the PDMP database may generate an automated alert to notify either health and/or law enforcement agencies of suspicious prescribing." (Haines, Savic, et al., 2022, p. 2)

Under these circumstances, it is difficult to see to what extent a physician actively influenced the decision-making process when acting according to the risk score generated by PDMPs. Nevertheless, she will indeed be *directly affected* by the consequences of the decision taken because she is likely to be considered blameworthy if the decision made has negative consequences for the patient. Therefore, due to her stakeholder rights, she is supposed to have the possibility to be involved in the decision-making process in a meaningful, genuinely agential way. This does not seem to be the case due to the incontestability of PDMPs and their law enforcement power. These considerations shed light on a further dimension of the system's hermeneutical appropriation to the extent that physicians are, at least partially, deprived of their stakeholder rights. The fulfillment of the stakeholder condition points to the fact that also physicians are, to a certain extent, experiencing epistemic injustice.[19]

The experience of seeing their epistemic authority undermined by an inscrutable and incontestable epistemic entity in a systematic way could have disruptive consequences for physicians' professional identity. Moreover, it also entails the possibility of their deskilling and disengagement, eliciting the ten-

[19] Of course, the consequences of the injustice experienced by physicians in terms of stakeholder rights limitations are not comparable to the hermeneutical disadvantages that afflict patients since the former are still members of a socially privileged group.

dency to evade responsibilities that would otherwise be a constitutive part of her professional role (it has been already pointed out in the discussion of condition 2 how the role of PDMPs leads to physicians' disengagement with their patients).[20]

Physicians' epistemic dependence upon PDMPs has a considerable impact on patients. While ML-based PDMPs deprive patients of the conceptual tools needed to understand why they are red-flagged in the case that the *knowledge* they possess about themselves is not aligned with the score they are stigmatized by, there are no options available to the patient to come out from this circle exactly because, crucially, physicians themselves are epistemically dependent on them. For this reason, their ability to potentially counteract a hermeneutical injustice suffered by their patients is very much constrained.

This indicates that a virtue theoretical approach (Fricker, 2007, p. 174) towards the mitigation of ML-induced hermeneutical injustice suffered by patients is insufficient. Fricker sees the virtue of hermeneutical justice as fundamental to opposing epistemic injustice. The author takes this virtue to be corrective in nature to the extent that a virtuous attitude of a hearer showing awareness of the social situation of a speaker and "a more pro-active and more socially aware kind of listening" is able to partially compensate disadvantages emerging from hermeneutical injustice (Fricker, 2007, p. 174). In healthcare encounters, this would require physicians to be particularly aware of a patient's possible hermeneutical marginalization and their active effort to show understanding for their situation to overcome it.

However, a solution to the issues pointed out in this paper needs to go beyond the virtue theoretical approach indicated by Fricker. In fact, her approach seems to be limited to cases of epistemic injustice emerging in exclusively human-centric epistemic environments. In the case of interest, the

[20]While a thorough analysis of these systems' impact on physicians' professional role is paramount to capture the nature of the injustice they experience, I cannot pursue this issue further here.

difficulty in identifying the oppressing agent and the impossibility for patients to seek recourse to epistemically authoritative agents such as medical professionals (since they are themselves epistemically dependent on the systems) renders the ML-induced injustice they experience even more wide-ranging and difficult to mitigate.

5.6. FINAL REMARKS

The overarching goal of this paper was to shed light on issues understood in terms of hermeneutical injustice and brought about by ML systems implemented in medicine and healthcare. To substantiate my argumentative aims, I analyzed in detail a particularly concerning ML-based system currently deployed throughout the USA to produce patients' risk scores of opioid addiction and misuse. Since physicians are expected to consider these systems' outputs to inform their medical decisions, it is paramount to critically scrutinize whether they increase patients' vulnerability to forms of epistemic injustice.

In order to convincingly argue that this is the case, I showed that three main conditions to recognize instances of hermeneutical injustice are met in the case under scrutiny. PDMPs hold an unwarranted epistemic privilege (condition 1) that impairs understanding and fundamental communication practices among patients and physicians (condition 2), and finally, constitutes hermeneutical disadvantages, particularly for vulnerable social categories (condition 3).

I further argued that ML-induced hermeneutical injustice is to be directly traced back to an automated hermeneutical appropriation from the side of the system. The latter reveals in the way in which hermeneutical resources are established by the system and how it deprives human agents of understanding and hinders their communication practices. On top of this, it deprives physicians of the possibility to actively safeguard patients who are victims of the injustice the ML system brings about since the former are themselves

subordinated to the system's epistemic authority. Crucially, this strongly limits physicians' possibility to resist hermeneutical injustice through virtuous behavior, as Fricker conceptualizes in her human-centric approach.

More needs to be said regarding how these issues take shape in epistemic practices that see ML systems as powerful and ubiquitous epistemic entities. However, I hope this paper could show the importance of further pursuing the highlighted issues and encourage further research to work towards technically feasible solutions with the aim of overcoming the difficulties recognized.

6

Participatory Injustice in Conversational AI[1]

6.1. Introduction

Advances in applications of artificial intelligence and the use of data analytics technology in biomedicine are creating optimism for mental health care. With the launch of new large language models, such as GPT-4, this excitement is mounting. For instance, smartphones are believed to have the potential to aid researchers and therapists in comprehending, predicting, and intervening in human psychological phenomena by monitoring the mental states and actions of their users. One particularly promising resource in this regard is the smartphone psychotherapy chatbot—an artificially intelligent bot that provides cognitive behavior therapy to users, aiming to enhance their mental well-being

[1]This chapter is based on the following article:

Pozzi, G. & De Proost, M. (under review). Keeping an AI on the mental health of vulnerable populations: reflections on the potential for participatory injustice.

(Luxton et al., 2011). Several benefits are commonly mentioned concerning the use of these chatbots for mental health issues, including their cost-effectiveness, widespread accessibility, and availability in various languages (Tekin, 2021). As a result, they are considered an ideal tool, particularly in regions where there is a scarcity of therapists who can communicate in the native language of individuals in need of mental healthcare, such as refugees.

Sedlakova and Trachsel (2022) have considered how the use of conversational artificial intelligence (CAI) tools raises difficult ethical questions related to issues of authenticity, autonomy, and expanding access for vulnerable populations. The vulnerable groups that Sedlakova and Trachsel highlight are the elderly, adolescents, and underdiagnosed people (Sedlakova & Trachsel, 2022). However, other vulnerable populations[2], such as refugees, who lack access to mental healthcare due to historical and cross-cultural treatment gaps, ought to be more central to the discussion of CAI (Knox et al., 2023). As research indicates, there is a general paucity of literature and a lack of evidence available regarding the uptake of mHealth interventions among refugees and other vulnerable populations (Ashfaq et al., 2020).

When specific attention is paid to such vulnerable populations and the use of CAI, various ethical concerns come to light. Principles of biomedical and AI ethics, such as beneficence, non-maleficence, explainability, and justice, are applied in the literature as well (Ursin et al., 2022; Vilaza & McCashin, 2021). However, the latter ethical value has only been limitedly explored thus far. As one of the few studies in AI ethics on this concept, Gabriel (2022) investigates the relationship between artificial intelligence and principles of distributive justice. However, the rationale of 'ideal theory', famous from John Rawls' *A Theory of Justice*, could be radically put into question as non-ideal societies with injustices have historically been the norm rather than the other way

[2]We acknowledge that there is a great deal of vagueness in the definition of vulnerable populations. Refugees should not be considered a vulnerable group *per se*, as each individual should be evaluated based on his/her inherent and situational fragilities and needs (Mendola & Pera, 2022).

around (Mills, 2005). Moreover, recent philosophical writing on the scope of justice has also drawn attention to forms of injustice that do not involve material redistribution, but the harms persons could suffer through failures of recognition and discrimination (Giovanola & Tiribelli, 2022).

Our focus in this paper is on a less familiar kind of harm that CAI can cause in health care, namely the harm to individual human persons as knowers. Thus, we put forward an analysis of the epistemo-ethical impact of CAI on vulnerable populations through the lens of the analytic framework of *epistemic injustice* (Fricker, 2007). These generally underserved populations ought to be central to our discussion of the medical ethics of CAI. Our considerations aim to offer a novel perspective under which the fundamental biomedical principle of justice needs to be scrutinized in broader terms in the face of the role acquired by these systems in crucial practices, such as mental health support to be widely delivered to vulnerable populations. Considerations of epistemo-ethical difficulties in mental health are not new (Kidd et al., 2022; McCradden et al., 2023; Sakakibara, 2023), though little attention has been given to the digital context and the epistemic consequences of CAI for the therapist–patient interaction.

Miranda Fricker recognizes two main forms of epistemic injustice that are to be considered the building blocks of her framework: *testimonial* and *hermeneutical injustice*. In general terms, testimonial injustice occurs at the interpersonal level when a hearer attributes to a speaker a reduced level of credibility for epistemically invalid reasons (e.g., due to identity prejudices). Hermeneutical injustice is a more structural notion that aims to capture a wrong done to someone when, due to their marginalization, they do not have the conceptual resources to make sense of and express to others their social experience. Both testimonial and hermeneutical injustices could play significant roles in CAI for mental health care.[3] However, in what follows, we frame our discus-

[3]Initial considerations on the potential for testimonial and hermeneutical injustice through the general use of mental health chatbots can be found in (De Proost & Pozzi, 2023).

sion in terms of one broad form of epistemic injustice that such practices are especially prone to, given the technology's nascent status: *participatory injustice*. This injustice tracks one comprehensive category of epistemic encounters: engagement as participants in knowledge generation (Hookway, 2010). So understood, participatory injustice takes place among two or more interlocutors but is not restrained to purely testimonial interactions in which epistemic exchanges are limited to conveying and receiving information. Participatory injustice aims to capture a whole range of epistemic activities in which a knower is (un)accounted for in their capacities to make hypotheses, contribute to the formation of knowledge, and to acquiring self-knowledge, among others (Hookway, 2010).

The dialogical nature of psychotherapeutic encounters is not only aimed at a transfer of information between patient and psychotherapist (Miner et al., 2019). The therapist's role is also to accompany the patient through self-reflection and, ultimately, self-understanding, leading them to rethink and re-evaluate certain possibly detrimental beliefs and form new ones (Tekin, 2021). Thus considered, the range of epistemic activities associated with a therapeutic relationship is wide and requires the full and active participation of the patient. This should take place in an environment in which they feel acknowledged, taken seriously in their concerns, and capable of successfully engaging in relevant epistemic activities. Hence, the extent to which the epistemic participation of patients in this rich sense is possible through the use of mental health chatbots needs to be critically scrutinized.

In this article, we proceed as follows. Section 6.2 discusses a case of a mental health Chatbot, Karim, deployed to deliver mental health support to Syrian refugees. This case substantiates our theoretical considerations and the epistemo-ethical concerns brought about by the use of mental health applications among vulnerable populations. In Section 6.3, we introduce the theoretical framework of participatory injustice. In Section 6.4, we consider how conceptualizing epistemic participation as a capability to be accounted

for through the framework of Capability Sensitive Design could lead to the mitigation of participatory injustice when it emerges in connection with the use of mental health chatbots among vulnerable populations.

6.2. AI-mediated mental health support for vulnerable populations: the Karim Chatbot

The risks of using chatbots in vulnerable populations have received limited attention in the ethics of AI literature. Scholars have devoted particular attention to the potential accountability gap created by such systems where there is an absence of a therapist. For example, in the case that someone expresses suicidal ideation, the system lacks the capacity to react appropriately, as dramatic consequences of chatbot responses in such delicate situations have sadly shown (Graber-Stiehl, 2023). It is unclear that a CAI can be trained to handle such a crisis situation, and even more unclear who should take responsibility if the CAI fails to mitigate this harm as well as a human could have. In another case, a company called Koko, provided emotional support chat services based on GPT-3 for 4000 people in distress without asking for consent. When users learned of this unauthorized experiment, many felt betrayed. The division of responsibilities in such an experimental situation was once again ambiguous (Haupt & Marks, 2023).

Apart from responsibility concerns, the main argument put forward in the literature is the need for greater efficiency: estimations suggest that for every 100,000 people worldwide, there are about 4 psychiatrists on average; that number is much lower in most low- and middle-income countries with about 1 psychiatrist for the same amount of people (Rathod et al., 2017). In the face of the overall shortage of therapists to meet the psychological needs of vulnerable populations, the hype surrounding AI-based technologies is often seen as a possible remedy as the quest to automate therapy could democratize access.

Particularly smartphone apps or chatbots are increasingly used to offer mental health support, mostly through cognitive behavior therapy (CBT). As Tekin (2021) points out, the enthusiasm revolving around the use of these systems can be interpreted to be based on three main promises that these chatbots seem to be able to uphold. The first is that digital phenotyping allows early diagnoses and treatments, thus improving patients' chances of early recovery (this is arguably also a good way for patients that do not recognize alarming symptoms themselves to become aware of them and seek support). The second is that they represent an alternative solution for people who do not feel comfortable seeking psychological support due to the stigma attached to it. Arguably, sharing intimate concerns with a chatbot instead of a human agent could decrease patients' fear of being judged by their therapist. And the third, more general promise, is that access to psychotherapy is increased through the use of these technologies, supporting populations whose mental health needs would not be otherwise met.

In this paper, we focus particularly on the third promise mentioned, i.e., the fact that these systems are supposed to provide mental health support to populations whose mental health issues would otherwise remain unaddressed, thus enabling broader access to mental health support. Considering the case of a mental health chatbot introduced to provide mental support to Syrian refugees,[4] we build upon Tekin's skeptical arguments regarding the efficacy of such technology by broadening the landscape of ethical and epistemological issues connected to it. In particular, we do so by elucidating how these systems can be used with the risk of bringing about an epistemic injustice, more specifically, a participatory injustice (a form of epistemic injustice) in Hookway's sense (Hookway, 2010). We argue that it is not epistemically and ethically justified to try to resolve a problem in terms of distributive justice (i.e., the fact that human therapists are a scarce resource, particularly in the

[4]It should also be noted that there is a fine line between the definition of refugee, migrant, and asylum-seeker. This is important to consider because each immigration status causes different ethical concerns.

context of refugee mental health) at the cost of causing new issues in terms of epistemic injustice. Moreover, we show that chatbots such as Karim will likely disappoint the expectations created by the third promise mentioned since, as we will argue, it considerably impairs the epistemic participation of refugees in therapeutic communication.

In order to bridge the gap to participatory injustice, let us reconstruct some characteristics of the chatbot of interest. In March 2016, the Silicon Valley start-up X2AI (now Cass) launched "Karim", a psychotherapy chatbot, to support Syrian refugees in Lebanon (Solon, 2016). The chatbot uses natural language processing, a form of AI, to simulate human conversations in Arabic through existing communication channels such as SMS texts or Facebook Messenger. This chatbot was reportedly piloted on 60 Syrians "mostly men and boys". This is a strikingly small pilot for scaling up to a large and vulnerable population: there are over one million Syrian refugees in Lebanon. X2AI developed the pilot in partnership with "Field Innovation Team", a non-profit specializing in technology in disaster recovery, and the so-called "Singularity University", the Silicon Valley business incubator and consultancy service. In the report of the Field Innovation Team, it is mentioned that the chatbot encountered issues with translation because of the many variations in Arabic dialects. Instead of using standard Arabic and google translate, they hired Syrians to resolve translation issues to the Damascus (Levantine) dialect (Field Innovation Team, 2016).

Karim is not explicitly marketed as a psychotherapeutic tool but rather as a "friend" (Solon, 2016). Several issues can emerge against this background. One has to do with how these systems should be conceived of in the first place. The FDA recently relaxed regulations regarding how mental health chatbots can be sold as medically grounded devices. In fact, in the face of the mental health crisis brought about by the COVID pandemic, what was previously conceived as a "wellness" application can now be rebranded as a proper medical intervention (Mattioli, 2021). So, even if the line between the extent to which

chatbots similar to Karim can be considered proper medical devices is quite blurry, they are *de facto* used to provide mental health support. In the case of Karim, this applies to particularly vulnerable populations whose mental health needs differ substantially from other, more privileged populations.

The latter point seems particularly relevant in the face of the fact that Karim has been developed as a version of Tess, a chatbot used in the USA to support people with an anxiety disorder or mild depression. While Tess serves as a therapeutic tool supplementing and not replacing a human-human psychotherapeutic relation, the use of Karim among refugee populations is unsupervised by trained professionals (Madianou, 2021). However, particularly for refugees who have most probably experienced traumatic or even life-threatening events, the presence of a human psychologist is even more crucial in order to be able to intervene in a situation of emergency.

Moreover, as with many other mental health support tools (Tekin, 2023), Karim has not been subjected to empirical scrutiny. The few empirical studies that tested mental health applications report positive results on patients' mental health, however always if the tool is a supplement and not a replacement for the psychological support that human therapists can provide. It is largely recognized that using mental health apps in an unsupervised setting is quite controversial, and its effect and possible perils are untested (Manríquez Roa et al., 2019).

While all these considerations are central in the analysis of the ethical impact of these systems, in this paper, we aim to elucidate a more subtle issue related to the effective possibility of epistemic participation that people using a chatbot like Karim have. To this goal, consider the following points that, as we argue, lay the ground for the occurrence of epistemic injustices in refugee mental health through the use of chatbots similar to Karim.

The first important point is that, as Tekin (2021) points out, sociocultural factors have an impact on mental health and illness. That is to say, the imposition of Western criteria of how psychotherapy should work upon Mid-

dle Eastern populations with a different sociocultural background encodes an *identity bias* into the technology, thus excluding people who do not identify the standards it follows. This can create fundamental difficulties in making one's experience accessible to the technology. The possible mismatch between users' experience and the concepts available to the systems can be seen as a first step toward epistemic injustices arising in connection with the use of these technologies among vulnerable populations. In fact, users' possibility to properly engage with these tools can be constrained due to a gap between their lived experiences and the pre-determined options encoded into the system (De Proost & Pozzi, 2023; Pozzi, 2023a, 2023b).

To support these claims, consider the use of chatbots in refugee support as described by an interviewee in Madianou (2021): "All chatbots are about pushing information out. Even 'Refugee Text' is: 'tell us your status and we'll give you some information on that basis'. Maybe at best it's tailored information, but it's not a conversation. [...] Participation is hard to do. It's easy to push out information". Hence, people's possibility to participate in an epistemically meaningful communicative experience can be strongly impaired. As we argue in the next section, this paves the way for instances of participatory injustice to emerge.

On a similar note, Sedlakova and Trachsel (2022) argue that "the CAI as an algorithm-driven system is good in providing quantified data or factual information which are limited in range. This type of knowledge can be categorized as third-person knowledge that can inform patients about relationships, human mind, or psychological processes. However, this type of knowledge is insufficient to gain new self-understanding and constitute a therapeutic change." Here it becomes clear that therapeutic interactions are not limited to passing on and receiving information, that is, to testimonial exchanges in a restrictive sense of the term.[5] In contrast, therapy entails epistemically richer interac-

[5]Admittedly, Fricker's definition of testimony in her discussion of testimonial injustice is quite encompassing, being understood in the broadest sense of the transmission of knowl-

tions and activities in which understanding, self-understanding, hypothesizing, and critically analyzing are only some of the many relevant ones.

Against this background, our central aim in this article is to provide a theoretically informed analysis of these issues and make more explicit their ethical and epistemological consequences. In the following section, we spell out the notion of participatory injustice against the backdrop provided by the case of the chatbot Karim just discussed.

6.3. Epistemic harm beyond testimony: toward participatory injustice

There is still precaution on therapeutic possibilities of chatbots due to the preliminary nature and the early stage of research in this area. Moreover, it is unclear that Chatbots as technological artifacts can constitute a testifier since such technologies, unlike people, lack moral character and well-being. In the literature on social epistemology, an "anthropocentric view of testimony" is commonly held based on the presupposition that only persons can participate in the act of testimony because only humans, in principle, can qualify as testifiers (Fricker, 2007).

Because of this fact, we want to focus on the early phases of knowledge production and possible related harms. In Fricker's standard view of testimonial and hermeneutical injustice, knowledge transmissions, in the form of credibility deficit and interpretative obstacles, are central. However, there are many other core epistemic activities related to the generation of knowledge beyond

edge (Fricker, 2007, 2010). However, Hookway takes the argumentation a step further considering epistemic activities that do not necessarily rely on receiving or transmitting information.

giving testimonies and conceptual interpretation. Other kinds of epistemic injustice are thus possible beyond those focused on by Fricker. For instance, Dotson (2011) conceptualizes a preemptive self-censoring of the content and expression of speakers' testimonies as "testimonial smothering". Especially in (digital) therapeutic conversations, the epistemic subject is not offering a piece of knowledge or opinion. Rather, they are attempting to move the discussion down some particular line of inquiry to see what results.

Christopher Hookway was the first to make this point, in his critical commentary of Fricker's monograph, where he emphasized the central importance of cooperative epistemic endeavors and argued that there is a wide variety of types of participant contributions that lead to the success of cooperative epistemic pursuits. These contributions reach well beyond offering or seeking testimony. He introduced the concept of the participatory perspective in epistemic injustice to describe how knowers could be unfairly excluded from participating in non-testimonial epistemic practices such as those involved in querying, conjecturing, and imagining (rather than a mere "informational perspective"). Hookway argues that a wide range of distinctively epistemic harms can occur when participation in inquiry is unfairly compromised. As Hookway puts it, "the resources we make use of in exercising our epistemic agency are richer and more varied than is often supposed. [...] Someone may not be credited as sufficiently trustworthy as an 'epistemic agent', and this judgment may reflect identity prejudices, even if their evaluation as unreliable is not made in the context of a straightforward testimonial exchange." (Hookway, 2010, p. 153)

He offers the example of a teacher who, although willing to take students' informational questions seriously in their role as students, does not give a student uptake when they ask a question that is intended as a contribution to the inquiry itself. What happens in such cases is that someone who wishes

"to be recognized as a member of a community of people collaborating in the attempt to improve understanding or advance knowledge" fails to be so recognized (Hookway, 2010, p. 155). When one is not taken seriously as a participant in inquiry, one can lose epistemic confidence or self-trust, becoming too tentative in one's contributions. When one's questions are ignored, one may develop a habit of silencing oneself, not asking relevant questions that might forward the investigation (Hookway, 2010). Hookway's approach broadens the very concept of epistemic injustice in a helpful way and underscores that what is common to a wide range of cases of unfair epistemic treatment that falls under the category of epistemic injustice is the compromise of the epistemic agency of a marginalized group.

Based on the above-described case of Karim, one could imagine a similar scenario as in the classroom. The refugees were not treated as potential participants in discussions on the development of the application but just as testing subjects who can ask for and provide additional information. This could be based upon a stereotypical view of the value of refugees' contributions to the debate. Due to prejudice, the company fails to respect the refugee as a potential contributor to the discussion (or participant in the discussion). The result is that the refugees can no longer think of themselves as a participant in inquiry and discussion. They become epistemically disabled or, what Medina describes as "epistemically disempowered", because the company fails to take the refugees' mental health questions seriously (Medina, 2022).

We argue that the situation of participatory injustice just described predominantly occurs as a consequence of two assumptions seemingly built into the design of the chatbot under scrutiny. The first has to do with a participatory prejudice that amounts to regarding the intended users of the system as *objects* and not as participants (Carel & Kidd, 2017) in the epistemic activities ensuing from the use of the chatbot. Carel and Kidd consider this

form of prejudice related to more general medical practices in which the role of patients in interactions with medical professionals is often restricted to reporting or confirming symptoms or anagraphic information, excluding a more substantial epistemic involvement. These considerations can be transferred to the case of interest since the user interaction with the system does not leave space for the kind of "cooperative epistemic inquiry" (Grasswick, 2018, p. 316) that would be needed for a successful interaction geared toward mental health support.

The second consideration pertains to assumptions related to the trust that end users can, indeed, participate in an epistemically substantial way. As Medina (2020) points out, participatory justice "involves being trusted in one's overall epistemic competence and participatory skills, and not just as a possessor of knowledge but also as a *producer* of knowledge" (our emphasis). According to this view, the failure to design the CAI in such a way as to allow genuine epistemic participation of the user could underlie the failure to entrust them with the capacity of genuine epistemic participation. Both assumptions are detrimental to users' epistemic standing in the ways previously described.

Let us also point out that the latter observation has a bearing on whether epistemic subjects interacting with the chatbot can fulfill their role as epistemically autonomous agents. As Tanesini (2022) points out, epistemic objectification in Fricker's sense, i.e., understood as being denied the possibility to convey knowledge and testimony, hampers the epistemic value of intellectual autonomy since epistemic agents are effectively constrained in their role of informants. The same issue can arguably also occur under a broader definition of epistemic injustice as participatory injustice in the case under scrutiny. Being the person interacting with the CAI the object rather than the subject of the interaction, her possibility to be an active and autonomous enquirer toward the purpose of establishing a therapeutic exchange aiming at mental health support remains precluded to her.

6.4. Mitigating participatory injustice through Capability Sensitive Design

It is sometimes suggested that the remedy to problems of participatory injustice is the development of individual virtues (Hookway, 2010). Fricker proposes "virtuous listening" as a helpful corrective but partial solution to issues of epistemic injustice. However, some scholars stress the need to go beyond the dyadic instances of epistemic injustice on which Fricker often focuses, aiming for more encompassing solutions (Sherman, 2016). Particularly in the context of systematic epistemic injustices such as the ones brought about by AI-based systems, it seems appropriate to explore principles for the cultivation of epistemically just technologies, as well as social and political institutions (Anderson, 2012; Symons & Alvarado, 2022). In a similar vein, we believe that the more comprehensive approach of Value Sensitive Design (VSD), especially its development through the capabilities approach, i.e., Capability Sensitive Design as outlined by Jacobs (2020), can support epistemically just CAI deployed in mental health. Let us first reconstruct some main characteristics of the Value Sensitive Design approach.

The need to couple ethical considerations with design choices ensues from the consideration that a system's design can bring about a positive and/or negative change (Mink et al., 2014) and that technological artifacts function as "agentive amplifiers" in that they can create possibilities that were previously unavailable to the agent (Van den Hoven, 2012). It is widely agreed upon in the current debate that technologies are not value-neutral but rather the product of choices encoded into a system's design. Therefore, it is paramount to shape technological developments with shared moral values (Van de Poel & Kroes, 2014; Veluwenkamp & Van den Hoven, 2023). The overall aim of the Value Sensitive Design framework is thus to translate core values into normative considerations, which are further concretized into precise design

requirements that can be implemented. Following a tripartite methodology, the design process is considered from three interconnected levels: a conceptual, an empirical, and a technical level of investigation (Friedman et al., 2013). In an iterative process of moral inquiry and interdisciplinary deliberation, this methodology is used to define values to be translated into a technology's design by establishing standards and analyzing technical requirements that guarantee the practical implementation of the values. In this line of argument, the focus is not on a retrospective ethical analysis but rather on proactively shaping and anticipating the development of a value-sensitive design: ethical and social considerations are incorporated into the design process and thus also support the technical conditions for ethical development at an early stage (Bleher & Braun, 2023).

However, scholars considering the application of VSD have pointed out limitations pertaining to this approach in its standard formulation (Jacobs & Huldtgren, 2018). For example, one problem highlighted is the identification of stakeholders (Manders-Huits, 2011), resulting in the central question of whose values should be effectively included in the design process in the first place. A second criticism indicates that the normative dimension often remains underdetermined in this approach. As Jacobs and Huldtgren point out: "VSD makes no explicit commitment to particular ethical theories." (Jacobs & Huldtgren, 2018, p. 1) Umbrello and van de Poel (2021) address this issue by linking it to a further relevant shortcoming that comes to light also once VSD is applied to AI technologies (such as machine learning systems), namely its lack of sensitivity for political and social contexts. These authors propose a human rights framework as a possible solution in which a context analysis precedes the identification of relevant values.

A further widely discussed approach that aims to ameliorate the issues briefly described is Capability Sensitive Design (CSD), a framework combining the method of VSD with the capability theory advanced by Martha Nussbaum, thus backing up VSD with a needed theoretical underpinning. Jacobs (2020)

has recently considered the application of this framework to AI systems in well-being and health. In this section, we build upon Jacobs' work and argue that CSD can be useful in addressing the problem of participatory injustice in the mental health CAI application of interest in this paper. To achieve this goal, we proceed as follows. First, we motivate why it can be fruitful to conceive of *epistemic participation* as a capability in the first place. In the second step, we show that seeing epistemic participation as a capability embedded in the context of CSD has two beneficial effects. The first is that it provides us with the theoretical tools needed to spot a participatory injustice, since, as we have seen, these can occur in a rather subtle manner. Second, it provides the theoretical basis needed for designers to critically question whether a particular CAI could bring about these issues, thus anticipating possible problematic outcomes in terms of participatory injustice.

Introducing in detail CSD and Nussbaum's capability approach goes way beyond the scope of this paper, so we just focus on a few key aspects. The primary aim of this approach is to design technologies that enhance and expand users' fundamental capabilities. Nussbaum lists ten central capabilities,[6] (1) being able to live a normal length of lifespan; (2) having good health; (3) maintain bodily integrity; (4) being able to use the senses, imagination, and think; (5) having emotions and emotional attachments; (6) possess practical reason to form a conception of the good; (7) have social affiliations that are meaningful and respectful; (8) express concern for other species; (9) being able to play; and (10) have control over one's material and political environment. The assumption underlying Nussbaum's capabilities list is that every individual has a right to pursue a life worth living and, to this end, they should be able to exercise these basic capabilities (Jacobs, 2020).

[6] Nussbaum's approach that provides a finite list of human capabilities arguably applicable to any individual irrespective of societal, cultural etc. differences does not remain uncriticized in debates revolving around the capability approach. See, for example, Claassen (2011) for a critical assessment.

Fricker explored the link between epistemic injustice and the capabilities approach and argues that epistemic contribution can be conceived of as a fundamental human capability, thus deserving to be included in Nussbaum's capabilities list (Fricker, 2015).[7] More specifically, Fricker conceives of epistemic contribution as a "combined capability" following Nussbaum's tripartite definition of different capabilities (Fricker, 2015). A combined capability is one that is developed and trained, but that requires certain social conditions to be in place for it to effectively flourish. An example would be the capacity to express one's sexuality. The internal capacity to do so can be developed by an individual. However, it can turn into concrete expression only as long as suitable external conditions are in place. For instance, in a situation of oppression and/or discrimination, a person's capability of expressing their sexuality would not acquire the status of a combined capability since disruptive societal mechanisms would prevent them from effectively expressing this capability. In a similar vein, Fricker claims that wrongful exclusion or lack of credibility for unjustified reasons means that a person does not receive the social uptake needed to transform her innate ability to transmit knowledge, transforming it into a capability that she can successfully exercise.

In her analysis of epistemic contribution as a capability, Fricker focuses particularly on social reciprocity, insisting on the fact that we are not only epistemic receivers but also epistemic givers. Informational material and interpretative material are the two forms of epistemic giving constituting the epistemic capabilities that are constrained in cases of testimonial and hermeneutical injustice (Fricker, 2015). In framing epistemic contribution as a capability in these terms, it is evident that Fricker's approach remains at the level of

[7] Fricker's claim that this capability deserves an extra spot on Nussbaum's list needs more substantiation than we could possibly offer in this paper if we were to argue that the capability of epistemic participation should be added as well. Probably less controversially, we think the latter can be subsumed under Nussbaum's formulation of the capability to use the sense, imagination, and think. This is the case because we have a broader conception of epistemic participation in mind than Fricker's informational view. We expand on this point later in the section.

receiving and conveying information. From an informational perspective, a person's capability of contributing epistemically would be limited if she was, due to prejudicial considerations, deemed as an untrustworthy informant, for instance.

However, the participatory perspective we are interested in goes beyond this more restrictive understanding of a subject's epistemic contribution, and so does the capability ensuing from it. One of Nussbaum's listed capabilities is particularly noteworthy in relation to participatory injustice, i.e., "*being able to use the senses, imagination, and think.*" (Nussbaum, 2000) As pointed out in the previous section, Hookway's account of participatory injustice captures forms of epistemic injustice that exceed informational exchanges between two or more interlocutors. In fact, the conceptualization of this form of injustice aims to shed light on practices that unfairly limit the subject in their possibilities not only to share information and knowledge but to create knowledge or gain a deeper understanding, among others. For example, in a therapeutic relationship, the patient does not only need conditions in place for her to be able to pass on information to her therapist through testimony (in a descriptive fashion, e.g., subject X is experiencing anxious feelings) but to hypothesize, challenge, and possibly change her beliefs. There are thus richer epistemic activities that are crucial to a successful therapeutic relationship and do not necessarily involve transmitting and acquiring information. We maintain that these activities can be performed if users have the possibility to exercise the capability of *epistemic participation*. This encompasses the central epistemic activities mentioned and could be thus considered a subcategory of the more general ability of imagining and thinking, as recognized by Nussbaum.

We thus conceive of epistemic participation as a more encompassing combined capability that exceeds Fricker's informational approach. In fact, epistemic participation requires having trust in the fact that a subject is competent in their ability to ask pertinent questions, advance understanding of a certain subject matter through critical scrutiny, inquiring into a problem's solution.

The activities that these capabilities comprehend go beyond the informational ability of receiving and sharing information. Nevertheless, similarly to the capability of epistemic contribution envisaged by Fricker, epistemic participation also requires appropriate development that can succeed through societal uptake. Recalling Hookway's example in the classroom, a positive, unbiased disposition of a teacher with respect to the epistemic competencies of her student of being able to advance knowledge and understanding of a particular subject matter are necessary conditions for their capability of epistemic participation to flourish. Therefore, both an informational perspective and a participatory perspective, i.e., the one under scrutiny in this article, can be captured by a capability approach that highlights individuals' epistemic agency and focuses on the external conditions that allow the subject to realize their capability.

Now that we have clarified in which sense we can conceive of epistemic participation as a human capability, we need to consider how CAI for mental health support can endanger and/or enhance it. Jacobs (2020) conceives of CSD as following the tripartite division of VSD in conceptual, empirical, and technical investigation. The steps pertaining to each investigation are not to be understood as a linear process but rather as a continuous back and forth in which these investigations reciprocally inform one another in a process of constant re-evaluation. Therefore, it is not possible to analyze these dimensions in isolation. However, due to the limited scope of this paper, we cannot expand on each component of CSD, so let us advance some initial considerations pertinent to the conceptual investigation in relation to the case previously discussed.

As Jacobs points out, the goal of the conceptual investigation is threefold: individuating the capabilities pertaining to the technology of interest, focusing on the stakeholders affected by the technology, and recognizing relevant conversion factors. In the case under scrutiny, the capability that we are interested in analyzing is epistemic participation as previously described, and the stakeholders involved are, very broadly, vulnerable populations receiving

mental health support through a CAI application (such as in the case of Karim previously analyzed). Let us turn to some considerations related to conversion factors.

Conversion factors encompass the degree to which a person is able to transform a resource, in this case, an AI-based technology, into a capability (i.e., epistemic participation) (Jacobs, 2020). For the CAI under scrutiny to support refugees' mental health, we thus need to consider which factors would prevent them from using the technology to enhance their capability of epistemic participation. Against the background provided in the previous sessions, two main factors need to be scrutinized.

The first relates to assumptions built into the technology regarding the role that users can play in their interaction with the mental health support app. As previously pointed out in Section 6.3, the assumption that refugees are objects instead of subjects of mental health support comes to light in the case in which they are confronted with information outputted by the system but do not get to participate effectively in an exchange in a more epistemically substantial way. Changing this assumption and conceiving of users as subjects of an interaction geared toward mental health support is the first conversion factor that we need to account for to enhance their capability of epistemic participation. Otherwise, this risks to remain, *by design*, precluded to the user.

The second consideration has to do with a contextualization of the use of these systems for a particular population. To transform the CAI into an exploitable resource, we need to consider cultural diversity as a paramount conversion factor. The fact that Karim is the follow-up version of an app developed and implemented in the USA for people with mild depression or anxiety (see Section 6.2) presupposes that the way in which psychotherapy is done in Western countries can be applied to a population with a completely different cultural background and mental health needs. This can result in a built-in bias (Luxton, 2020) that imposes Western values onto a culturally

different population. Context-sensitive considerations pertaining to the societal values, background knowledge, and expectations of these systems' target population are thus paramount to designing for their active epistemic participation and avoiding that they cannot effectively realize their capability of epistemically participating *through* the technology.

6.5. Final remarks

The main goal of this paper was to expand on the ethical and epistemological assessment of the use of mental health chatbots among vulnerable populations. More specifically, we aimed to show that using these systems to mitigate issues of distributive injustice due to the scarcity and/or unavailability of human therapists can, as a downside, bring about less explicit but not less harmful forms of epistemic injustice.

Drawing on the case of the chatbot Karim used to provide mental health support to Syrian refugees, we showed that these systems could lay the ground for a particularly harmful form of epistemic injustice, i.e., participatory injustice. As we have argued, this amounts to the fact that these systems' users are fundamentally constrained in many crucial epistemic activities that we would otherwise consider central to successful therapeutic interactions. These amount to the possibility of gaining self-understanding, inquiring into one's own mental health situation, modifying a set of disruptive beliefs leaving space for new ones, hypothesizing, and critically questioning, among many others. Against the backdrop provided by our analysis, this paper's contribution to ongoing discussions on the ethics and epistemology of mental health chatbots is threefold.

First, our analysis provides reasons why, to achieve an ethically sound use of mental health chatbots, we need to ensure that the users' epistemic status as autonomous knowers and inquirers is not endangered, crucially, *through* the

use of these systems. To achieve this goal, the framework of participatory injustice was applied to a novel field of inquiry, i.e., mental health chatbots.

Second, we shed light on the ethical issues of using these technologies, specifically among vulnerable and generally underserved populations whose specific circumstances often remain under-researched in their specificity. The case of the chatbot Karim brings thus to the forefront problems related to the effort of finding a technological solution to important societal problems, such as refugees' mental health.

Third, this paper provides initial considerations on how to conceive of epistemic participation as a central human capability and how the Capability Sensitive Design framework can ameliorate issues of participatory injustice in CAI technologies. Thus, we provide insights into initial considerations on how to address the epistemological and ethical issues identified, and that can be encountered using these technologies, specifically among vulnerable populations.

Further research is needed to consider how the capability of epistemic participation can be translated into appropriate norms and design requirements to be built into CAI technologies in support of the mental health of vulnerable populations. However, the initial considerations advanced in this paper hopefully show the relevance of this approach analyzed in connection with CAI technologies and the risks that epistemic injustices, in general, and participatory injustice in particular, represent for epistemic agents.

7

Social Causes and Epistemic (In)justice in Medical Machine Learning[1]

7.1. Introduction

The social aspects of causality in medicine and healthcare have been emphasized in recent debates in the philosophy of science as crucial factors that need to be considered to enable, among others, appropriate interventions in public health (Russo, 2023). Therefore, it seems central to recognize the bearing of social conditions (broadly understood, e.g., social inequalities, limited access to healthcare resources, (lack of) social inclusion in causing certain concrete

[1]This chapter is based on the following article:

Pozzi, G. & Durán, J. M. Social Causes and Epistemic (In)justice in Machine Learning-Mediated Medical Practices. Forthcoming in *The Routledge Handbook of Causality and Causal Methods* (Illari, P. and Russo, F. eds.).

pathologies. In this chapter, we frame our discussion around different causal levels by considering the synergy between the biological and social levels of disease causation, as these often intersect. Consider, for example, the fact that black people in the USA often tend to suffer from respiratory diseases, such as asthma and pneumonia, at a considerably higher rate compared to white people (Gaffney, 2021). Although the causes of these respiratory pathologies need to be determined in their *biological* etiology and manifestation, leaving out the bearing that *social* factors have in causing this higher incidence of respiratory diseases would provide us with an incomplete and thus less actionable picture (see, again, Russo (2023)). Social conditions, such as exposure to environmental pollution (as predominant in poor neighborhoods), limited access to healthcare resources, material deprivation and chronic stress, contribute to worse respiratory outcomes that disparately affect black people in the USA (Gaffney, 2021). This means that the prevalence of pulmonary diseases in a particular part of the population is also largely caused by social factors that perpetuate conditions of oppression and inequality, such as a "racialized geography." (Gaffney, 2021)

In this case, the social factors that contribute to causing respiratory diseases among black people also have a very strong epistemic component; recognizing them as effective causes contributes to the knowledge and understanding not only of the pathology itself but also of the social situatedness and epistemic standing of underrepresented social groups. Thus, it has relevant ethical implications that need to be addressed. To determine how questions of causality can impact considerations related to individuals' epistemic standing, we think it is fruitful to explore possible connections with debates on epistemic injustice. Let us notice that although connections between causal and normative questions are usually sidelined in the philosophy of causality, with our consideration of epistemic injustice, they acquire a central stage in this chapter.

One of the merits of the framework of epistemic injustice (Fricker, 2007) is that it motivates the need to consider epistemic subjects as embedded in specific social contexts, as these inevitably affect their epistemic status. Authors who have applied this framework to the field of medicine and healthcare have shown that the social identity of patients (their race, gender, social status and health condition) has a considerable bearing on crucial healthcare procedures (Carel & Kidd, 2014). For instance, it has been shown that attributing less credibility to patients for epistemically invalid reasons pertaining to prejudices regarding their race or gender can have very damaging practical consequences (e.g., misdiagnoses) above and beyond the fact that their epistemic status is unjustifiably deflated (Kidd & Carel, 2017).

The causes sought by medical professionals have a bearing on the information that patients share with their physicians. This clearly impacts patients' role as knowing subjects and their possibility of sharing relevant knowledge through their testimony and lived experiences. In turn, as we argue, overfocusing on individuating biological causes while neglecting possible social factors that contribute to a certain pathology can bring about epistemic injustices in medicine and healthcare.[2] Thus, we aim to analyze the normative implications underlying different approaches to disease causation through the analytical lens of epistemic injustice, highlighting the importance of considering the causal role played by social factors, such as racial and gender inequalities. We argue that accounts of disease causation that consider social factors causally relevant on the individual level are, *ceteris paribus*, well positioned to allow patients' testimonial contributions to the medical discourse. Thus, they are an important step toward epistemic justice in medical encounters.

[2]Let us point out that our discussion takes notice of underlying issues of causal metaphysics—such as the nature of causes, causation, and disease causation—as pertinent to the questions of use and normative considerations discussed here (see Illari and Russo's causal mosaic (Illari & Russo, 2014)). While our focus in this chapter is particularly on the latter issues, it is paramount to recognize the close connection between these philosophical questions of causality. We are thankful to Federica Russo for encouraging us to clarify this point.

After establishing a link between questions of causality and epistemic (in)justice, we maintain that being aware of the relevance of social causes in medicine and healthcare (Kelly et al., 2015; Kelly & Russo, 2018) is particularly important in the face of the role that artificial intelligence (AI)-based systems (such as machine learning (ML) algorithms) are increasingly playing in these fields. These systems have the dangerous potential to ignore relevant social aspects and thus conceal relevant social causes. As we show, this is problematic not only because it reinforces issues of distributive injustice but also because it can pave the way to forms of epistemic injustice.

Given this background, the central aim of this chapter is to make the first effort to point out possible intersections between the importance of recognizing social causes in medicine and healthcare and forms of epistemic injustice in ML (Pozzi, 2023a, 2023b; Symons & Alvarado, 2022). Thus, this first exploratory work will analyze how these debates intersect, opening up new venues for fruitful research.

7.2. Accounting for social causation

It is commonly acknowledged that health and disease are crucially shaped by social factors. Conditions related to socio-economic situations of inequality, limited access to education, and problematic family circumstances, among other social factors, play a role in determining health outcomes and the incidence of certain diseases. The bearing of so-called social determinants of health on pathologies across different demographics is also accounted for by the World Health Organization (WHO) (https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1). Although social factors have been recognized as strongly correlated to diseases, the potentially causal functions of social aspects are often disregarded in favor of individuating biological causal mechanisms (Russo, 2023). This may be because a conceptualization of disease

causation that includes social and biological aspects can be highly demanding (Kelly et al., 2015; Russo, 2023). Beyond theoretical challenges, proximate factors in the causal chain are often more easily actionable, requiring pertinent action at the individual level.

However, this can contribute to neglecting more fundamental aspects that are also seen as causally relevant (Henderson, 2022; Link & Phelan, 1995). For instance, in the face of a considerably higher rate of COVID mortality in black populations in the USA, Curry points out that "Black Americans are vulnerable to the bio-social impositions of racial oppression and marginalization. The economic and political segregation of Black people in the U.S. results in demographies that are particularly vulnerable to disease and health related death." (Curry, 2020, p. 262-263) The quoted passage highlights how deeply rooted social issues, such as clear manifestations of systemic racism (unequal healthcare support, marginalization) in the USA, can bring about concrete biologically observable health issues (e.g., a higher rate of COVID cases and deaths among black people compared to the general population).

To tackle these issues, it seems insufficient to focus exclusively on biological mechanisms in disease causation that reduce health disparities to contingent happenings to be treated in their singularity. At a more general level, the latter approach fails to explain why certain pathologies occur more predominantly in certain social subgroups than in others (Russo, 2023). On the individual level, it fails to contextualize a person's experience of illness within broader social structures of power and inequalities; this point is of particular interest in this chapter. Thus, it is crucial to consider how social aspects can play a role in causing diseases, not only to provide a more comprehensive disease etiology. As we argue later in this chapter, this is also important when accounting for individuals' social situatedness, i.e., the need to consider them as embedded into a specific social context in which social identities and structures impact them as *epistemic* subjects.

There are efforts in current debates aimed at conceptualizing the coexistence of biological and social factors in disease causation. For example, Ghiara and Russo (2019) advance a conceptualization of disease causation, aiming to capture the hybrid nature of social and biological causes (see also Russo (2021)). In the remainder of this section, we briefly reconstruct these authors' accounts to, in a further step, highlight relevant normative implications.

Ghiara and Russo develop an account aimed at effectively capturing when social factors "get under the skin" and have an impact at the biological level (Ghiara & Russo, 2019, p. 6). We take this account of social causation as a starting point for our normative analysis because the authors gear efforts toward a more precise conceptualization of the causal role played by social aspects. They do so by introducing the concept of *socio-markers* to be identified at the *individual level*.

According to Ghiara and Russo, in measuring the social dimension of disease causation, socio-markers differ from indicators and proxies in that the latter are correlated to something that cannot be directly measured: using these, we are limited to saying that a strong correlation exists between certain social factors and health issues. However, we cannot say *how* the causal relationship functions (Ghiara & Russo, 2019). The role of socio-markers is more specific in individuating concrete happenings or features of an individual's living circumstances in bringing about a concrete pathology. As the authors point out, "(t)he aim, when using sociomarkers, is to pick up signals to reconstruct the continuum from social factors to disease, in analogy with how biomarkers help pick up signals from exposure to disease at the biological level." (Ghiara & Russo, 2019, p. 13) Thus, just as biomarkers are used to reconstruct the connection between exposure and a certain pathology at the biological level (Illari & Russo, 2016), socio-markers have the same function but are related to social factors.

Consider the following example put forward by the authors. Adverse Childhood Experience (ACE) indicators have been developed to point out

causal connections between traumatic experiences during childhood and adverse health outcomes.[3] Some ACE indicators include physical or emotional abuse or whether a household member was suicidal or imprisoned during childhood. At a general level, these indicators aim to show that there is indeed a strong correlation between relevant socioeconomic factors and health outcomes. This correlation can be made explicit by measuring the incidence of ACEs across populations with low socio-economic status. However, in more contextual terms related to the experiences made by individuals, ACE indicators also work as proper socio-markers in that they can be applied to a case-based approach to inquire about events that crucially mark the causal continuum between experiencing a traumatic event during childhood and the occurrence of a specific pathology. Using ACEs as socio-markers, it would be possible to determine whether they play a difference-making role in the emergence of certain pathologies. As the authors point out, empirical studies "have provided evidence of the associations between ACEs and allostatic load biomarkers, inflammatory biomarkers and cancer biomarkers (...) support(ing) the idea that ACEs could be the link connecting 'the social' and 'the biological'." (Ghiara & Russo, 2019, p. 15)

The importance of this line of research is to bring nearer to medical practice the need to uncover causal mechanisms that are bio-social in nature. According to this interpretation, social aspects are more than just fundamental but distal factors in the causal chain that cannot be actionable.

Now, having considered the importance of bringing social causation into medical practice, we need to take a step further to analyze its normative implications. In fact, which account of disease causation underlies medical decision-making in crucial practices, such as diagnostic procedures, has a direct bearing on what information physicians admit to the medical discourse.

The remainder of this chapter is dedicated to explicitly stating these im-

[3]For our purposes, this example aims to make the idea behind socio-markers as a central component in the conceptualization of social factors in disease causation more explicit.

plications, highlighting how hybrid accounts of disease causation constitute an important step toward epistemic justice. As we argue, these approaches are better positioned to facilitate inclusion in the medical discourse of information that needs to be provided *directly* by patients. In the next section, we briefly introduce the framework of epistemic injustice, detailing, in a further step, its link to questions related to uncovering social markers in disease causation. We then turn, in Section 7.4, to the analysis of these aspects in ML-mediated medical practices.

7.3. Epistemic (in)justice in the (un)recognition of social causes

As pointed out in the previous section, the central aim of this chapter is to argue that the causes sought after in medical practice have normative implications that can be analyzed through the analytical lens of epistemic (in)justice. To establish these connections, we briefly reconstruct this framework and discuss what it aims to tackle.

In its broadest meaning, epistemic injustice refers to a variety of practices that constrain a subject's epistemic standing, understood as their capacity to receive, elaborate and convey information (Kidd et al., 2017). In more specific terms, Fricker, who coined the concept of epistemic injustice, conceptualizes it as taking two main concrete expressions: *testimonial* and *hermeneutical injustice* (Fricker, 2007). Testimonial injustice occurs at the interpersonal level between a speaker and their interlocutor when the former is attributed less credibility than they deserve due to epistemically unjustified reasons (e.g., gender, race and being in a vulnerable health condition). As such, testimonial injustice implies wrongfully withholding credibility due to a hearer's prejudices. In contrast, hermeneutical injustice occurs at a more structural level when certain groups of people do not have access to the linguistic or conceptual

tools needed to make sense of or express what they are experiencing because of their marginalization. Due to the limited scope of this chapter, we restrict our focus to issues of testimonial injustice.[4]

In the medical context, patients can often be in a fragile position due to knowledge asymmetries and physicians' epistemic authority (Carel & Kidd, 2014) or simply for being in a vulnerable health situation in which they depend on other people's care (Kidd & Carel, 2017). For instance, cases in which patients' pain has been underestimated based on epistemically irrelevant considerations, such as prejudices pertaining to their gender or race, abound in the literature (Kidd & Carel, 2017; Trawalter et al., 2012). In such situations, a patient not only risks experiencing under-treatment but is also unjustifiably diminished in their epistemic status since their testimony is groundlessly dismissed. Fricker conceives these forms of harm as leading epistemic subjects to be "undermined or otherwise wronged in a capacity essential to human value." (Fricker, 2007, p. 44) As previously suggested in footnote 2, questions of disease causation analyzed at the biological *and* social levels are relevant for the analysis of the role that patients' knowledge can play in medical encounters. As we show, it is also paramount to consider how the social conditions that characterize patients' experiences at the individual level have a bearing on disease causation. This is, in turn, closely related to the possibilities that individuals have of expressing their social situatedness through their testimony and is thus central to questions of testimonial (in)justice.

Testimonial injustice can emerge in contextualized instances of prejudice and bias. However, at its root, there can be more fundamental questions regarding which *forms of knowledge* are deemed suitable to inform medical decision-making practices (Moes et al., 2020). It can be the case that patients' knowledge is, more generally, disregarded as irrelevant for diagnostic purposes (Epstein & Gramling, 2013). Kidd and Carel (2017) point out that patients are often perceived by medical professionals as incapable of participating in med-

[4]For a thorough discussion of hermeneutical injustice, see Mason (2021).

ical discourse in an epistemically meaningful way, either because they cannot distinguish between relevant and irrelevant information for diagnostic purposes or due to their idiosyncratic way of conveying their testimony. These interactions take place in situations in which practical limitations (such as time constraints) play an important role and can lead to an overall devaluation of patients' active contributions. This is particularly the case for patients who belong to underrepresented or stigmatized social groups in which credibility attribution is wrongfully withheld even more frequently (Carel & Kidd, 2014).

Considering this background, it seems crucial to encourage medical practices that, by default, need to attribute epistemic value to what patients can actively contribute to medical discourse. The nature of the causes underlying patients' symptoms that physicians seek is central in this respect. Further elaboration is needed on this important point.

Often, it is assumed that health problems are accompanied by objective findings (Malterud, 2000, p. 603) and that these can be traced by uncovering biological markers and mechanisms that causally connect patients' symptoms with an explicit biological manifestation (Malterud et al., 2015). In this respect, patients often play the passive role of sources of information from which relevant data for diagnostic purposes can be successfully inferred without them playing an active role. Information that is perhaps no more substantial than merely reporting the main symptoms or confirming anagraphic data (Kidd & Carel, 2017).

Approaches to disease causation that focus *exclusively* on uncovering causes of disease at the biological level presuppose that the information needed about a patient's health condition can be indirectly extrapolated (e.g., through a CT scan or blood test). Hence, patients are considered passive sources of information for the diagnostic process. In contrast, conceiving a patient's disease as caused by the interplay of both biological *and* social markers requires positioning the patient's experience in a broader social context in which factors pertaining to their social situatedness potentially play a role in causing a

particular disease. From this perspective, the patient is no longer exclusively considered a passive source. Instead, they are a necessary source of information for the diagnostic process. The reason is that patients can directly, meaningfully and actively provide their experience to reconstruct the relevant social markers in the causal continuum leading to disease.

Against this background, it is paramount to consider that the conceptualization of disease causation as bio-social has different epistemological implications from purely biological accounts (Russo, 2023), particularly regarding the nature of the information needed to reconstruct the causal pattern leading to a patient's disease. Consequently, a bio-social account of disease causation influences the role that patients play as active informants, whose testimony is crucial to reconstructing their experience of disease (Malterud et al., 2019).

Under this heading, looking exclusively for biological causes by default constrains the role that a patient's testimony can play in medical decision-making. In this respect, issues in terms of testimonial injustice emerge since patients' testimony can be disregarded as irrelevant to the purpose of unveiling biological causes altogether. Consequently, a form of silencing occurs in which the patient is simply and plainly excluded from playing an active role in the medical discourse.

In contrast, accounts of disease causation that admit the interplay between biological *and* social causes are well positioned in counteracting these forms of epistemic injustices that emerge through the exclusion of patients' testimony by means of informing medical decision-making. Accounts that recognize that social markers play a relevant causal role *must* allow patients to actively contribute to medical discourse. Looking for causally relevant social markers thus requires patients' testimony to be included in medical decision-making. In fact, patients are uniquely positioned to provide healthcare professionals with epistemic access to their social situatedness as a direct means to uncover past events or experiences that play a role in bringing about a problematic health outcome (Rogers, 2002).

Consider, for example, chronic or psychosomatic diseases that are often extremely burdensome to patients, even though they lack an objectively recognizable manifestation at the biological level (Blease et al., 2017). In many cases, patients are disparately affected by these pathologies across different populations. For instance, a chronic disease, such as fibromyalgia, is more likely to affect women than men, i.e., people who are often victims of unjustified credibility deficits (Carel & Kidd, 2014). Since "(t)here is no test to detect fibromyalgia; the disease is diagnosed based on the patient's self-report of symptoms" (Heggen & Berg, 2021, p. 2), in the absence of an evident manifestation of the disease, admitting patients' testimony to the pool of relevant information for this to be included in medical decision-making is paramount to uncovering socially relevant markers that would otherwise go unnoticed. Likewise, it is more probable to dismiss patients' discomfort as "exaggerating pain" or even go as far as doubting that their problematic situation is real because it is not biologically verifiable. Investigating the socially relevant markers that can cause the pathology (e.g., past traumatic/stressful events) is thus even more relevant, considering that women are more likely to suffer from fibromyalgia compared to men. This means that the members of a population group who already have a harder time getting their testimony and credibility acknowledged and taken seriously are disparately affected by this pathology (Zhang et al., 2021).

Hence, we need accounts of disease causation that: 1. support the necessity to include various sources of information pertaining to biological and social markers in disease causation that have been analyzed at the group and individual levels in medical decision-making (i.e., coming from patients' testimony); and 2. do not reduce disease causation to mere biological causes because, otherwise, diseases that are not explainable at the biological level will remain unconsidered altogether. Overall, a bio-social account of disease causation—such as the one advanced by Ghiara and Russo (2019) and discussed in the previous section—can play a relevant role in "legitimizing user

knowledge in decision-making processes" (Grim et al., 2019), thus mitigating forms of testimonial injustice in medical encounters.

Consider again the role of ACEs as socio-markers, as they have been previously discussed. Arguably, the most immediate—or, on occasion, even the only possible—way to access them is through patients' testimony of their lived experiences. Hence, a bio-social account of disease causation is well suited to admit patients' direct contributions to medical discourse because patients' testimony is the most direct and effective way of accessing the socio-markers of disease.

Of course, a bio-social conceptualization of disease causation does not offer a solution to the more structural problems of identity bias and discriminatory practices that give rise to the problem of epistemic injustice. However, *ceteris paribus*, a medical practice that accounts for a patient's testimony as a relevant source of knowledge is more likely to reduce the incidence of cases in which patients' testimony is preemptively excluded from medical decision-making.

7.4. Machine learning systems and social causes

In the previous sections, we argued that a conceptualization of disease causation that properly considers the incidence of social markers is a relevant step toward epistemic justice. Thus far, we have considered medical interactions taking place between patients and physicians, i.e., in the context of exclusive human–human interactions. However, the introduction of technology into medical practice raises new challenges that need to be accounted for. Indeed, as they are embedded in more complex social structures, technological artifacts can reflect and further perpetuate social inequalities. Liao and Carbonell (2022) cogently argue that mechanisms of injustice and oppression can materialize into technological devices used daily by physicians. Referring to the

biases built into commonly used devices, such as oximeters and spirometers, they argue that social differences are often treated as genetic differences, and this "underwrites a bias built into medical tools and technologies that treat the dominant group as the norm and the subordinated groups as deviations." (Liao & Carbonell, 2022, p. 6) This shows that biases and social inequalities can become materialized realities through technology. Recently, these kinds of social and ethical issues have experienced a surge in the field of AI-based systems, such as ML systems.

In the remainder of the chapter, we argue that the use of ML in healthcare brings about a shift back to an investigation of causation in medical practice that, more often than not, ignores considerations of social markers. We maintain that this is the case for many medical ML systems for a number of reasons. Sometimes designers find it easier to operationalize biological markers than social markers simply because the former do not necessarily require interpretation and are not socially constructed, and we already have models of what the average physical body looks like (or should look like) (Van de Poel, 2020). On occasion, identifying the right social markers and potential conflicts among them is not possible at the design stages; therefore, ML systems simply do not include them (Karaca, 2021). Sometimes designers ignore details about social markers relevant to a particular patient by assuming that they homogeneously apply across patients (Obermeyer et al., 2019). For instance, differences in terms of employment conditions, insurance and access to healthcare are applied indistinguishably to a large population, regardless of their particularities or relations to other social markers.[5] Next, we analyze a case where ML implements health costs as the primary data regarding patients' medical expenses, ignoring that this factual information is strongly correlated with racial disparities and job opportunities, among other data.

Under these headings, it seems unrealistic to expect physicians to be cognitively and otherwise able to identify which social markers are absent from

[5]We thank Emanuele Ratti for pointing this out.

the ML system's design. Moreover, even under the assumption that physicians know that a given marker is missing, they are unable to determine the extent to which that social marker affects a patient, how to reassess the ML output using this information, and how to act in accordance with this new information. At best, physicians can reject the output outright, given, for example, obvious conflicts with medical practice. However, they are not in an epistemic position to aggregate the social marker to the ML output as part of an ulterior assessment of the patient's condition. As a result, ML that falls within these situations effectively displaces, by default, the social markers collected through the patient's testimony. We believe that this can ultimately constitute forms of testimonial injustice in at least two respects. First, because markers accessed through patients' testimony tend to be dismissed by ML designers as irrelevant, for the reasons listed before. Second, and as a consequence, patients' testimony is not admitted to the discourse in the first place, and, thus, the ML system plays a considerable role in preemptively silencing patients as epistemic agents.

Interestingly, these automated decision-making systems are typically displayed as conveying objectivity and neutrality in their role of providing (factual) information. To this end, the operationalization of concepts in their socially relevant meaning is often neglected or glossed over. However, and as argued before, opting for concept definitions deflated of socially charged meaning carries the detrimental risk of further promoting the shift toward a biologization of health and disease that does not do justice to complex social markers pervaded by inequalities, injustices and failed mechanisms for the distribution of wealth.

To render these considerations more concrete, consider the study advanced by Obermeyer et al. (2019). These authors scrutinize an algorithmic system widely used in the USA to allocate medical resources more efficiently by identifying patients with significant health needs. This study shows that patients' past health costs have been used as a proxy to determine which patients are in

need of extra care due to complex health conditions. As such, the algorithm highly influences the distribution of crucial medical resources and thus plays a socially consequential role in paramount medical decisions. *Prima facie*, past health costs can be seen as an efficient and allegedly neutral proxy. It is easily quantifiable and refers to factual information regarding patients' medical expenses. However, it disregards socially relevant considerations pertaining to patients' possibility of receiving medical support in the first place due to racial disparities afflicting the US medical system. Under this heading, the ML system reinforces injustices and inequalities. This fact was pointed out by Obermeyer et al. (2019) in their study, in which black people systematically resulted in being wrongly considered less sick and, consequently, less in need of medical attention simply because they had lower medical expenses. Consequently, black people have a lower qualification rate to be eligible to receive crucial medical resources. For example, in a situation in which a black person and a white person are both excluded from receiving extra care, the black person is likely to be considerably more in need of medical support than the white person (Benjamin, 2019).

The fact that black people are usually granted less access to medical care due to the higher frequency of being under-insured and having a lower socioeconomic status, on average, is fully disregarded. Black people have lower medical costs not because they are less in need of medical support but rather because they are prevented from accessing medical services in the first place, which further hinders their access to care when these ML systems make predictions, thus perpetuating social discrimination and distributive injustices. Consequently, they continue to receive less medical attention than they deserve. From the case description, it follows that similar ML systems bring about biased outcomes that have disparate effects on population subgroups. In fact, Obermeyer's case has often been discussed in the literature in the context of the perpetuation of biases and risks of AI (e.g., Aquino et al. (2023) and Benjamin (2019)).

We can add a further perspective on these analyses by highlighting the disparate effects that an ML system can produce when the definition of a complex, abstract and socially loaded concept, such as "health", is operationalized. From the analysis of this case, it emerges that this ML system reduces what it means to be healthy or sick to a fixed metric that does not consider relevant, socially loaded markers (e.g., racial inequalities in access to healthcare, socio-economic circumstances, cultural backgrounds and habits). As a result, social conditions that play a causally relevant role for people with low medical costs are simply missing from this picture.

However, it seems promising to try to circumvent these issues by shedding light on the causal role of social markers in ML-mediated contexts. In fact, as Russo points out, the "concepts of health and disease are not 'causally neutral'. (...) [D]epending on how we conceptualize them, this will impact what causes we are looking for and what actions could/should follow." (Russo, 2023, p. 6) In the case previously considered, ML systems that have empty or unrealistically homogenized socially constructed concepts are utilized. Hence, it should not be a surprise that social inequalities are left unconsidered, even if they are the real causes of why certain population subgroups are sicker than others.

7.5. FINAL REMARKS

The overarching goal of this chapter was to analyze the normative implications arising in connection with different accounts of disease causation. More precisely, we considered discussions in the relevant literature revolving around the bearing of social markers in causing certain pathologies, particularly for individuals belonging to vulnerable populations. We advanced the claim that the causes sought by medical professionals (either only biological or bio-social) in the context of crucial medical interactions, such as diagnostic procedures, are relevant in terms of broader normative implications connected to patients'

epistemic standing. This is the case because, as we argue, the causes that are sought are inextricably connected to the sources of information and knowledge admitted to medical discourse. Concretely, hybrid accounts of disease causation that recognize the importance of identifying social factors in the causal continuum leading to a certain pathology are better positioned to prevent forms of epistemic—and, in particular, testimonial—injustice. Our conclusion in the first part of the chapter was that hybrid accounts of disease causation are an important step toward epistemic justice because they allow room for patients' participation, acknowledging the epistemic value of their testimony.

With the increasing introduction of ML systems in medical care, medical interactions are often no longer exclusive to the human–human context. The fact that these systems mediate crucial medical practices and the patient–physician relationship considerably complicates the picture previously analyzed.

In the final part of the chapter, we provided an—admittedly initial—discussion of the danger of shifting back to the exclusion of relevant social markers from medical practices based on the use of ML systems. Through the consideration of a case study, we aimed to show that relevant social markers often fail to be operationalized. We hinted at the possibility that this represents an obstacle to the meaningful consideration of social causes, thus reintroducing issues in terms of testimonial injustice in ML.

In conclusion, this chapter aimed to provide some initial considerations of the epistemic and ethical implications of different accounts of disease causation along different levels of causation (from group- to individual-level considerations and from biological to social causation). The current analysis provides a further theoretical underpinning to motivate a shift of attention to causally relevant social markers, particularly when ML systems, as socially consequential technologies, are involved in decision-making processes.

8

Conclusion

My main goal with this dissertation has been to propose a novel framework to capture epistemic injustices that arise in connection with the introduction of machine learning (ML) systems in medicine and healthcare. While extensive research efforts into the ethics and epistemology of ML are geared toward addressing issues at the intersection of these two dimensions, such as algorithmic bias, privacy, and responsibility, epistemic injustice has been, so far, neglected. However, dedicating research efforts to understanding the nature of epistemic injustices that arise due to the role of ML systems in medical settings is of utmost importance. The discussions in the previous chapters reveal the wide-ranging nature of the harm they cause and, thereby, highlight the need for timely approaches to identify and counteract ML-induced epistemic injustice. To this end, the main research question (*In which ways does machine learning-induced epistemic injustice in medicine and healthcare emerge, and how can it be mitigated?*) has been answered in two steps, based on which this dissertation is divided into two main parts. The first part aimed to highlight two research gaps that have led to the general neglect of issues of epistemic injustice in the current debate.

In the first part of the dissertation, I explicated how the first research gap is related to the treatment of trust in ML that bypasses these systems' mediating role in patient-physician interactions. This sidelines how physicians' assessment of patients' credibility is often determined by ML systems that mediate medical interactions. Even though these systems are meant to be of assistance to medical professionals in performing central medical tasks (such as diagnoses and treatment recommendations), in practice, many ML systems are displacing them from their epistemically authoritative role. As I argued in Chapter 4, this analysis is crucial to identifying ML-induced testimonial injustice and exposing how these systems weaken patients' epistemic standing. The second gap is connected to a general tendency in current discussions to conceive of the ethics and epistemology of ML as compartmentalized dimensions. I defended the claim that this approach is problematic because it has led to a consideration of the epistemology of ML as emptied of relevant normative elements related to patients' personal values. As a consequence, forms of epistemic objectification emerge at the patients' expense. I showed that the epistemic objectification experienced by patients manifests in ways that cannot be captured by the standard framework of epistemic injustice. The problematic nature of these research gaps and the need to fill them to ensure a more comprehensive analysis of the ethics and epistemology of ML motivated my inquiry into *ML-induced epistemic injustice* and, thus, expansion of the original framework of epistemic injustice.

The second part of the dissertation has been dedicated in its entirety to identifying and mitigating different forms of epistemic injustice in medical ML. This has been done through the examination of various cases. As specified in Chapter 1, an analysis of epistemic injustice is always context-specific, so ensuring that my approach is case-based has been a way to avoid abstract considerations that are not relatable to the lived experiences and social situatedness of the individuals affected by the injustices studied in this dissertation.

In the following sections, I briefly summarize the key findings ensuing from the study I put forward in the previous chapters. Subsequently, I conclude this dissertation by pointing out aspects of my analysis that require further attention. Finally, I lay out perspectives for future research based on the framework I have proposed.

8.1. KEY FINDINGS

Five key findings emerge from this dissertation. The first is a philosophically informed and systematic analysis of trustworthy AI. This analysis has been fundamental for the overall project of the dissertation. This is because trust is a central concept at the intersection of the ethics and epistemology of AI and is essential to a thorough study of epistemic injustice, as I have demonstrated in the previous chapters. While the literature on this topic is vast and fragmented, Chapter 2 provided a structured way of conceiving trust in AI systems by setting the conditions for trustworthy AI. This chapter also showed that while the question of what trust in ML systems directly amounts to is prominent in the current debate, the question of how AI systems can negatively affect trust relationships between patients and physicians is sidelined. Showing the need to address this research gap was key to tackling trust dysfunctions occurring due to AI in medicine that are conducive to epistemic injustice (see the analysis presented in Chapter 4).

The second key finding emerges from Chapter 3, which challenged the logic underlying the relationship between the ethics and epistemology of ML in the current debate and showed that these two dimensions are often considered in a compartmentalized way. Particularly, it has been pointed out that ethically relevant elements are not considered in terms of their regulatory role of central epistemic functions (e.g., explanations). After making it explicit that this is the default approach adopted in current discussions, I went a step further

in this chapter to depict its shortcomings. Notably, I showed that an epistemological assessment of ML that is considered to be fixed and unmodifiable by information relevant to a patient's epistemic and moral values can lead to their epistemic objectification. Importantly, accounting for the specific ways in which epistemic objectification manifests in medical ML represented the first step in the development of this framework specifically tailored to ML systems.

With regard to the third key finding, this dissertation has provided the first systematic analysis of different forms of epistemic injustice, specifically in ML-mediated medical practices. Based on the considerations presented in Chapters 4 and 5, a framework emerges through which the characteristics of testimonial and hermeneutical injustice brought about by ML in medicine are presented. More specifically, the locus of ML-induced testimonial injustice has been defined as the risk that ML systems become the main markers of trustworthiness in situations in which the assessment of patients' credibility is at stake. The advanced analysis serves as a guideline to identify these forms of injustice and can be applied to the assessment of other cases. With regard to hermeneutical injustice, the three conditions outlined in Chapter 5 dissected this phenomenon according to its main manifestations and provided a systematic way for individuating this form of injustice in medical ML. The concept of *automated hermeneutical appropriation* that I advanced effectively shows that these forms of injustice have more far-reaching and difficult-to-control consequences in ML-mediating scenarios than in human-human settings. On a more general level, the research put forward in these chapters illustrates the societal relevance of a systematic analysis of epistemic injustice that needs to be treated in its own right. Issues of epistemic injustice could be mistakenly conflated with issues of bias and discrimination. However, as became clear throughout this dissertation, epistemic injustice is not reducible to these issues and deserves separate treatment.

The fourth key finding pertains to the analysis of participatory injustice presented in Chapter 6. This chapter sheds light on epistemological and ethical issues that mental health chatbots can bring about, specifically among generally under-served populations. The analysis of these population subgroups' specific circumstances in connection with the use of AI technologies is often under-researched. Therefore, this chapter contributed to addressing a significant and socially relevant research gap. Moreover, in this chapter, I examined how to mitigate participatory injustice related to conversational AI (CAI) technologies for mental health support. This has been done through the novel approach of framing *epistemic participation* as a capability that should be accounted for in the design of CAI systems. I showed that including epistemic participation as a capability in the Capability Sensitive Design framework can lead to the mitigation of participatory injustice.

The fifth key finding emerges from Chapter 7. In this chapter, I provided arguments for bringing together debates in social epistemology with debates in the philosophy of science on causality that are usually independent of each other. I have shown that considering the intersection of these debates is a crucial step to effectively grasp the influence that patients' social situatedness has on their health outcomes. Moreover, it is also fundamental to highlight that causal questions underlying medical practice are relevant to patients' epistemic participation and, consequently, their risk of being the victims of epistemic injustice. Considered positively, the arguments advanced provide reasons to include social markers of disease *explicitly* in the development of AI technologies in medicine. While this point is of great importance to achieve epistemic justice, these markers often fail to be operationalized in ML systems. This chapter provided reasons as to why this should be of high priority in the design of ML systems in medicine.

8.2. Prospects for future research

As pointed out in Chapter 1, this work was, first and foremost, geared towards a clear conceptualization of ML-induced epistemic injustice. The focus was, thus, on two related goals. First, I aimed to show that these forms of injustice can arise in healthcare practice in which ML systems play a role in supporting medical decision-making. This adds a further dimension to debates on the ethics of AI that have been, so far, focused on the widely discussed issues of privacy, bias, discrimination, and responsibility, among others. Second, I aimed to provide the framework needed to identify these forms of injustice since they can easily go unnoticed. In doing so, I expanded the standard framework of epistemic injustice, which is not comprehensive enough to account for the role of ML systems as epistemically powerful technologies. Therefore, my overarching goal was to show that there *is* an epistemic and moral problem that is currently overlooked and to analyze its nature and the different forms it can assume, specifically in medicine and healthcare. As mentioned in the previous section, I also presented some considerations related to how such forms of ML-induced epistemic injustice can be mitigated.

However, starting from the analysis I provided in this dissertation, more needs to be said on how to achieve epistemic justice in medical ML. That is, more research work is needed to translate my framework into operationalizable design requirements. Although this dissertation abounds in cases where my framework is applied, case-sensitive considerations should be included at the design stage of the development of ML technologies to mitigate or avoid altogether the issues outlined in this dissertation. More precisely, reflections pertaining to avoiding ML-induced epistemic injustice should be included in design methodologies that view the epistemological and ethical assessment of AI systems not as an afterthought but, rather, as a constitutive part of technological advancement aimed at ensuring the societal acceptability of ML

systems. For example, my framework can be incorporated in approaches such as Value Sensitive Design, in which the conceptual specification of fundamental principles, such as justice, acquires a central stage. Overall, my analysis underpins the need to design for justice, specifically in terms of its epistemic nature.

Initial considerations gearing toward this research direction have been presented in Chapter 6, in which epistemic participation has been framed as a capability to be incorporated in the Capability Sensitive Design framework. However, this analysis was limited to the conceptual level. Like Value Sensitive Design, the Capability Sensitive Design approach follows a tripartite division in the form of iterative conceptual, empirical, and technical investigations. Future research efforts taking into account empirical studies on the needs of the stakeholders directly and indirectly affected by a medical ML system are central to preemptively evaluating and mitigating concerns related to ML-induced epistemic injustice.

These considerations show that efforts toward the mitigation and anticipation of ML-induced epistemic injustice require interdisciplinary collaborations. Such efforts should involve theoretical analyses to specify the nature of the injustice, as specified in my framework, and empirical studies to understand the needs and vulnerabilities, together with the social identities of the populations affected by the technology. This is especially relevant if vulnerable populations are the target of a particular technology, as in the case of the chatbot Karim that has been analyzed in Chapter 6.

Another aspect of this study that needs more attention pertains to the possibility of generalizing the analysis of ML-induced epistemic injustice beyond the field of medicine and healthcare as it has been developed in this dissertation. In fact, my framework is tailored to the medical domain, in which the moral salience of patient-physician interactions needs special scrutiny, together with a well-defined and context-specific series of moral norms and values. The application of my considerations and concepts (such as the concept

of *automated hermeneutical appropriation* I advanced in Chapter 5) beyond the contexts of medicine and healthcare might need further specification or adaptation. Future research efforts are needed to apply my framework to high-stakes fields beyond healthcare in which ML systems pose a considerable risk to individuals' epistemic standing. This holds particularly true, for example, for judicial cases in which ML systems could undermine the legitimacy of individuals' testimony and bring about potentially profoundly unjust life-determining decisions. The often-discussed case of the COMPAS algorithm used in US courts to predict defendants' likelihood to re-offend is a case in point (Dieterich et al., 2016). My framework could be adapted to evaluate and preemptively counteract ML-induced epistemic injustices emerging in these ethically salient contexts.

In conclusion, the analysis I provided represents an important step toward greater awareness of epistemological, ethical, and socially relevant problems connected to the use of ML systems in the highly sensitive fields of medicine and healthcare. I hope that, through this dissertation, I have compellingly demonstrated that these issues need timely attention. Importantly, because they manifest in a rather subtle manner and affect mostly disadvantaged and under-represented social groups, these issues should be brought to the foreground of debates on the ethics and epistemology of AI.

A

Appendix

A.1. Further remarks on testimonial injustice in medical machine learning: a response to commentaries[1]

A.1.1. Introduction

In my paper entitled 'Testimonial injustice in medical machine learning' (Pozzi, 2023b), I argued that machine learning (ML)-based Prediction Drug Monitoring Programmes (PDMPs) could infringe on patients' epistemic and moral standing inflicting a testimonial injustice (Fricker, 2007). I am very grateful for all the comments the paper received, some of which expand on it while others take a more critical view. This response addresses two objections raised to my consideration of ML-induced testimonial injustice in order to clarify the po-

[1] This appendix is based on the following publication:

Pozzi, G. (2023). Further remarks on testimonial injustice in medical machine learning: a response to commentaries. *Journal of Medical Ethics*, 49: 551-552.

sition taken in the paper. The first maintains that my critical stance toward ML-based PDMPs idealises standard medical practice. Moreover, it claims that the ML-induced testimonial injustice I discuss is not substantially different from situations in which it emerges in human–human interactions. The second claims that my analysis does not establish a link to issues of automation bias, even if these are to be considered the core of testimonial injustice in ML. In the following, I address each objection in turn.

A.1.2. A misguided equivalence

Gillett (2023) argues that my critical stance towards using risk prediction tools such as PDMPs implies the idealisation of standard (i.e., non-ML-mediated) modes of clinical practice. Considering certain uses of ML in a different setting, that is, psychiatry, the author goes as far as claiming that 'traditional models of clinical practice in psychiatry are far from a utopia, free from epistemic injustice, which Pozzi's argument risks proposing'. Since this statement does not represent what I intend to suggest, I am glad to have the possibility to clarify this point.

It is not a matter of contention that standard models of medical practice are fraught with epistemic injustices. In my paper, I reconstruct Fricker's account, following authors who apply it to the field of medicine and healthcare (Kidd & Carel, 2017). Here, it becomes clear that testimonial injustice can occur in healthcare encounters due to epistemically unjustified identity prejudices. For example, I consider empirical studies to substantiate the claim that racial prejudices are often at the root of misguided beliefs conducive to testimonial injustices in pain management (Trawalter et al., 2012). These concerns refer to traditional models of clinical practice. It is thus absolutely uncontroversial that episodes of testimonial injustice are recurrent in human–human interactions in clinical care.

More interesting is a discussion about what makes an ML-induced testimonial injustice different from one in standard medical practice. Gillett equates the two, seemingly denying that there is something essentially unique to ML-induced testimonial injustices. I maintain that this equivalence is misguided for at least two reasons.

The first is a point I briefly touched on in the paper, but that is worth restating, that is, the scale of propagation of the epistemic harm caused by PDMP risk scores that do not adequately represent a patient's clinical situation. As I pointed out, these systems disproportionately miscategorize patients from under-represented or stigmatized social categories. Thus, they reinforce and further propagate existing inequalities (Pozzi, 2023a, 2023b). This does not mean that the solution to these problems is to be found at the individual level. However, we need to be aware that epistemic harms are systematized through ML-based PDMPs, strongly constraining patients' possibilities of recourse if they were unfairly attributed a high-risk score. So, not only is the harm these systems cause more difficult to identify, but also the possibility that patients have of counteracting it, thus reaffirming their epistemic agency, is very much constrained.

The second point relates to clinicians' possibilities to evaluate the trustworthiness of a patient's testimony in ML-mediated medical practices. The possibility of meaningfully making this assessment is paramount to avoid a potentially premature dismissal of a patient's testimony in the face of a red flag outputted by the ML-based PDMP. However, this requires considering system-independent epistemic resources and being able to use these to motivate a medical decision that contradicts the system's risk score. The current use of PDMPs discussed in my paper does not seem to allow this fundamental activity, thus laying the ground for pervading testimonial injustices.

Moreover, these are not arguments against ML in medicine altogether. The example of ML in psychiatry mentioned by the author to enhance patients' understanding of their condition could have beneficial effects. However, this is

a very different use of ML than the one I analyze, in which ML scores represent the most immediate base for physicians' decision-making regarding opioid prescriptions. Ultimately, my analysis should contribute to efforts discouraging the attempt to advance seemingly rapid technological solutions to deeply rooted societal issues like the opioid crisis in the USA. But at no point do I intend to discourage the general use of ML in clinical practice.

In conclusion, Gillet and I agree that medical ML should be explicitly designed for epistemic justice. Further research efforts heading in this direction are more needed than ever.

A.1.3. More than automation bias

Nguyen (2023) claims that my analysis of ML-induced testimonial injustice fails to establish a link to the problem of automation bias. As I explicitly recognized in my paper, automation bias is crucial in bringing about testimonial injustice in an ML-mediated context. However, I refrained from making it the central piece of my examination to avoid obscuring other equally important factors that impair patients' credibility assessment.

I maintain that even though automation bias is an important element in the occurrence of ML-mediated testimonial injustices, the latter is not to be reduced to the former. Automation bias is a psychological attitude of healthcare practitioners towards automated systems that needs to be corrected through appropriate training, as Nguyen rightly points out, particularly by clarifying the limitations of ML systems. However, it is crucial to highlight that in the PDMP case, even for a physician with a critical attitude towards the output of ML-based PDMP, it would be extremely hard not to follow through with the system's recommendation. This is for reasons that exceed issues of automation bias, and I elaborate on them in the following.

As pointed out in the paper, physicians are legally required to observe PDMP recommendations (Oliva, 2022). The higher-level setup does not allow

for a meaningfully critical approach toward ML-based PDMPs, and consequently, patients are likely to be withdrawn credibility in the face of a high-risk score. The broader context in which these systems' use is regulated thus leads physicians to consider patients' risk score as a marker of trustworthiness rather than their testimony. Moreover, there are constrained possibilities to assess the epistemic value of a risk score outputted and evaluate whether it reliably depicts a patient's clinical condition and need for opioid medication if the risk score contradicts the patient's testimony. These considerations show that the risk of bringing about testimonial injustices through the current use of PDMPs is present beyond the crucial problem of automation bias.

Nguyen's conclusion that these systems can cause not only epistemic but also physical harm undoubtedly deserves attention. From the case of Kathryn discussed in the paper, it is evident that these systems can also cause tangible and real physical harm. I decided to highlight epistemic harms as these tend to go unnoticed. Nevertheless, their analysis is central to preserving patients' epistemic standing.

Bibliography

Alfano, M., & Huijts, N. (2019). Trust in Institutions and Governance. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 256–270). Routledge.

Alpaydin, E. (2014). *Introduction to Machine Learning*. Massachusetts Institute of Technology.

Alvarado, R. (2023). AI as an Epistemic Technology. *Science and Engineering Ethics*, *29*(5). https://doi.org/10.1007/s11948-023-00451-3

Anderson, E. (2012). Epistemic Justice as a Virtue of Social Institutions. *Social Epistemology*, *26*(2), 163–173. https://doi.org/10.1080/02691728.2011.652211

Aquino, Y. S. J., Carter, S. M., Houssami, N., Braunack-Mayer, A., et al. (2023). Practical, epistemic and normative implications of algorithmic bias in healthcare artificial intelligence: a qualitative study of multidisciplinary expert perspectives. *Journal of Medical Ethics*. https://doi.org/10.1136/jme-2022-108850

Ashfaq, A., Esmaili, S., Najjar, M., Batool, F., et al. (2020). Utilization of Mobile Mental Health Services among Syrian Refugees and Other Vulnerable Arab Populations—A Systematic Review. *International Journal of Environmental Research and Public Health*, *17*(4). https://doi.org/10.3390/ijerph17041295

Babushkina, D., & Votsis, A. (2022). Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. *Ethics and Information Technology*, *24*(2), 22. https://doi.org/10.1007/s10676-022-09629-y

Beisbart, C. (2021). Opacity thought through: On the intransparency of computer simulations. *Synthese*, *199*, 11643–11666.

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA - Journal of the American Medical Association*, *318*(22), 2199–2210. https://doi.org/10.1001/jama.2017.14585

Benjamin, R. (2019). Assessing risk, automating racism. *Science*, *366*(6464), 421–422. https://doi.org/10.1126/science.aaz3873

Bjerring, J. C., & Busch, J. (2021). Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy & Technology*, *34*, 349–371. https://doi.org/https://doi.org/10.1007/s13347-019-00391-6

Blease, C., Carel, H., & Geraghty, K. (2017). Epistemic injustice in healthcare encounters: Evidence from chronic fatigue syndrome. *Journal of Medical Ethics*, *43*(8), 549–557. https://doi.org/10.1136/medethics-2016-103691

Bleher, H., & Braun, M. (2023). Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice. *Science and Engineering Ethics*, *29*(3). https://doi.org/10.1007/s11948-023-00443-3

Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, *32*(1), 43–75. https://doi.org/10.1007/s11023-021-09569-4

Bryer, E., & Henry, D. (2018). Chemotherapy-induced anemia: Etiology, pathophysiology, and implications for contemporary practice. *International Journal of Clinical Transfusion Medicine*, *6*, 21–31.

Bryson, J. (2018). AI & Global Governance: No One Should Trust AI [Accessed: 12-12-2023]. *UNU-CPR (blog)*. https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html

Buchman, D. Z., Ho, A., & Goldberg, D. S. (2017). Investigating Trust, Expertise, and Epistemic Injustice in Chronic Pain. *Journal of Bioethical Inquiry*, *14*(1), 31–42. https://doi.org/10.1007/s11673-016-9761-x

Bugel, S. (2023). Fake doctor who worked in NHS for 20 years found guilty of fraud [Accessed: 10-12-2023]. *The Guardian*. https://www.theguardian.com/uk-news/2023/feb/15/fake-doctor-zholia-alemi-nhs-guilty-fraud

Buijsman, S. (2022). Defining Explanation and Explanatory Depth in XAI. *Minds and Machines*, *32*, 563–584. https://doi.org/10.1007/s11023-022-09607-9

Bulloch, M. (2018). The Evolution of the PDMP [Accessed: 25-10-2022]. *Pharmacy Times*. https://www.pharmacytimes.com/view/the-evolution-of-the-pdmp

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1). https://doi.org/10.1177/2053951715622512

Byskov, M. F. (2021). What Makes Epistemic Injustice an "Injustice"? *Journal of Social Philosophy*, *52*(1), 114–131. https://doi.org/10.1111/josp.12348

Carel, H., & Kidd, I. J. (2014). Epistemic injustice in healthcare: a philosophial analysis. *Medicine, Health Care and Philosophy*, *17*(4), 529–540. https://doi.org/10.1007/s11019-014-9560-2

Carel, H., & Kidd, I. J. (2017). Epistemic injustice in medicine and healthcare. In I. J. Kidd, J. Medina, & G. J. Pohlhaus (Eds.), *The Routledge Handbook of Epistemic Injustice* (pp. 336–346). Routledge.

Chen, M. (2021). Trust and Trust-Engineering in Artificial Intelligence Research: Theory and Praxis. *Philosophy & Technology*, *34*(4), 1429–1447. https://doi.org/10.1007/s13347-021-00465-4

Cho, J. H., Xu, S., Hurley, P. M., Mackay, M., et al. (2019). STRAM: Measuring the trustworthiness of computer-based systems. *ACM Computing Surveys*, *51*(6). https://doi.org/10.1145/3277666

Choung, H., David, P., & Ross, A. (2022). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction*, 1–13.

Chung, R. (2021). Structural health vulnerability: Health inequalities, structural and epistemic injustice. *Journal of Social Philosophy*, 1–16. https://doi.org/10.1111/josp.12393

Claassen, R. (2011). Making Capability Lists: Philosophy versus Democracy. *Political Studies*, *59*(3), 491–508. https://doi.org/10.1111/j.1467-9248.2010.00862.x

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1–17.

Craig, E. (1990). *Knowledge and the state of nature: An essay in conceptual synthesis*. Clarendon Press.

Creel, K. A. (2020). Transparency in Complex Computational Systems. *Philosophy of Science*, *87*(4), 568–589. https://doi.org/10.1086/709729

Curry, T. J. (2020). Conditioned for Death: Analysing Black Mortalities from Covid-19 and Police Killings in the United States as a Syndemic Interaction. *Comparative American Studies*, *17*(3-4), 257–270. https://doi.org/10.1080/14775700.2021.1896422

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE symposium on security and privacy (SP)*, 598–617.

De Proost, M., & Pozzi, G. (2023). Conversational Artificial Intelligence and the Potential for Epistemic Injustice. *The American Journal of Bioethics*, *23*(5), 51–53. https://doi.org/10.1080/15265161.2023.2191020

Decamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *The Lancet Digital Health*, *1*(8), e390.

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy and Technology*, *31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

Demir-Doğuoğlu, H., & McLeod, C. (2023). Towards a Feminist Theory of Distrust. In M. Alfano, D. Collins, & I. V. Javonovic (Eds.), *The moral psychology of trust*. https://works.bepress.com/carolyn-mcleod/62/

Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, *7*(4).

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608

Dotson, K. (2011). Tracking Epistemic Violence, Tracking Practices of Silencing. *Hypatia*, *26*(2), 236–257. https://www.jstor.org/stable/23016544?seq=1&cid=pdf-

Douglas, H. E. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Durán, J. M. (2018). *Computer Simulations in Science and Engineering: Concepts - Practices - Perspectives*. Springer.

Durán, J. M. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence*, *297*. https://doi.org/10.1016/j.artint.2021.103498

Durán, J. M., & Formanek, N. (2018). Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines*, *28*(4), 645–666. https://doi.org/10.1007/s11023-018-9481-6

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329–335. https://doi.org/10.1136/medethics-2020-106820

Durán, J. M. (manuscript). *Machine learning, justification, and computational reliabilism*. https://philarchive.org/rec/DURMLJ

Elgin, C. Z. (1996). *Considered judgement*. Princeton University Press.

Emanuel, E. J., & Emanuel, L. L. (1992). Four Models of the Physician-Patient Relationship. *Jama*, 2221–2226. https://jamanetwork.com/journals/jama/article-abstract/396718

Epstein, R. M., & Gramling, R. E. (2013). What Is Shared in Shared Decision Making? Complex Decisions When the Evidence Is Unclear. *Medical Care Research and Review*, *70*(1), 94S–112S. https://doi.org/10.1177/1077558712459216

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

European Commission. (2019). High-level expert group on artificial intelligence. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Favalli, E. G., Biggioggero, M., Maioli, G., & Caporali, R. (2020). Baricitinib for covid-19: A suitable treatment? *The Lancet Infectious Diseases*, *20*(9), 1012–1013.

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI We Trust Incrementally: A Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philosophy & Technology*, *33*(3), 523–539. https://doi.org/10.1007/s13347-019-00378-3

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: It is possible to trust medical AI. *Journal of Medical Ethics*, *47*(6), 437–438. https://doi.org/10.1136/medethics-2020-106922

Field Innovation Team. (2016). *Lebanon After Action Report* (tech. rep.). http://fieldinnovationteam.org/wp-content/uploads/2014/09/Lebanon-After-Action-Report-2016.pdf

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Fricker, M. (2010). Replies to Alcoff, Goldberg, and Hookway on Epistemic Injustice. *Episteme*, *7*(2), 164–178.

Fricker, M. (2015). Epistemic Contribution as a Central Human Capability. In G. Hull (Ed.), *The equal society: Essays on equality in theory and practice* (pp. 73–90). Lexington Books.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. Van de Poel, & M. E. Gorman (Eds.), *Philosophy of engineering and technology* (pp. 55–95). Springer Nature.

Frost-Arnold, K. (2020). Trust and epistemic responsibility. In J. Simon (Ed.), *The routledge handbook of trust and philosophy* (pp. 64–75). Routledge.

Gabriel, I. (2022). Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, *151*(2), 218–231. https://doi.org/10.1162/daed_a_01911

Gaffney, A. (2021). Racism and Respiration [accessed: 02.11.2023]. *Boston Review*. https://www.bostonreview.net/articles/adam-gaffney-racism-and-respiration/

Ghiara, V., & Russo, F. (2019). Reconstructing the mixed mechanisms of health: the role of bio-and socio-markers. *Longitudinal and Life Course Studies*, *10*(1), 7–25.

Gillett, G. (2023). Testimonial injustice in medical machine learning: a perspective from psychiatry. *Journal of Medical Ethics*, 2023–109059. https://doi.org/10.1136/jme-2023-109059

Giovanola, B., & Tiribelli, S. (2022). Weapons of moral construction? On the value of fairness in algorithmic decision-making. *Ethics and Information Technology*, *24*(1). https://doi.org/10.1007/s10676-022-09622-5

Goldberg, S. C. (2020). Trust and Reliance. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 97–108). Routledge.

Golden, J. A. (2017). Deep Learning Algorithms for Detection of Lymph Node Metastases From Breast Cancer. Helping Artificial Intelligence Be Seen. *JAMA - Journal of the American Medical Association*, *318*(22), 2184–2186. https://doi.org/10.1001/jama.2017.14580

Goldman, A. I. (2019). The What, Why, and How of Social Epistemology. In M. Fricker, P. J. Graham, D. Henderson, & N. J. L. L. Petersen (Eds.), *The Routledge Handbook of Social Epistemology* (pp. 10–20). Routledge.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Graber-Stiehl, I. (2023). Is the world ready for ChatGPT therapists? *Nature*, *617*(7959), 22–24.

Grasswick, H. (2018). Understanding Epistemic Trust Injustices and Their Harms. *Royal Institute of Philosophy Supplement*, *84*, 69–91. https://doi.org/10.1017/s1358246118000553

Greene, C. (2019). Big Data and the Reference Class Problem. What Can We Legitimately Infer about Individuals? *Computer Ethics-Philosophical Enquiry (CEPE) Proceedings*, *2019*(1), 7.

Grim, K., Tistad, M., Schön, U. K., & Rosenberg, D. (2019). The Legitimacy of User Knowledge in Decision-Making Processes in Mental Health Care: An Analysis of Epistemic Injustice. *Journal of Psychosocial Rehabilitation and Mental Health*, *6*(2), 157–173. https://doi.org/10.1007/s40737-019-00145-9

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., et al. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Haines, S., Lam, A., Savic, M., & Carter, A. (2022). Patient experiences of prescription drug monitoring programs: a qualitative analysis from an Australian pharmaceutical helpline. *International Journal of Drug Policy*, *109*, 1–9. https://doi.org/10.1016/j.drugpo.2022.103847

Haines, S., Savic, M., Nielsen, S., & Carter, A. (2022). Key considerations for the implementation of clinically focused Prescription Drug Monitoring Programs to avoid unintended consequences. *International Journal of Drug Policy*, *101*, 1–5. https://doi.org/10.1016/j.drugpo.2021.103549

Hao, K. (2020). Doctors are using AI to triage covid-19 patients. the tools may be here to stay [accessed: 10-07-2023]. *MIT Technology Review*, *27*. https://www.technologyreview.com/2020/04/23/1000410/ai-triage-covid-19-patients-health-care/

Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation.

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics*, *46*(7), 478–481. https://doi.org/10.1136/medethics-2019-105935

Haupt, C. E., & Marks, M. (2023). AI-generated medical advice - GPT and beyond. *Jama*, *329*(16), 1349–1350.

Hawley, K. (2014). Trust, Distrust and Commitment. *Noûs*, *48*(1), 1–20. https://doi.org/10.1111/nous.12000

Hawley, K. (2015). Trust and distrust between patient and doctor. *Journal of evaluation in clinical practice*, *21*(5), 798–801. https://doi.org/10.1111/jep.12374

Hawley, K. (2017). Trust, Distrust and Epistemic Injustice. In I. J. Kidd, J. Medina, & G. Pohlhaus (Eds.), *Routledge Handbook of Epistemic Injustice* (pp. 69–78). Routledge.

Hawley, K. (2019). *How To Be Trustworthy*. Oxford University Press.

Heaven, W. D. (2021). Hundreds of AI tools have been built to catch covid. None of them helped. [accessed: 10.12.2023]. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

Heggen, K. M., & Berg, H. (2021). Epistemic injustice in the age of evidence-based practice: The case of fibromyalgia. *Humanities and Social Sciences Communications*, *8*(1), 1–6. https://doi.org/10.1057/s41599-021-00918-3

Henderson, G. E. (2022). What Bioethicists Need to Know About the Social Determinants of Health - and Why. *Perspectives in Biology and Medicine*, *65*(4), 664–671. https://doi.org/10.1353/pbm.2022.0058

Hildebran, C., Cohen, D. J., Irvine, J. M., Foley, C., et al. (2014). How Clinicians Use Prescription Drug Monitoring Programs: A Qualitative Inquiry. *Pain Medicine*. https://academic.oup.com/painmedicine/article/15/7/1179/1878292

Hildebran, C., Leichtling, G., Irvine, J. M., Cohen, D. J., et al. (2016). Clinical Styles and Practice Policies: Influence on Communication with Patients Regarding Worrisome Prescription Drug Monitoring Program Data. *Pain Medicine*, *17*, 2061–2066. https://doi.org/10.1093/pm/pnw019

Hookway, C. (2010). Some Varieties of Epistemic Injustice: Reflections on Fricker. *Episteme*, 151–163.

Humphreys, P. W. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Humphreys, P. W. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2

Illari, P., & Russo, F. (2014). *Causality: Philosophical Theory Meets Scientific Practice*. OUP Oxford.

Illari, P., & Russo, F. (2016). Information Channels and Biomarkers of Disease. *Topoi*, *35*(1), 175–190. https://doi.org/10.1007/s11245-013-9228-1

Jacobs, N. (2020). Capability Sensitive Design for Health and Wellbeing Technologies. *Science and Engineering Ethics*, *26*(6), 3363–3391. https://doi.org/10.1007/s11948-020-00275-5

Jacobs, N., & Huldtgren, A. (2018). Why value sensitive design needs ethical commitments. *Ethics and Information Technology*, *23*(1), 23–26. https://doi.org/10.1007/s10676-018-9467-3

Jones, K. (1996). Trust as an Affective Attitude. *Ethics*, *107*(1), 4–25.

Karaca, K. (2021). Values and inductive risk in machine learning modelling: the case of binary classification models. *European Journal for Philosophy of Science*, *11*(4). https://doi.org/10.1007/s13194-021-00405-1

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR)*, *55*(2), 1–38.

Kelly, M. P., Kelly, R. S., & Russo, F. (2015). The Integration of Social, Behavioral, and Biological Mechanisms in Models of Pathogenesis. *Perspectives in Biology and Medicine*, *57*(3), 308–328. https://doi.org/10.1353/pbm.2014.0026

Kelly, M. P., & Russo, F. (2018). Causal narratives in public health: the difference between mechanisms of aetiology and mechanisms of prevention in non-communicable diseases. *Sociology of Health and Illness*, *40*(1), 82–99. https://doi.org/10.1111/1467-9566.12621

Kidd, I. J., & Carel, H. (2017). Epistemic Injustice and Illness. *Journal of Applied Philosophy*, *34*(2), 172–190. https://doi.org/10.1111/japp.12172

Kidd, I. J., Medina, J., & Pohlhaus, G. (2017). Introduction to The Routledge Handbook of Epistemic Injustice. In *The Routledge Handbook of Epistemic Injustice* (pp. 1–9). Routledge.

Kidd, I. J., Spencer, L., & Carel, H. (2022). Epistemic injustice in psychiatric research and practice. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2022.2156333

Knox, B., Christoffersen, P., Leggitt, K., Woodruff, Z., et al. (2023). Justice, Vulnerable Populations, and the Use of Conversational AI in Psychotherapy. https://doi.org/10.1080/15265161.2023.2191040

Latour, B. (2000). The Berlin Key or How to Do Words With Things. In Graves-Brown (Ed.), *Matter, Materiality and Modern Culture* (pp. 10–21).

Lawrence, D. J. (2007). The Four Principles of Biomedical Ethics: A Foundation for Current Bioethical Debate. *Journal of Chiropractic Humanities*, *14*, 34–40. https://doi.org/10.1016/S1556-3499(13)60161-8

Le Bui, M., & Noble, S. U. (2020). We're missing a moral framework of justice in artificial intelligence. On the limits, failings, and ethics of fairness. In M. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 163–179). Oxford University Press Oxford.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50–80.

Lehrer, K. (2006). Testimony and Trustworthiness. In J. Lackey & E. Sosa (Eds.), *The Epistemology of Testimony* (pp. 145–159). Oxford University Press.

Leichtling, G. J., Irvine, J. M., Hildebran, C., Cohen, D. J., et al. (2017). Clinicians' Use of Prescription Drug Monitoring Programs in Clinical

Practice and Decision-Making. *Pain Medicine*, *18*(6), 1063–1069. https://doi.org/10.1093/pm/pnw251

Lewis, P. R., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, 33–49. https://doi.org/10.1016/j.cogsys.2021.11.001

Li, B., Qi, P., Liu, B., Di, S., et al. (2023). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3555803

Liao, S., & Carbonell, V. (2022). Materialized Oppression in Medical Tools and Technologies. *American Journal of Bioethics*, 9–23. https://doi.org/10.1080/15265161.2022.2044543

Lim, C.-P., Chen, Y.-W., Vaidya, A., Mahorkar, C., et al. (2022). *Handbook of Artificial Intelligence in Healthcare Vol 2: Practicalities and Prospects*. Springer.

Link, B. G., & Phelan, J. (1995). Social Conditions as Fundamental Causes of Disease. *Journal of Health and Social Behavior*, 80–94.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Lloyd, E. P., Paganini, G. A., & ten Brinke, L. (2020). Gender Stereotypes Explain Disparities in Pain Care and Inform Equitable Policies. *Policy Insights from the Behavioral and Brain Sciences*, *7*(2), 198–204. https://doi.org/10.1177/2372732220942894

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1).

Longino, H. E. (2004). How Values Can Be Good for Science. In P. Machamer & G. Wolters (Eds.), *Science, Values, and Objectivity* (pp. 127–142). University of Pittsburgh Press Pittsburgh.

Luxton, D. D. (2020). Ethical implications of conversational agents in global public health. *Bulletin of the World Health Organization*, *98*(4), 285.

Luxton, D. D., McCann, R. A., Bush, N. E., Mishkind, M. C., et al. (2011). mHealth for mental health: Integrating smartphone technology in behavioral healthcare. *Professional Psychology: Research and Practice*, *42*(6), 505–512. https://doi.org/10.1037/a0024485

Madianou, M. (2021). Nonhuman humanitarianism: when 'AI for good' can be harmful. *Information Communication and Society*, *24*(6), 850–868. https://doi.org/10.1080/1369118X.2021.1909100

Malterud, K., Guassora, A. D., Graungaard, A. H., & Reventlow, S. (2015). Understanding medical symptoms: a conceptual review and analysis. *Theoretical Medicine and Bioethics*, *36*(6), 411–424. https://doi.org/10.1007/s11017-015-9347-3

Malterud, K., Reventlow, S., & Guassora, A. D. (2019). Diagnostic knowing in general practice: interpretative action and reflexivity. *Scandinavian Journal of Primary Health Care*, *37*(4), 393–401. https://doi.org/10.1080/02813432.2019.1663592

Malterud, K. (2000). Symptoms as a Source of Medical Knowledge: Understanding Medically Unexplained Disorders in Women. *Family Medicine*, *32*(9), 603–611.

Manders-Huits, N. (2011). What Values in Design? The Challenge of Incorporating Moral Values into Design. *Science and Engineering Ethics*, *17*(2), 271–287. https://doi.org/10.1007/s11948-010-9198-2

Manríquez Roa, T., Biller-Andorno, N., & Trachsel, M. (2019). The Ethics of Artificial Intelligence in Psychotherapy. In M. Trachsel, J. Gaab, N. Biller-Adorno, & S. Tekin (Eds.), *The Oxford Handbook of Psychotherapy Ethics* (pp. 613–625). Oxford University Press.

Mason, R. (2021). Hermeneutical Injustice. In J. Khoo & R. Sterken (Eds.), *The routledge handbook of social and political philosophy of language* (pp. 247–258). Routledge.

Mattioli, M. (2021). Second Thoughts on FDA's Covid-Era Mental Health App Policy. *Articles by Maurer Faculty*. https://www.repository.law.indiana.edu/facpub/3033

McCradden, M., Hui, K., & Buchman, D. Z. (2023). Evidence, ethics and the promise of artificial intelligence in psychiatry. *Journal of Medical Ethics*, *49*(8), 573–579. https://doi.org/10.1136/jme-2022-108447

McDougall, R. J. (2019). Computer knows best? the need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

McLeod, C. (2021). Trust. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy (fall 2021 edition)*. https://plato.stanford.edu/archives/fall2021/entries/trust/

Medeiros, J. (2021). How tech is changing healthcare. From rapid development and rollout of the Covid-19 vaccines to the science of isolation, machine-learning-enabled gene editing and digitised medicine [accessed: 01.12.2023]. *Wired*. https://www.wired.co.uk/article/future-health-trends

Medina, J. (2022). Group agential epistemic injustice: Epistemic disempowerment and critical defanging of group epistemic agency. *Philosophical Issues*, *32*(1), 320–334. https://doi.org/10.1111/phis.12221

Medina, J. (2020). Trust and Epistemic Injustice. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 52–63). Routledge New York.

Mendola, D., & Pera, A. (2022). Vulnerability of refugees: Some reflections on definitions and measurement practices. *International Migration*, *60*(5), 108–121. https://doi.org/10.1111/imig.12942

Mills, C. W. (2005). "Ideal Theory" as Ideology. *Hypatia*, *20*(3), 165–183. https://doi.org/10.1111/j.1527-2001.2005.tb00493.x

Miner, A. S., Shah, N., Bullock, K. D., Arnow, B. A., et al. (2019). Key Considerations for Incorporating Conversational AI in Psychotherapy. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00746

Mink, A., Parmar, V. S., & Kandachar, P. V. (2014). Responsible Design and Product Innovation from a Capability Perspective. In *Responsible Innovation 1. Innovative Solutions for Global Issues* (pp. 113–148). Springer.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., et al. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Moes, F., Houwaart, E., Delnoij, D., & Horstman, K. (2020). Questions regarding 'epistemic injustice' in knowledge-intensive policymaking: Two examples from Dutch health insurance policy. *Social Science and Medicine*, *245*, 1–9.

Morley, J., Machado, C. C., Burr, C., Cowls, J., et al. (2020). The ethics of AI in health care: A mapping review. *Social Science and Medicine*. https://doi.org/10.1016/j.socscimed.2020.113172

Mossey, J. M. (2011). Defining Racial and Ethnic Disparities in Pain Management. *Clinical Orthopaedics and Related Research*, *469*(7), 1859–1870. https://doi.org/10.1007/s11999-011-1770-9

Nguyen, T. (2023). PDMP causes more than just testimonial injustice. *Journal of Medical Ethics*, 2023–109112. https://doi.org/10.1136/jme-2023-109112

Nickel, P. J. (2009). Trust, staking, and expectations. *Journal for the Theory of Social Behaviour*, *39*(3), 345–362.

Nickel, P. J. (2022). Trust in medical artificial intelligence: A discretionary account. *Ethics and Information Technology*, *24*(7). https://doi.org/10.1007/s10676-022-09630-5

Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can we make sense of the notion of trustworthy technology? *Knowledge, Technology & Policy*, *23*(3), 429–444.

Nussbaum, M. (2000). *Women and Human Development: The Capabilities Approach*. Cambridge University Press.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Oliva, J. D. (2022). Dosing Discrimination: Regulating PDMP Risk Scores. *California Law Review*, *110*(1), 47–115. https://doi.org/10.15779/Z38Z31NP8J

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, *29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Pearl, J. (2000). *Causality: Models, reasoning and inference* (Vol. 110). Cambridge University Press. https://doi.org/10.2307/3182612

Phillips, A. (2017). 'They're rapists.' President Trump's campaign launch speech two years later, annotated [accessed: 01.10.2023]. *The Washington Post*. https://www.washingtonpost.com/news/the-fix/wp/2017/06/16/theyre-rapists-presidents-trump-campaign-launch-speech-two-years-later-annotated/

Picco, L., Lam, T., Haines, S., & Nielsen, S. (2021). How prescription drug monitoring programs influence clinical decision-making: A mixed methods systematic review and meta-analysis. *Drug and Alcohol Dependence*, *228*, 1–19. https://doi.org/10.1016/j.drugalcdep.2021.109090

Pohlhaus, G. (2017). Varieties of Epistemic Injustice. In I. J. Kidd, J. Medina, & G. Pohlhaus (Eds.), *The Routledge Handbook of Epistemic Injustice* (pp. 13–26). Routledge.

Pozzi, G. (2023a). Automated opioid risk scores: a case for machine learning-induced epistemic injustice in healthcare. *Ethics and Information Technology*, *25*(1). https://doi.org/10.1007/s10676-023-09676-z

Pozzi, G. (2023b). Testimonial injustice in medical machine learning. *Journal of Medical Ethics*, *49*, 536–540. https://doi.org/10.1136/jme-2022-108630

Pozzi, G. (2023c). Further remarks on testimonial injustice in medical machine learning: a response to commentaries. *Journal of Medical Ethics*, 2023–109302. https://doi.org/10.1136/jme-2023-109302

Pozzi, G., & Durán, J. M. (2024). From ethics to epistemology and back again: informativeness and epistemic injustice in explanatory medical machine learning. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-01875-6

Pozzi, G., & Van den Hoven, J. (2023). Physicians' Professional Role in Clinical Care: AI as a Change Agent. *The American Journal of Bioethics*, *23*(12), 57–59. https://doi.org/10.1080/15265161.2023.2272924

Rampasek, L., & Goldenberg, A. (2018). Learning from Everyday Images Enables Expert-like Diagnosis of Retinal Diseases. *Cell*, *172*(5), 893–895. https://doi.org/10.1016/j.cell.2018.02.013

Rathod, S., Pinninti, N., Irfan, M., Gorczynski, P., et al. (2017). Mental Health Service Provision in Low- and Middle-Income Countries. *Health Services Insights*, *10*. https://doi.org/10.1177/1178632917694350

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

Richardson, P., Griffin, I., Tucker, C., Smith, D., et al. (2020). Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *Lancet (London, England)*, *395*, e30–e31.

Richman, K. A. (2004). *Ethics and the Metaphysics of Medicine. Reflections on Health and Beneficence*. MIT Press.

Rogers, W. A. (2002). Is there a moral duty for doctors to trust patients? *Journal of Medical Ethics*, *28*(2), 77–80. https://doi.org/10.1136/jme.28.2.77

Russo, F. (2021). Value-promoting concepts in the health sciences and public health. *Preprint*. http://philsci-archive.pitt.edu/19287/

Russo, F. (2023). Causal Pluralism and Public Health. In S. Venkatapuram & A. Broadbent (Eds.), *The Routledge Handbook of Philosophy of Public Health* (pp. 98–114). Routledge.

Russo, F., Schliesser, E., & Wagemans, J. (2023). Connecting ethics and epistemology of AI. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01617-6

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26*, 2749–2767.

Sakakibara, E. (2023). Epistemic injustice in the therapeutic relationship in psychiatry. *Theoretical Medicine and Bioethics*. https://doi.org/10.1007/s11017-023-09627-1

Sand, M., Durán, J. M., & Jongsma, K. R. (2021). Responsibility beyond design Physicians: Requirements for ethical medical AI. *Bioethics*, 162–169. https://doi.org/https://doi.org/10.1111/bioe.12887

Sedlakova, J., & Trachsel, M. (2022). Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent? *American Journal of Bioethics*, *23*(5), 4–13. https://doi.org/10.1080/15265161.2022.2048739

Sherman, B. R. (2016). There's No (Testimonial) Justice: Why Pursuit of a Virtue is Not the Solution to Epistemic Injustice. *Social Epistemology*, *30*(3), 229–250. https://doi.org/10.1080/02691728.2015.1031852

Simonite, T. (2020). How an algorithm blocked kidney transplants to black patients [accessed: 14.12.2023]. *Wired*. https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/

Singh, R., Kalra, M. K., Nitiwarangkul, C., Patti, J. A., et al. (2018). Deep learning in chest radiography: Detection of findings and presence of change. *PLoS ONE*, *13*(10), 1–12. https://doi.org/10.1371/journal.pone.0204155

Solon, O. (2016). Karim the AI delivers psychological support to Syrian refugees [Accessed: 25-09-2022]. *The Guardian*. https://www.theguardian.com/technology/2016/mar/22/karim-the-ai-delivers-psychological-support-to-syrian-refugees

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search*. The MIT Press.

Starke, G., van den Brule, R., Elger, B. S., & Haselager, P. (2022). Intentional machines: A defence of trust in medical artificial intelligence. *Bioethics*, *36*(2), 154–161. https://doi.org/10.1111/bioe.12891

Sutrop, M. (2019). Should we trust artificial intelligence? *Trames: A Journal of the Humanities and Social Sciences*, *23*(4), 499–522.

Symons, J., & Alvarado, R. (2022). Epistemic injustice and data science technologies. *Synthese*, *200*(2), 1–26. https://doi.org/10.1007/s11229-022-03631-z

Szalavitz, M. (2021). The Pain Was Unbearable. So Why Did Doctors Turn Her Away? [accessed: 23.08.2023]. https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/

Tallant, J. (2019). You can trust the ladder, but you shouldn't. *Theoria*, *85*(2), 102–118.

Tanesini, A. (2022). Intellectual Autonomy and Its Vices. In J. Matheson & K. Lougheed (Eds.), *Epistemic Autonomy*. Routledge.

Tekin, Ş. (2021). Is Big Data the New Stethoscope? Perils of Digital Phenotyping to Address Mental Illness. *Philosophy and Technology*, *34*(3), 447–461. https://doi.org/10.1007/s13347-020-00395-7

Tekin, Ş. (2023). Ethical Issues Surrounding Artificial Intelligence Technologies in Mental Health: Psychotherapy Chatbots. In *Technology ethics: A philosophical introduction and readings* (pp. 152–159). Routledge.

Thomas, A., Kuper, A., Chin-Yee, B., & Park, M. (2020). What is "shared" in shared decision-making? Philosophical perspectives, epistemic justice, and implications for health professions education. *Journal of Evaluation in Clinical Practice*, *26*(2), 409–418. https://doi.org/10.1111/jep.13370

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Trawalter, S., Hoffman, K. M., & Waytz, A. (2012). Racial Bias in Perceptions of Others' Pain. *PLoS ONE*, *7*(11). https://doi.org/10.1371/journal.pone.0048546

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., et al. (2021). The ethics of algorithms: key problems and solutions. *AI and Society*. https://doi.org/10.1007/s00146-021-01154-8

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3

Ursin, F., Timmermann, C., & Steger, F. (2022). Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics*, *36*(2), 143–153. https://doi.org/10.1111/bioe.12918

Vadivelu, N., Kai, A. M., Kodumudi, V., Sramcik, J., et al. (2018). The Opioid Crisis: a Comprehensive Overview. *Current Pain and Headache Reports*, *22*(3), 2–6. https://doi.org/10.1007/s11916-018-0670-z

Van de Poel, I., & Kroes, P. (2014). Can Technology Embody Values? In P. Kroes & P.-P. Verbeek (Eds.), *The Moral Status of Technical Artefacts* (pp. 103–124). Springer.

Van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

Van de Sande, D., van Genderen, M. E., Huiskens, J., Gommers, D., et al. (2021). Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Medicine*, *47*(7), 750–760. https://doi.org/10.1007/s00134-021-06446-7

Van den Hoven, J. (2012). Human Capabilities and Technology. In I. Oosterlaken & J. Van den Hoven (Eds.), *The Capability Approach, Technology and Design* (pp. 27–36). Springer.

Van den Hoven, J. (1998). Moral Responsibility, Public Office and Information Technology. In I. T. M. Snellen & W. B. van de Donk (Eds.), *Public Administration in an Information Age: A Handbook* (pp. 97–112). IOS Press.

Veluwenkamp, H., & Van den Hoven, J. (2023). Design for values and conceptual engineering. *Ethics and Information Technology*, *25*(1). https://doi.org/10.1007/s10676-022-09675-6

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 284–293. https://doi.org/10.1145/3351095.3372876

Vilaza, G. N., & McCashin, D. (2021). Is the Automation of Digital Mental Health Ethical? Applying an Ethical Framework to Chatbots for

Cognitive Behaviour Therapy. *Frontiers in Digital Health*, *3*. https://doi.org/10.3389/fdgth.2021.689736

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, *34*(4), 1607–1622.

Wardrope, A. (2015). Medicalization and epistemic injustice. *Medicine, Health Care and Philosophy*, *18*(3), 341–352. https://doi.org/10.1007/s11019-014-9608-3

Watson, D. S., & Floridi, L. (2021). The explanation game: A formal framework for interpretable machine learning. *Synthese*, *198*(10), 9211–9242. https://doi.org/10.1007/s11229-020-02629-9

Watson, L. (2021). *The Right to Know: Epistemic Rights and Why We Need Them*. Routledge.

Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, 4038–4052. https://www.researchgate.net/publication/310235081_Automated_Inference_on_Criminality_using_Face_Images

Wu, X., & Zhang, X. (2017). Automated inference on criminality using face images. https://doi.org/10.48550/arXiv.1611.04135

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science Technology and Human Values*, *41*(1), 118–132. https://doi.org/10.1177/0162243915605575

Zhang, L., Losin, E. A., Ashar, Y. K., Koban, L., et al. (2021). Gender Biases in Estimation of Others' Pain. *Journal of Pain*, *22*(9), 1048–1059. https://doi.org/10.1016/j.jpain.2021.03.001

About the author

Giorgia Pozzi (1995) was born in Cesena, Italy. She completed her Ph.D. on the ethics and epistemology of medical machine learning at Delft University of Technology between September 2020 and April 2024. Her Ph.D. project was embedded in the research infrastructure SoBigData++. Prior to her Ph.D., she completed an M.A. in Philosophy at Ludwig-Maximilians-Universität München (Germany) in 2020 with a focus on the ethics of artificial intelligence. She obtained a B.A. in Chinese Studies in 2018 and a B.A. in Philosophy at the same university in 2017, focusing on metaethics and moral philosophy.

List of Publications

- **Pozzi, G.** & Durán, J. M. *Social causes and epistemic (in)justice in machine learning-mediated medical practices.* Forthcoming in The Routledge Handbook of Causality and Causal Methods (P. Illari and F. Russo eds.).

- **Pozzi, G.** & Durán, J. M. (2024) From ethics to epistemology and back again: informativeness and epistemic injustice in explanatory medical machine learning. *AI & Society.* https://doi.org/10.1007/s00146-024-01875-6

- **Pozzi, G.** & Van den Hoven, J. (2023). Physicians' professional role in clinical care: AI as a change agent. *The American Journal of Bioethics*, 23:12, 57-59. https://www.tandfonline.com/doi/full/10.1080/15265161.2023.2272924

- **Pozzi, G.** (2023). Further remarks on testimonial injustice in medical machine learning: a response to commentaries. *Journal of Medical Ethics*, 49:551-552. http://dx.doi.org/10.1136/jme-2023-109302

- De Proost, M. & **Pozzi, G.** (2023). Conversational Artificial Intelligence and the Potential for Epistemic Injustice. *The American Journal of Bioethics*, 23:5, 51-53. https://doi.org/10.1080/15265161.2023.2191020

- **Pozzi, G.** (2023). Automated opioid risk scores: a case for machine learning-induced epistemic injustice in healthcare. *Ethics and Information Technology, 25*(1), 3. https://doi.org/10.1007/s10676-023-09676-z

- **Pozzi, G.** (2023). Testimonial injustice in medical machine learning. *Journal of Medical Ethics*, 49:536-540. https://doi.org/10.1136/jme-2022-108630

- **Pozzi, G.** & Durán, J. M. (2022). Ethics. In G. Comandé (ed.) *Encyclopaedia of Law and Data Science*, pp. 153–159, Edward Elgar Publishing.

- Durán, J. M. & **Pozzi, G.** (under review). What is Trustworthy AI?

- **Pozzi, G.** & De Proost, M. (under review). Keeping an AI on the mental health of vulnerable populations: reflections on the potential for participatory injustice.

- Van den Hoven, J., **Pozzi, G.**, Stauch, M. et al. (under review). The European approach to artificial intelligence across geo-political models of digital governance.