Contributory Injustice, Epistemic Calcification, and the Use of AI Systems in Healthcare

Mahi Hardalupas
Independent researcher
mchardalupas@gmail.com

Abstract

AI systems have long been touted as a means to transform the healthcare system and improve service user outcomes. However, these claims frequently ignore the social context that leaves service users subject to epistemic oppression. This paper introduces the term "epistemic calcification" to describe how the use of AI systems leads to our epistemological systems becoming stuck in fixed frameworks for understanding the world. Epistemic calcification leads to contributory injustice as it reduces the ability of healthcare systems to meaningfully consider alternative understandings of people's health experiences. By analysing examples of algorithmic prognosis and diagnosis, this paper demonstrates the challenges of addressing contributory injustice in AI systems and the need for contestability to focus on more than the AI system and on the underlying epistemologies of AI systems.

1 Introduction

Artificial Intelligence (AI) systems have long been heralded for their potential to "transform" medicine and healthcare (Topol 2019; Obermeyer, Ziad and Emanuel, Ezekiel J., M.D. 2016). This trend shows no signs of slowing in the era of generative AI systems (Zhang and Boulos 2023). In a global survey, more than two in three clinicians expressed excitement around the use of AI in healthcare with a third claiming to have used AI in the last year (Sooch and Pateraki 2023). However, despite the optimism from some corners on the potential of AI systems, there are several accompanying criticisms of the hype surrounding medical AI such as the difficulties of implementing medical AI in practice, reduction of trust between patients and doctors, and turning aspects of care into unexplainable black box decisions (Cabitza et al., 2017; Chin-Yee & Upshur, 2019; Grote & Berens, 2020; Lee Peter et al., 2023; Morley, 2023; Starke et al, 2020).

This paper focuses on another kind of criticism of medical AI: its potential to exacerbate contributory injustice, a form of epistemic oppression, in medicine. Epistemic oppression, as introduced by Dotson (2012), describes the ways people are systematically excluded and dismissed from producing knowledge. One type of epistemic oppression is contributory injustice, when marginalized groups are excluded from contributing equally to dominant understandings of their experiences. Most work on epistemic oppression in medical AI systems has focused on Fricker's (2007) examples of testimonial and hermeneutical injustice (Chin-Yee and Upshur 2019; Faissner et al. 2024; Pozzi 2023; Walmsley 2020), which are well-documented in medicine (Carel and Kidd 2014; Blease, Carel, and Geraghty 2017). Less work has focused on understanding how medical AI systems can perpetuate other forms of epistemic oppression, as explored in Herzog (2022) and Faissner et al. (2024). This paper builds on this research by analysing how AI systems in medicine exacerbate contributory injustice.

This paper argues medical AI systems will amplify contributory injustice due to their tendency for 'epistemic calcification', when knowledge systems "harden" and become set in a fixed framework for understanding a problem. Section 2 outlines Dotson's (2012, 2014) framework of epistemic oppression including contributory injustice and how it manifests in healthcare and medicine. Section 3 turns to how AI systems generally worsen contributory injustice and introduces the concept of 'epistemic calcification' to explain why this occurs. Section 4 demonstrates the risks of epistemic calcification leading to contributory injustice through two examples of applications of AI systems in medicine: prognosis and diagnosis. Section 5 reflects on the challenge of addressing this issue in a discourse that prioritises the values of patients over their knowledge, failing to treat them as equal members of knowledge production. This paper concludes that contributory injustice cannot be addressed through simple technical fixes and requires recognizing the use of AI systems as representative of underlying knowledge systems that themselves need to be challenged.

2 Epistemic Oppression and Contributory Injustice

Epistemic oppression describes the systematic exclusion of

individuals or groups from participating in knowledge production (Dotson, 2014). Theorising about epistemic oppression starts from the basic assumption that there are very different ways to conceive of and understand the world and that this knowledge is socially mediated. What we know or understand can often be intimately shaped by our social identities and experiences. For example, an asylum seeker fleeing war in their home country will likely have a deeper understanding of the situation in their home country than the asylum decision maker where they are seeking asylum.

While there are ways we can understand ourselves individually, Dotson (2014) explains how the process of coming to understand and assess our experiences is also dependent on collective epistemic resources, which are collective ways of understanding. Epistemic resources can be our language, conceptual frameworks or criteria of evaluation (Pohlhaus Jr. 2012) as well as our emotions and embodiments which can inform how we understand our experiences (Shotwell 2017).

Some epistemic resources will be more dominant than others, whether due to social, political or pragmatic reasons, and which are the dominant resources may vary depending on the community. For example, particular slang or idioms may be dominant in a regional or ethnic community but the dominant linguistic epistemic resources in a country may be more formal. This example helps demonstrate two points about epistemic resources. First, dominant epistemic resources won't necessarily be those that are most popular or widely used. Second, and relatedly, dominant epistemic resources may become so due to institutional adoption. In the language example, the style guide of a major newspaper or news outlet could play a significant role in establishing what counts as dominant epistemic resources. In Doston's terms, epistemic resources prevail within an epistemological system, which is a "holistic concept that refers to all the conditions for the possibility of knowledge production and possession" (2014: 121). Because epistemic resources shape how we understand the world, the epistemological system and dominant epistemic resources within it can play a profound role in constraining how we act or make decisions. For example, the asylum seeker is subject to how dominant epistemic resources in the country she's seeking asylum in categorise what is happening in her home country, what the risks are to her, and whether she's believed or not.

Epistemic oppression occurs because not everyone is equally able to utilise or contribute to dominant epistemic resources, leading to potential injustice and oppression. Dotson (2014) introduces three kinds of epistemic oppression. This paper is primarily concerned with third-order epistemic oppression, where the exclusion of peoples' ways of understanding themselves is due to dominant epistemic resources being inadequate or irrelevant to capturing an individual or group's experience. One form of epistemic oppression is 'contributory injustice' described in Dotson (2012).

Contributory injustice occurs when people rely on structurally prejudiced dominant epistemic resources and choose to ignore alternative epistemic resources in ways that systematically exclude and harm an individual or group (Dotson 2012). Ignoring alternative ways of understanding experiences can sometimes lead to material harms for individuals or groups. For example, if a decision is made to deny asylum to an asylum seeker due to dominant prejudiced epistemic resources not recognising the risks to the asylum seeker even though they can point to alternative epistemic resources showing clear evidence of harms, contributory injustice is responsible for both epistemic and material harms to the individual. The failure to understand a person's experience results from the choice to ignore resources that explain their experience and continue to use the dominant prejudiced epistemic resources that fail to accommodate the speaker's experience.

One context where contributory injustice can lead to material harms is healthcare. Miller Tate (2019) explains how contributory injustice manifests in psychiatry in the experiences of service users who hear voices and their interactions with the medical system. There are several other examples of contributory injustice in healthcare such as the experiences of intersex and trans communities, who both have been and are subject to epistemic oppression due to dominant epistemic resources in medicine understanding their social identities as a medical condition to be treated (Spade 2003; Merrick 2019; Keyes, Hitzig, and Blell 2021; Faye 2021). Disabled people also can face contributory injustice in how medicine relies on dominant epistemic resources that ignore alternative understandings of their abilities and experiences that have been developed by disabled communities (Blease, Carel, and Geraghty 2017; Amundson 2000).

Another recent example is the patient-led research introducing the medical condition 'Long COVID' – the persistence of COVID-19 symptoms for more than 28 days, which can occur in some patients. Patients experiencing these symptoms found that doctors were at a loss to help them as their experiences did not fit into a medical framework that believed someone could not suffer ongoing COVID-19 symptoms. This resulted in patients developing their own resources to catalogue their experiences and culminated in a report defining this condition as "Long COVID" (Mccorkell et al. 2021; Callard and Perego 2021). The inability for these patients to be understood by doctors was partly due to intentional or non-intentional ignorance of these alternative resources and dismissal of anecdotal evidence. While now the "Long COVID" reports have been discussed in medical journals and mainstream media outlets, these patients were subject to contributory injustice and may continue to face epistemic oppression from medical professionals who are unaware or choose to ignore these epistemic resources or the condition itself.

Addressing contributory injustice is made especially difficult due to the resilience of our epistemological systems. Doston (2014) describes epistemological resilience as the ability of an epistemological system to adapt to changes without having to be restructured. For example, a scientific

theory can demonstrate epistemological resilience when new scientific discoveries are made without the theory having to be radically overhauled. As Dotson emphasises, resilience is not itself a bad feature – the stability of resilient epistemological systems allows us to understand change and share a framework. In the scientific theory example, a resilient theory allows scientists to share new insights and develop applications based on this theory. Similarly, having a stable understanding of the world, allows us to identify shortcomings in our communities and societies and reason about interventions.

But when our dominant epistemic resources are structurally prejudiced, their resilience makes it difficult to overcome contributory injustice as it is harder to acknowledge and address the inadequacy of dominant resources. Addressing third-order epistemic oppression requires knowers to grapple with the limitations of their own epistemological systems and understand the merits of an outside perspective on them. For example, Miller Tate (2019) argues addressing contributory injustice requires the medical system and medical professionals to be aware of and take seriously alternative epistemic resources for understanding service user's experiences, such as hearing voices. In summary, one step towards avoiding contributory injustice requires embracing the plurality of alternative epistemic resources for understanding our experiences and the world.

The next section introduces epistemic calcification to explain why AI systems are particularly likely to exacerbate contributory injustice.

3 Epistemic Calcification

There is a developing literature exploring epistemic injustice (a subset of epistemic oppression) and AI in various contexts, including healthcare (Faissner et al. 2024; Pozzi 2023; Symons and Alvarado 2022). Herzog (2022) and Miragoli (2024) both discuss the potential for AI systems to cause epistemic oppression. Herzog (2022) argues that the lack of explainability of AI systems can in some cases cause third-order epistemic oppression in medicine. Miragoli (2024) argues epistemic oppression is related to the tendency to only treat statistically dominant information as epistemically relevant. This section argues we can better understand the risks of contributory injustice and epistemic oppression from AI systems through the lens of "epistemic calcification".

Epistemic calcification is the tendency of epistemological systems and epistemic resources to "harden" and become stuck in a fixed framework for structuring and understanding the world. In epistemic calcification, the practical ways an epistemological system is embedded and implemented in a person or institution's activities can leave parts or all of the epistemological system fixed and resistant to change. While there may be ways an individual's epistemic resources could become epistemically calcified, here the main concern is epistemic calcification in institutions or communities. The more epistemically calcified a system is, the more barriers there will be to shifting the system to consider and include alternative epistemic resources.

Classification systems are a key example of how epistemic calcification can occur. Bowker and Star (1999) provide a detailed analysis of how classification systems act as powerful infrastructures shaping our experiences in the world through multiple domains. For example, the categories included on the census represent a particular epistemological system for understanding the world. In theory, the categories on the census ought to provide a comprehensive description of the population they survey but in practice, the decisions made around what categories to include and how they're defined are laden with social and political values. Because of how the census feeds into political decision-making, the categories chosen can become epistemically calcified as they reinforce the ongoing reification of those categories by other political bodies, institutions and individuals who are forced to express their identities in ways that comply with those categories. For example, Guyan (2022) critically discusses the framing of sexual orientation questions added to Scotland's 2022 census and the exclusion of non-binary identities as shaping national conceptions of the LGBTQ+ community.

Epistemic calcification contrasts with Dotson's (2014) epistemological resilience, which emphasises the ability for epistemic resources to adapt to change without having to be restructured. Epistemologically resilient systems will likely retain some degree of adaptivity as a means of remaining dominant. However, in epistemic calcification, an epistemological system remains dominant despite a lack of adaptability. Instead, the epistemological system remains dominant as it is embedded in systems in ways that introduce barriers to adapting and restructuring these resources. In this way, epistemic calcification can also lead to 'value capture' as defined by Nguyen (forthcoming). In value capture, "a person or group adopts an externally-sourced value as their own, without adapting it to their particular context" (7). Nguyen gives an example of the introduction of law school rankings fundamentally shifting how students make decisions and how departments organise their missions. Prior to the USN&WR law school rankings, there was more value pluralism in how people approached deciding which law school to go to as there was a recognition that different law schools may be more or less aligned with what a student wanted from their studies. The introduction of the ranking diminished this value pluralism as the rankings came to dominate students' decision process as well as incentivising law schools to abandon their specific missions in pursuit of what would help them climb the rankings. Just as value capture reduces value pluralism, epistemic calcification hinders epistemic pluralism. This consequence of epistemic calcification heightens the risk of contributory injustice.

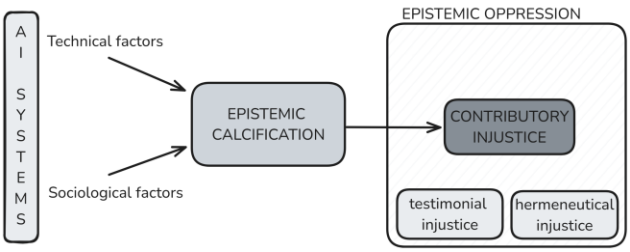

Figure 1. Diagram illustrating the relationship between AI systems, epistemic calcification and contributory injustice.

The next sections build on this definition of epistemic calcification to show how AI systems exacerbate it. Figure 1 illustrates this paper's central argument mapping how AI systems, due to technical and sociological factors, lead to epistemic calcification which exacerbates contributory injustice, a form of epistemic oppression.

3.1 Epistemic Calcification in AI Systems

This section argues AI systems cause epistemic calcification because using AI systems keeps an epistemological system calcified by introducing barriers to adapting it. Digital systems leading to epistemic calcification is not a new phenomenon. In an interview from 1985, Joseph Weizenbaum remarks, "I think the computer has from the beginning been a fundamentally conservative force. It has made possible the saving of institutions pretty much as they were, which otherwise might have had to be changed […] the computer has acted as […] a force which kept power or even solidified power where is[sic] already existed" (ben-Aaron 1985). This centralization of power highlights the risks of epistemic calcification through digital systems – by entrenching epistemological frameworks in a system, institutions can further reinforce power imbalances that continue to disadvantage those with experiences that aren't represented or understood through those systems. AI systems are tools for epistemic calcification which increase the likelihood of contributory injustice.

The roots of epistemic calcification from AI stem from foundational assumptions in 20th century AI research. In knowledge-based AI (or Good Old-Fashioned AI), an approach more popular in the early days of AI based on the work of Newell and Simon, the goal was to mimic expert-level human problem-solving by explicitly creating a 'knowledge base' of relevant facts that AI systems would then infer new knowledge from. Alison Adam (1998) shows how the knowledge base of these early AI systems made implicit assumptions about who counted as knowers and prioritised their knowledge systems. One example is the still-ongoing Cyc project that aims to create a knowledge base for AI systems that includes all the basic knowledge about the world needed to capture human common sense. The question: whose common sense does this represent? Adam analyses Cyc to show the worldview represented in the system is significantly influenced by the largely white middle-class male developers, who ultimately decide which facts about the world constitute common sense knowledge. Systems like Cyc lead to epistemic calcification, where a particular framework for understanding the world becomes fixed in the system thereby influencing how it will utilise knowledge to solve problems.

Another example analysed by Adam is Soar, a problem-solving AI project that worked by formally defining a set of problem states that the AI system would then "search" to find a solution. In this case, Adam traces how Soar's problem-solving approach was based on empirical studies on human problem-solving conducted on male US college students. Thus, "the theory of human problem solving[…] which has strongly influenced not just the development of Soar but of symbolic AI in general, is based on the behavior of a few technically educated, young, male, probably middle-class, probably white college students working on a set of rather specialised tasks" (Adam, 2004, 339). Here epistemic calcification is brought about, not just from the knowledge embedded into the system, but by fixing a particular approach to problem-solving in the first place.

In these historic examples, epistemic calcification occurred due to the explicit project of incorporating particular types of knowledge into AI systems. However, one might be tempted to think that newer approaches to AI like machine learning (ML), which avoids explicit articulation of knowledge in lieu of allowing the algorithm to "learn" for itself, would thereby be less susceptible to epistemic calcification. The next two sections outline how ML systems can introduce epistemic calcification due to technical and sociological factors.

3.2 Technical Factors Contributing to Epistemic Calcification

Epistemic calcification can result from factors related to the technical development of AI systems.

First, while there is no explicit knowledge representation in machine learning (ML) systems, there are several other ways that epistemic calcification can occur in machine learning. Many AI systems used in medicine for diagnostic or prognostic purposes use supervised learning. In supervised learning, an algorithm is trained on datasets labelled by human annotators, so the desired input-to-output mapping is known. In other words, if you are training an image recognition algorithm to be able to diagnose skin conditions, then the images it will be trained on will be labelled appropriately according to whether they are examples of a particular skin condition or not. Supervised learning imposes a structure of categories that are decided by the developers in a similar way to the development of Cyc. In contrast to assumptions that labels represent a ground-truth about the real

world, they are shaped by those defining the labels and require negotiation and iteration to decide on consistent categories (Birhane et al. 2022; Miceli, Posada, and Yang 2021; D. Wang, Prabhat, and Sambasivan 2022). Because of this, ML systems will still tend to epistemic calcification as they will entrench the specific categories and labels determined during training or fine-tuning.

Second, the dataset used for training can result in epistemic calcification. It is well-known that algorithms, such as facial recognition algorithms, can exhibit bias due to the lack of diversity in the datasets they are trained on (Buolamwini and Gebru 2018). However, epistemic calcification considers further what frameworks for problem-solving can be reinforced by datasets used for training. For example, if a diagnostic tool for skin conditions is based on image recognition, then this promotes the idea that images are all that is needed for diagnosis. The privileging of particular modes of evidence for diagnosis obscure other kinds of evidence that are more holistic, such as non-visual properties of the skin, texture and how it feels to the patient. The same is true for datasets that rely on standardised demographic categories for patient classification (e.g. sex, race, gender). Boulicault (2023) shows how data classifications can both essentialise social differences (e.g. when the use of sex categories in COVID-19 health outcomes data led researchers to look for biological instead of social explanations for infection disparities) or 'invisibilize' groups (e.g. when COVID-19 data classification failed to explicitly collect data on trans/non-binary people). Here the chosen datasets will dictate a framework for approaching a problem, just as in Soar, leading to epistemic calcification. This shows the limits of approaches which change the composition of datasets for addressing epistemic calcification. More diverse datasets may reduce bias but it will not normally shift the implicit approach to problem-solving used by the system or who holds the power to impose this view on others (Burrell 2024; Miceli, Posada, and Yang 2021).

3.3 Sociological Factors Contributing to Epistemic Calcification

Epistemic calcification can also be the result of sociological factors in the wider AI ecosystem.

First, there is a broad commitment to the project of scalability. The enthusiasm for algorithms that have the ability "to scale" encourages ML systems that can be developed in one domain and then be transferred for application in new domains and new markets (Sloane et al. 2022). This means that programmers are incentivised to strive for systems with fixed definitions of notions like 'fairness'. These operationalisations are then assumed to be universal and thus generalisable to multiple contexts (Selbst et al. 2019). However, this scalability presumes a homogeneity of users that is harmful to those who are already marginalized (Hanna and Park 2020). In other words, the underlying goal of having a neutral universal subject has not changed from the days of Cyc and expert systems research in AI. In this way, the motivating goals of the field continue to encourage epistemic calcification.

Second, there are factors reducing the likelihood of challenges to the use of AI systems. Automation bias, the overreliance on automated systems despite contrary evidence, further entrenches the dominance of the "knowledge" of AI systems. This can be particularly harmful when information from automated systems contradicts the testimony of already marginalized groups, as occurred in the UK's Post Office Horizon scandal, where a flawed software system led to hundreds of Post Office workers being wrongfully prosecuted for fraud (Gray 2023). Even when digital systems are known to be flawed and in need of update, there is a deep inertia to updating legacy systems with a recent survey suggesting a majority of healthcare organisations still use legacy operating systems ("2021 HIMSS Healthcare Cybersecurity Survey"). Through increased reliance on automated systems and long-standing inertia to updating these digital systems, even when presumed or found to be flawed, epistemic calcification can be perpetuated even under conditions where people are aware of potential issues.

4 How AI Systems Reinforce Contributory Injustice

So far this paper has discussed contributory injustice in healthcare and introduced how AI systems are used for epistemic calcification. In this section, those two strands are combined to explain how the use of AI systems in healthcare leads to epistemic calcification, which consequently reinforces contributory injustice.

4.1 Prognostic Algorithms

This section considers the use of a prognostic algorithm for risk prediction. Obermeyer et al. (2019) examine a healthcare risk prediction algorithm used by hospital systems for millions of patients annually to decide where to invest their resources. The purported goal of these systems is to identify those with the greatest healthcare needs so they can be placed into care management programs. Based on a set of inputs such as insurance type, prescribed medications, and demographic information like age and sex (but specifically excluding race), the algorithm is trained to generate a risk score that evaluates an individual's healthcare needs. To assess healthcare needs, the algorithm uses past healthcare costs as a proxy. This means that a person who has higher predicted medical costs in the next year will be given a higher risk score. These risk scores are then used to flag people for enrolment into extra healthcare programs (at 97th percentile) and refer people to their PCPs (at 55th percentile).

In their paper, Obermeyer et al. (2019) set out to answer: does this algorithm's risk score accurately predict the healthcare needs of patients? To test this, they define a new measure for healthcare needs, the 'comorbidity score',

which quantifies the number of active chronic conditions a patient has. When Black and white patients with the same risk score are compared, the Black patients have a much higher comorbidity score. This means that the algorithm is systematically underestimating how sick Black patients are when predicting healthcare needs, leading to bias in who is recommended to healthcare programs by the algorithm. A large proportion of Black patients will not be recommended for the healthcare program on the basis of their risk scores, even though a white patient with the same number of healthcare issues would be recommended. Indeed, when Obermeyer et al. (2019) control for this, they show that the fraction of Black patients recommended for the healthcare program would more than double (from 17.7% to 46.5%).

This is attributed to the issue of 'label bias'. Since healthcare providers spend less on Black patients in general, the use of healthcare costs as a proxy for healthcare needs disadvantages Black patients who will have lower predicted healthcare costs. As expected, this effect persists despite race not being an explicit variable during training. Obermeyer et al (2019) suggest that this bias is due to the different kinds of costs that emerge, with Black patients more likely to have emergency visits and dialysis treatments than inpatient and outpatient costs. Benjamin (2019) connects this to the historical context of the medical data used to train these algorithms, which requires reckoning with "segregated hospital facilities, racist medical curricula, and unequal insurance structures" (422).

As demonstrated by their research, the epistemological system that the prognostic algorithm is part of is structurally prejudiced both due to biased data and label bias. Epistemic calcification occurs due to the embedding of this algorithm based on a fixed knowledge system into the medical triage process in a way that fixes the understanding of people's healthcare experience to a narrow set of epistemic resources. By choosing predicted healthcare costs as the appropriate proxy in the algorithm, the system prioritises epistemic resources which favour groups for whom greater healthcare need will generally result in increased expenditure of healthcare resources, such as white patients. This conceptual framework for understanding health obscures the challenges with accessing healthcare in the first place, racial inequalities in how healthcare resources are distributed, who has health insurance in the US, and lack of trust in medical professionals. When Benjamin (2019) situates this research in its historical context, she is demonstrating the inadequacy of the dominant epistemic resources shaping the AI system to reckon with Black peoples' experiences in the healthcare system. In this way, the system results in contributory injustice.

One consequence of the epistemic calcification of the prognostic algorithm is that it provides no opportunity for patients to contribute their own understanding of their healthcare needs on their terms. Their health experiences are categorised into conditions (known or unknown) to create labels for machine classification without ability to appeal or contest them. This increases the likelihood of epistemic oppression, where patients' testimony regarding their health status could be dismissed in lieu of a risk score that underestimates how sick they are, mirroring the issues of the UK Post Office Horizon scandal (Gray 2023). This prognostic system bypasses the need to rely on peoples' testimony at all. Similarly, it serves as a source of epistemic oppression and contributory injustice in that it excludes alternative labels and proxies for understanding healthcare needs.

Obermeyer et al. (2019) show that a new variable for 'healthcare needs' that combines health prediction with cost prediction results in an 84% reduction in bias in the algorithm. The dramatic reduction demonstrates how important the selection of labels is in determining the outcomes from algorithms. But the fact that there is still bias in the algorithm shows that overcoming these issues is not as simple as changing the label. The use of an epistemically calcified algorithm will force a fixed mode of problem-solving and triage that still perpetuates harms, even if there is a reduction in bias.

Ultimately, we should challenge the purpose of this algorithm in the first place. Is it there to help patients or to reduce costs for the hospital system, hence explaining the choice of healthcare costs as the label? This question is in line with scholarship documenting the deep connections between the historical development of AI and administrative bureaucracy, such that AI "thinks like a corporation" (Penn 2018; Cave 2019). The issue at the heart of the matter is how AI systems preserve bureaucracy, which we already know is oppressive and harmful to marginalized communities (Alkhatib 2021).

To meaningfully address contributory injustice, redress and the ability to contest the decision of an AI system is not enough. It requires contesting the underlying structurally prejudiced epistemological system that delivers these decisions at scale and introducing ways for alternative epistemic resources to be adopted.

4.2 Diagnostic Algorithms

This section considers a second example of diagnostic algorithms. In medical AI literature, diagnosis is sometimes presented as a value-neutral activity with a focus on the potential for ML systems to improve diagnostic accuracy and discover "objective" measures for disorders such as schizophrenia (Starke et al. 2020). This obscures the social, ethical, and political implications of diagnosis. Consider Esmé Weijun Wang's discussion of her diagnosis: "[s]ome people dislike diagnoses, disagreeably calling them boxes and labels, but I've always found comfort in preexisting conditions. I like to know that I'm not pioneering an inexplicable experience" (E. W. Wang 2019). For her, diagnosis can be an empowering process providing someone with epistemic resources through which to understand their experiences and situate them in an established community. For others, diagnoses can be oppressive forcing trans people to conform to

certain medicalised understandings of their experiences in order to access resources (Spade 2003). Similarly, diagnoses are frequently considered to be value-laden and may carry stigma or negative associations (Ho 2004). These nuances fail to be understood in diagnostic algorithmic systems.

More generally, the narratives that inform and shape diagnoses are bound up in the collective epistemic resources that we use and develop. To this end, as Keyes (2020) argues, we must pay attention to the discursive harms of diagnostic AI tools: "what algorithms do is not just a question of material goods and (direct) material harms, but a question of the discourses and narratives they depend on, perpetuate and legitimise" (5). Through critical discourse analysis, Keyes illuminates how the researchers developing diagnostic AI tools for autism frame autism from a highly medicalised perspective that portrays autistic people as socially abnormal, lacking communicative competence, and lacking understanding of themselves and others. They show how this denies autistic people agency and personhood as well as excluding their knowledge of their own experiences from narratives that will become calcified in these AI systems. Instead, the research guiding diagnostic AI developers centres the perspectives of caregivers and healthcare professionals over autistic people.

Here the scientific research that underlies the development of diagnostic AI systems for autism is structurally prejudiced against autistic people's ability to contribute their own resources for understanding their experiences. Epistemic calcification occurs here as the embedding of these diagnostic AI systems will further disempower people who are already excluded from processes of medical knowledge production. Service users in a healthcare system are often considered sources of data and not participants in the epistemic process of decision-making (Carel and Kidd 2017; Scrutton 2017). While the structural prejudice may already exist in the scientific research independent of AI systems, the use of an algorithmic system increases the inertia inherent in the system and provides further barriers to challenging and adapting this framework. In doing so, it leads to contributory injustice as the algorithm will systematically fail to account for alternative accounts of autism that don't pathologise their experiences. The contributory injustice occurs since "one valid perspective on an experience is lost because another […] perspective is dominant and exclusive, giving rise to a one-sided interpretation" (Scrutton, 2017, 350).

In a non-AI context, Scrutton (2017) explains how service users are often forced to articulate their experiences to fit into an existing framework, which can potentially exclude other factors that are important to the patient. Furthermore, the very structure of contemporary healthcare can privilege certain modes of sharing knowledge or evidence in ways that disadvantage different modes of understanding illness (Carel and Kidd 2014; Shotwell 2017). For example, medicine often privileges quantitative measures for understanding disease at the risk of ignoring subjective measures that can help one understand the experience of illness in order to better help patients (Carel 2007). These risks will only be exacerbated through the use of AI systems.

5 Challenges

So far, this paper has outlined the connection between epistemic calcification in AI and the risk of contributory injustice when AI systems are used in prognosis and diagnosis. This section discusses a proposal to alleviate contributory injustice in medicine and explains why it fail for AI systems.

Recall that addressing contributory injustice requires being aware of alternative epistemic resources for understanding experiences and the world. This allows someone to have the flexibility to shift their perspective appropriately to engage with an individual or group on their epistemic terms. For healthcare professionals, this means engaging with alternative perspectives in a way that bridges the gap between them and the service user such that the latter is considered an equal participant in the medical decision-making process. Furthermore, it will also expose clinicians to differing non-medical perspectives disturbing the typically narrower views they have been exposed to throughout their training. In this way, clinicians can strive to avoid the epistemic calcification that leads to contributory injustice.

5.1 Value-Flexible AI Systems

When it comes to AI systems, the path to avoid epistemic calcification is unclear. One potential solution is value-flexible AI. McDougall (2019) argues that AI systems undermine shared decision-making as they impose particular values on how medical decisions are ranked. The example she uses is how IBM's Watson for Oncology ranked decisions based on maximising life, which may not align with patient preferences and values. She argues that to address this we need AI systems that are 'value-flexible' and responsive to different individual patient values to avoid reinforcing medical paternalism. Would value-flexible AI suffice to address epistemic calcification?

Let's start by accepting McDougall's suggestion that we need value-flexible AI. Immediately, it is hard to envisage how this would be implemented with our current paradigms of AI. Take her own example of Watson for Oncology, which uses life maximisation to assess and weight different cancer treatments. A value-flexible AI system would presumably allow for different values to be used for assessing and weighting cancer treatments. One plausible value that might be deemed important to this decision is quality of life. How would this be implemented in a value-flexible AI system? What constitutes quality of life itself will vary from individual to individual so there will be no simple operationalisation to implement in an AI system that would satisfy individual patient values. Ultimately, the values that can be incorporated in AI systems for medical decision-making will prioritise those that are quantifiable rather than qualitative measures that may be equally or more important to a

patient. In doing so, AI systems will still embody epistemic calcification, which excludes other perspectives on evaluating medical decisions.

However, aside from questions around implementation, the real danger of this framing of value-flexible AI is that it situates the risk of medical AI systems at the level of values focusing concern on the lack of alignment between AI and patient values. While conflict with patient values is definitely a significant issue (Birch et al. 2022), this emphasis has diminished the importance of patient knowledge as an equally pressing issue. When the potential for AI to conflict with someone's knowledge is discussed in the literature, it centres on the conflict between the clinicians' knowledge and AI system's recommendation rather than that of the patient. For example, Grote & Berens (2020) warn of how ML algorithms could undermine the epistemic authority of clinicians. Similarly, when Walmsley (2020) warns of the risk of epistemic injustice in AI medical diagnosis, it is the risk that "a GP's opinion may be discounted or rejected simply because the GP is human" (5). Framing the issue in this way assumes that the significance of patient participation in decision-making centres on making sure their values are reflected in medical decisions rather than acknowledging the patients' status as knowers. In the context of medical AI, rarely is it considered that patients themselves may be disadvantaged by the epistemic status accorded to AI systems and the forms of knowledge these systems privilege. This is despite patients being equally if not far more likely to deal with negative consequences of use of AI systems compared to doctors.

Kukla (2007) challenges this "fact-value division of labor" between doctor and patient, emphasising how medical knowledge is not value-neutral and that patient narratives can contribute valuable knowledge that is often missing from physicians' expertise. This echoes lessons from phenomenological accounts of illness, which argue that first-person experiences of illness are key to understanding illness, in contrast to naturalist accounts of disease that solely define disease as biological dysfunction (Carel 2007). In doing so, the worry is that naturalist accounts of disease ignore how patient experience and subjective measures of illness constitute valuable medical knowledge for understanding disease. To make this more concrete, consider the example in §4.2. The issues autistic people face with algorithmic diagnostic systems are not stemming from a conflict with their value preferences but a more fundamental issue with how their own understanding of their experiences are dismissed by those with power to create and sustain the epistemological framework of medical systems represented in AI systems. What is required to address this is the ability for flexibility in the knowledge represented and framings embedded in AI systems in the first place. In other words, challenging an AI system must allow a person to contest the epistemic resources underlying it.

When we consider the risk of medical AI systems at the level of knowledge rather than values, there is more reason to be skeptical about the potential for flexible AI that avoids epistemic calcification. It is not sufficient to change the explicit values that AI systems utilise in their decision-making without also interrogating the knowledge that medical AI systems rely on. Since the data that AI systems are trained on will necessarily be restricted to information that can be "datafied", they will struggle to accommodate phenomenological data, which is a long-standing challenge for AI (Chin-Yee and Upshur 2019). Chin-Yee & Upshur (2019) argue that the potential exacerbation of epistemic injustice with medical AI systems is a result of their failure to include phenomenological and qualitative data about illness. However, there is something to gain from recognising the role that epistemic calcification of AI systems plays in this too. Contributory injustice results not just from the kinds of data and evidence used in an AI system but from the calcification of this epistemic system introducing barriers to challenging it.

This problem is deeper than a simple technical fix. Even if there were ways to amend this on a technical level, attention to alternative hermeneutical resources does not fit with the universalising mindset that underlies AI systems. As we develop systems that are built to scale for more general use, we must contend with a loss of context and flexibility (Boyd & Crawford, 2012). Given this, those using AI systems will struggle to recognise the limitations of their own framework in such a way that allows them to shift to alternative ones.

One could argue that if AI systems were used in tandem with clinician judgment, they would be able to introduce some degree of flexibility that could prevent epistemic calcification. Given patients are often subject to epistemic oppression at the hands of clinicians, this should introduce some caution in relying on clinicians to mitigate contributory injustice from AI systems in medicine. Furthermore, the presence of AI systems can change the behaviour of those who use them where individuals may exhibit automation bias by being overly influenced by automated systems in spite of contradictory evidence (Skitka et al. 2000; Bond et al. 2018) or resist recommendations in biased ways (Selbst et al. 2019). The use of algorithms could also shift epistemic norms in healthcare by enforcing the standardisation of medical data to make it usable by algorithms or encouraging clinicians to align their conceptual frameworks with those used by algorithms to make better inferences from them (Grote and Berens 2020). Based on this, the epistemic calcification of AI systems may further calcify clinicians' epistemic resources resulting in healthcare infrastructures that are more susceptible to epistemic oppression.

6 Conclusion

The issue of epistemic oppression is not a new one in medicine and is not dependent on the use of AI systems in medicine. What this paper has aimed to do is provide an analysis to show why the use of AI systems in medicine is likely to

exacerbate contributory injustice rather than resolve it. Through epistemic calcification, AI systems promote a fixed and inflexible approach to understanding experiences and introduce barriers to engaging with alternative understandings. In medicine, where there is a need to accommodate alternative frameworks of understanding experience of illness to ensure epistemic justice, this quality of AI systems culminates in a situation liable to perpetuate epistemic oppression in medicine. This paper does not outline a clear solution. Instead, it calls for questioning what we want from AI systems in medicine, arguing for contestability to centre on more than just values and challenging whether AI systems can be developed in ways that better accommodate different epistemologies. If not, we must recognize the limitations of using AI systems in domains such as medicine without reinforcing harms to already marginalized communities.

Acknowledgements

I would like to thank Nedah Nemati, Os Keyes, Andrew Strait, Annika Froese, Morgan Thompson, Davide De Mola and Ioana Popisteanu for their encouragement, comments and discussions at different stages of working on this paper. Thank you also to the attendees of the 2021 workshop on Feminism, Social Justice and AI for their helpful feedback when an earlier version of this work was presented.

References

2021 HIMSS Healthcare Cybersecurity Survey. 2021. Healthcare and Information Management Systems Society. https://www.himss.org/sites/hde/files/media/file/2022/01/28/2021_himss_cybersecurity_survey.pdf

Alkhatib, A. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Yokohama: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445740

Amundson, R. 2000. Against normal function. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 31(1): 33–53. https://doi.org/10.1016/S1369-8486(99)00033-3

ben-Aaron, D. 1985. Weizenbaum examines computers and society. The Tech, 105(16). https://web.archive.org/web/20230217170819/http://tech.mit.edu/V105/N16/weisen.16n.html

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

Birch, J., Creel, K. A., Jha, A. K., & Plutynski, A. 2022. Clinical decisions using AI must consider patient values. Nature Medicine, 28(2): 229–232. https://doi.org/10.1038/s41591-021-01624-y

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. 2022. The Values Encoded in Machine Learning Research. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 173–184. https://doi.org/10.1145/3531146.3533083

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

Bond, R. R., Novotny, T., Andrsova, I., Koc, L., Sisakova, M., Finlay, D., Guldenring, D., McLaughlin, J., Peace, A., McGilligan, V., Leslie, S. J., Wang, H., & Malik, M. 2018. Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. Journal of Electrocardiology, 51(6): S6–S11. https://doi.org/10.1016/j.jelectrocard.2018.08.007

Boulicault, M. 2023. How Medical Technologies Materialize Oppression. The American Journal of Bioethics, 23(4): 40–43. https://doi.org/10.1080/15265161.2023.2186528

Bowker, G. C., & Star, S. L. 1999. Sorting things out: Classification and its consequences. Cambridge, MA: MIT Press.

Boyd, D., & Crawford, K. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information Communication and Society, 15(5): 662–679. https://doi.org/10.1080/1369118X.2012.678878

Buolamwini, J., & Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conference on Fairness, Accountability and Transparency 2018. New York: Proceedings of Machine Learning Research. 77–91.

Burrell, J. 2024. Automated decision-making as domination. First Monday. https://doi.org/10.5210/fm.v29i4.13630

Cabitza, F., Rasoini, R., & Gensini, G. F. 2017. Unintended consequences of machine learning in medicine. JAMA - Journal of the American Medical Association, 318(6), 517–518. https://doi.org/10.1001/jama.2017.7797

Callard, F., & Perego, E. 2021. How and why patients made Long Covid. Social Science and Medicine, 268. https://doi.org/10.1016/j.socscimed.2020.

Carel, H. 2007. Can I be ill and happy? Philosophia, 35(2): 95–110. https://doi.org/10.1007/s11406-007-9085-5

Carel, H., & Kidd, I. J. 2014. Epistemic injustice in healthcare: A philosophical 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 The Routledge handbook of epistemic injustice, edited by I. J. Kidd, J. Medina, & G. Pohlhaus Jr, 336–346. Routledge.

Cave, S. (2019). To save us from a Kafkaesque future, we must democratise AI. The Guardian. https://www.theguardian.com/commentisfree/2019/jan/04/future-democratise-ai-artificial-intelligence-power. Accessed: 2024-08-17.

Chin-Yee, B., & Upshur, R. 2019. Three problems with big data and artificial intelligence in medicine. Perspectives in Biology and Medicine, 62(2): 237–256. https://doi.org/10.1353/pbm.2019.0012

Dotson, K. 2012. A Cautionary Tale: On Limiting Epistemic Oppression. Frontiers: A Journal of Women Studies,

33(1): 24–47. https://doi.org/10.5250/fronjwomestud.33.1.0024

Dotson, K. 2014. Conceptualizing Epistemic Oppression. Social Epistemology, 28(2): 115–138. https://doi.org/10.1080/02691728.2013.782585

Faissner, M., Kuhn, E., Müller, R., & Laacke, S. 2024. Detecting your depression with your smartphone? – An ethical analysis of epistemic injustice in passive self-tracking apps. Ethics and Information Technology, 26(2): 28. https://doi.org/10.1007/s10676-024-09765-7

Faye, S. 2021. The Transgender Issue: An Argument for Justice. Penguin UK.

Fricker, M. 2007. Epistemic Injustice: Power and the Ethics of Knowing. Clarendon Press.

Gray, C. 2023. Testimonial Injustice in Governmental AI Systems. In KI-Realitäten: Modelle, Praktiken Und Topologien Maschinellen Lernens, edited by R. Gross & R. Jordan, 67–92. Transcript Verlag. https://doi.org/10.14361/9783839466605-004

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

Guyan, K. 2021. Constructing a queer population? Asking about sexual orientation in Scotland's 2022 census. Journal of Gender Studies 31(6): 782:792 . https://doi.org/10.1080/09589236.2020.1866513

Hanna, A., & Park, T. M. 2020. Against Scale: Provocations and Resistances to Scale Thinking arXiv:2010.08850

Herzog, C. 2022. Inexplicable AI in Medicine as a Form of Epistemic Oppression. In 2022 IEEE International Symposium on Technology and Society (ISTAS), 1, 1–5. https://doi.org/10.1109/ISTAS55053.2022.10227139

Ho, A. 2004. To be labelled, or not to be labelled: That is the question. British Journal of Learning Disabilities, 32: 86–92.

Keyes, O. 2020. Automating autism: Disability, discourse, and Artificial Intelligence. The Journal of Sociotechnical Critique, 1(1): 1-31.

Keyes, O., Hitzig, Z., & Blell, M. 2021. Truth from the machine: Artificial intelligence and the materialization of identity. Interdisciplinary Science Reviews, 46(1–2): 158–175. https://doi.org/10.1080/03080188.2020.1840224

Kukla, R. 2007. How Do Patients Know? Hastings Center Report, 37(5): 27–35. https://doi.org/10.1353/hcr.2007.0074

Lee Peter, Bubeck Sebastien, & Petro Joseph. 2023. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. New England Journal of Medicine, 388(13): 1233–1239. https://doi.org/10.1056/NEJMsr2214184

Mccorkell, L., Assaf, G. S., Davis, H. E., Wei, H., & Akrami, A. 2021. Patient-Led Research Collaborative: embedding patients in the Long COVID narrative. Pain reports 6(1): 913.

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

Merrick, T. 2019. From 'Intersex' to 'DSD': A case of epistemic injustice. Synthese 196(11): 4429–4447. https://doi.org/10.1007/s11229-017-1327-x

Miceli, M., Posada, J., & Yang, T. 2021. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? arXiv:2109.08131

Miragoli, M. 2024. Conformism, Ignorance & Injustice: AI as a Tool of Epistemic Oppression. Episteme: 1–19. https://doi.org/10.1017/epi.2024.11

Morley, J. 2023. Thinking Critically about AI in Healthcare. https://drive.google.com/file/d/15PiSt1fEuKX0SzCnnbudswA2stYUY6v/view. Accessed: 2024-08-18.

Nguyen, C. T. Forthcoming. Value Capture. Journal of Ethics and Social Philosophy. https://philpapers.org/archive/NGUVCH.pdf

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

Obermeyer, Ziad, M. D., & Emanuel, Ezekiel J., M.D., Ph. D. 2016. Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine 375(13): 1212–1216. https://doi.org/10.1056/NEJMp1606181.Predicting

Penn, J. 2018. AI thinks like a corporation—And that's worrying. The Economist. https://www.economist.com/open-future/2018/11/26/ai-thinks-like-a-corporation-and-thats-worrying. Accessed: 2024-08-18

Pohlhaus Jr., G. 2012. Relational Knowing and Epistemic Injustice: Toward a Theory of Willful Hermeneutical Ignorance. Hypatia 27(4): 715–735. https://doi.org/10.1111/j.1527-2001.2011.01222.x

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

Scrutton, A. P. 2017. Epistemic Injustice and Mental Illness. In The Routledge handbook of epistemic injustice, edited by I. J. Kidd, J. Medina, & G. Pohlhaus Jr., 347–355. Routledge.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598.

Shotwell, A. 2017. Forms of knowing and epistemic resources. In The Routledge handbook of epistemic injustice, edited by I. J. Kidd, J. Medina, & G. Pohlhaus Jr., 79–88. Routledge.

Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. 2000. Automation bias and errors: Are crews better than individuals? International Journal of Aviation Psychology 10(1): 85–97. https://doi.org/10.1207/S15327108IJAP1001_5

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. 2022. Participation Is not a Design Fix for Machine Learning. In Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–6. https://doi.org/10.1145/3551624.3555285

Sooch, R., & Pateraki, M. 2023. Ipsos Digital Doctor Survey 2023. Ipsos. https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-04/Ipsos%20Digital%20Doctor_v3_Apr2023.pdf. Accessed: 2024-08-18.

Spade, D. 2003. Resisting Medicine, Re/modeling Gender. Berkeley Women's Law Journal 18(1): 15–37.

Starke, G., De Clercq, E., Borgwardt, S., & Elger, B. S. 2020. Computing schizophrenia: Ethical challenges for machine learning in psychiatry. Psychological Medicine, May. https://doi.org/10.1017/S0033291720001683

Symons, J., & Alvarado, R. 2022. Epistemic injustice and data science technologies. Synthese 200(2): 87. https://doi.org/10.1007/s11229-022-03631-z

Tate, A. J. M. 2019. Contributory injustice in psychiatry. Journal of Medical Ethics 45(2): 97–100. https://doi.org/10.1136/medethics-2018-104761

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

Walmsley, J. 2020. Artificial intelligence and the value of transparency. AI and Society 36(2): 585-595. https://doi.org/10.1007/s00146-020-01066-z

Wang, D., Prabhat, S., & Sambasivan, N. 2022. Whose AI Dream? In search of the aspiration in data annotation. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1–16. https://doi.org/10.1145/3491102.3502121

Wang, E. W. 2019. Diagnosis. In The Collected Schizophrenias, by E. W. Wang. Graywolf Press. https://doi.org/10.1017/CBO9781107415324.004

Zhang, P., & Boulos, M. N. K. 2023. Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. Future Internet 15(9): 286.