Hermeneutic Harm, Epistemic Injustice, and Disrupted Sense-Making

REDACTED

REDACTED

REDACTED

Abstract: As AI becomes integrated into everyday life, it also becomes a powerful force in shaping whose voices are heard, whose knowledge is validated, and whose experiences are recognized as legitimate. While AI technologies are often portrayed as objective tools for decision-making, their development and deployment are deeply entangled with the social and political structures in which they operate. This makes them not only susceptible to contributing to epistemic injustice but capable of amplifying it at scale.

In this paper we draw on the recent theoretical innovation of Rebera *et al* who mobilize the idea of *hermeneutic harm* and use it to identify potential sites of epistemic injustice. Hermeneutic harm is "emotional and psychological pain caused by a prolonged inability to make sense of an event (or events) in one's life" [1]. While not specific to AI, hermeneutic harms can be exacerbated by AI-systems. Hermeneutic harm can arise in cases where agent-like AI-systems cause a harm, which under normal circumstances would be explainable by appeal to the intentions, motivations, or actions of human agents.

Drawing on the rich analytic resources of the epistemic injustice literature, we identify three forms of AI-induced hermeneutic harm: obscured self-understanding, asymmetrical testimonial obstruction, and bureaucratic displacement. We then highlight how each of these hermeneutic harms are trigged by specific kinds of epistemic injustice and can illustrate instances of wrongful exclusion from meaning and sense-making practices.

Keywords: Epistemic Injustice, Hermeneutic Harm, Sense-Making

Introduction

As AI becomes integrated into everyday life, it also becomes a powerful force in shaping whose voices are heard, whose knowledge is validated, and whose experiences are recognized as legitimate. While AI technologies are often portrayed as objective tools for decision-making, their development and deployment are deeply entangled with the social and political structures in which they operate. This makes them not only susceptible to contributing to epistemic injustice but capable of amplifying it at scale.

In this paper we draw on the recent theoretical innovation of Rebera *et al* [1] who mobilize the idea of *hermeneutic harm* and use it to identify potential sites of epistemic

injustice. Hermeneutic harm is "emotional and psychological pain caused by a prolonged inability to make sense of an event (or events) in one's life" [1]. While not specific to AI, hermeneutic harms can be exacerbated by AI-systems. Hermeneutic harm can arise in cases where agent-like AI-systems cause a harm, which under normal circumstances would be explainable by appeal to the intentions, motivations, or actions of human agents.

There is a clear conceptual link between the ideas of epistemic injustice and hermeneutic harm as both deal with the harm of having some element of one's sense making capacity obstructed. The wrong of epistemic injustice results in the primary harm of being undermined in one's capacity to participate in the practice of giving, receiving, or understanding knowledge. Hermeneutic harm, as a secondary harm, can thus result in any case where epistemic injustice is present and an agent is prevented from fairly participating in meaning or sense-making practices.

Drawing on the rich analytic resources of the epistemic injustice literature, we identify three forms of AI-induced hermeneutic harm: obscured self-understanding, asymmetrical testimonial obstruction, and bureaucratic displacement. We then highlight how each of these hermeneutic harms reveal and are trigged by specific kinds of epistemic injustice.

Hermeneutic Harm and Epistemic Injustice

Recently, Rebera et al. [1] have argued that the rapid increase and proliferation of AI products and services has led to an increase in instances of hermeneutic harm. Hermeneutic harm takes the form of "emotional or psychological pain" resulting from an inability to make sense of one's experiences, especially when these experiences are also harmful or otherwise unwelcome [1]. Hermeneutic harm is thus always a secondary kind of harm as it comes about after and because of a failed attempt to make sense of a primary harm. When these harmful experiences are in some way caused by an AI system, we are at an increased risk of suffering a hermeneutic harm as we do not, in any meaningful way, have access to the (AI) agent's reason for acting or an adequate explanation of their decision making, and are thus less able to fully make sense of the experience. Hermeneutic harm is particularly significant as it extends beyond the emotional and psychological pain suffered by someone who does not understand why something has happened. Hermeneutic harm can have far ranging consequences that affect our identity formation, how we see ourselves and participate in the world, and how we go on to develop epistemic virtues.

There are notable connections to be made between hermeneutic harm and the framework of epistemic injustice as they both deal with the harm of limiting an agent's ability to make sense of their experiences. Hermeneutic harm and epistemic injustice are also both indicators of a continued (situational or systemic) failure to ensure appropriate understanding and sense-making. Epistemic injustice is foundationally an issue of social power and occurs when a person is undermined in their capacity as a giver or receiver of knowledge or is prevented from fairly participating in practices of social meaning making.

Testimonial injustice, as introduced by Fricker, becomes apparent at the point of testimonial exchange, when a speaker's claim is considered less credible than it is because of a "negative identity prejudice" held by the hearer [2]. Hermeneutic[al] injustice occurs when an agent is prevented from fully understanding a significant part of their social experiences, owing to a gap in access to interpretive resources. This gap comes about as a result of structural identity prejudice that marginalises certain groups and "obscure[s]" their experiences from the dominant, collective social understanding [2].

The concept of epistemic injustice has developed widely beyond testimonial and hermeneutical injustice, introducing new varieties of epistemic injustice that allow us to better understand the continually evolving social, political, and technological landscape that influences epistemic participation and meaning-making [3], [4], [5], [6], [7]. Two further varieties of epistemic injustice will be discussed here, namely contributory injustice [8] and pre-emptive testimonial injustice [2], [9].

Contributory injustice is brought about by the willful hermeneutical ignorance [10] of dominant groups that is mobilized to maintain, utilize, and reinforce prejudiced epistemic resources at the expense of non-dominantly situated knowers . This creates a failure to recognise non-dominant epistemic resources used to explain or help understand the experiences of non-dominantly situated individuals [8]. Pre-emptive testimonial injustice was also conceptualized by Fricker as a structural form of testimonial injustice [2], [9]. Pre-emptive testimonial injustice manifests as a failure to seek out the testimony of an agent because of a structural prejudice against their social group [5]. This injustice is pre-emptive in that it occurs before any testimonial exchange. This is crucial as any epistemic injustice that occurred after the exchange would be considered a testimonial injustice.

It is from this point that we aim to explore hermeneutic harm through the lens of epistemic injustice as a distinctly epistemic wrong. This does not take away from hermeneutic harm as identified in cases where there is no epistemic injustice present. It does, however, allow us to home in on a kind of hermeneutic harm that affects the more marginalized groups involved in and who suffer from unjust epistemic practices within the interplay between organization, AI system, and end user. The concept of hermeneutic harm is specifically useful as an identifier of such epistemic injustice.

Epistemic Injustice and AI

In this section we offer some reasons that justify the usage of epistemic injustice as a framework for understanding the impacts of AI-systems. The value-ladeness of these systems is approaching a consensus view in the literature, and it is widely acknowledged that we ought to understand AI-systems as being *socio-technical*. That is, these systems are composed of both technical and social properties [11], [12], [13], [14]. This means going beyond looking only at the technical specifics or capabilities of an AI product or service: AI evaluation needs to include social context, human interaction, cultural values and systemic impacts if it is to be truly responsible.

This broadened scope of AI impact analysis also offers the opportunity to make use of methods and concepts from the social sciences and humanities to understand the ways these systems can shape our social, political, and epistemic lives. The

literature is replete with applications of epistemic injustice to AI, under the broad umbrella of 'algorithmic epistemic injustice' [17]. Glaberson notes how AI-enabled over policing of Black communities and single mothers is a form of testimonial injustice [18]. In healthcare, Pozzi suggests that the algorithmic determination of key medical concepts such as 'addiction' or 'pain' results in "automated hermeneutical appropriation" [19]. The now infamous COMPAS recidivism algorithm has also received sustained attention from scholars interested in epistemic injustice [20], [21], and there is also an emerging trend of the historically more technical field of 'AI fairness' engaging with social and political issues [22], [23], [24].

Epistemic Injustice and Hermeneutic Harm Revisited

We have introduced epistemic injustice, and shown how it is a useful lens with which to analyze AI systems and their impacts on marginalized groups. In what follows, drawing on the idea of hermeneutic harm outlined by Rebera *et al.* [1], we will outline three novel hermeneutic harms that may arise from the deployment of particular AI applications. We will then show that these hermeneutic harms can signal instances of epistemic injustice. We thus expand the work of Rebera *et al.*: We not only outline hermeneutic harms that arise in these cases, but trace how these harms may themselves be symptomatic of deeper epistemic injustices

To recap, Rebera *et al.* [1] argue that hermeneutic harm arises when we encounter harmful actions of a kind that are standardly associated with humans, but which turn out to be performed by an AI. The action provokes a reactive attitude, but we can't regulate it appropriately because we don't have access to the agent's reasons or inner states (because AIs don't, in any normal sense, *have* reasons or (relevant) inner states). This can cause hermeneutic harm.

There is thus a connection to be made between the ideas of epistemic injustice, especially hermeneutical and contributory injustice, and hermeneutic harm as we address the reasons for and consequences of AI-induced disruptions in experiential sense-making. However, Rebera et al. [1] make a point of saying that hermeneutic harm is most likely *not* a kind of epistemic injustice. Rather, hermeneutic harm would arise in response to ongoing epistemic injustice. This captures something distinct about both the ideas of hermeneutic harm and epistemic injustice—epistemic injustice is not a harm but rather causes harm. The wrong of epistemic injustice results in the primary harm of being undermined in one's capacity to participate in the practice of giving, receiving, or understanding knowledge. Hermeneutic harm, as a secondary harm, can thus result in any case where epistemic injustice is present and an agent is prevented from fairly participating in meaning or sense-making practices.

It is worth noting that hermeneutic harm itself does not constitute an epistemic injustice, as a harm does not necessarily imply that anything unjust has occurred. Moreover, hermeneutic harm is very likely to follow on from an epistemic injustice. Suffering an injustice creates a tension between the way we expect things should happen and the way they actually happen, or as Rebera *et al*. assert, it creates inconsistencies between "global meaning" and "situational meaning" [1]. If I am denied a loan to start my small business, even though I know I have a very strong application, it would make sense for me to feel frustrated, and to call the bank and request an explanation. If no

explanation is forthcoming, or if the bank stonewalls me or sends me a convoluted, highly technical account that I am unable to understand, it would make sense for me to feel frustrated and confused. As such, when we are unable to reasonably fit our experiences to our expectations, we are at risk of suffering a hermeneutic harm.

Epistemic injustice can similarly be a catalyst for creating this tension between global and situational meaning because it prevents people from engaging in the epistemic practices that they should reasonably expect to be able to participate in [25]. Take again the example above. If I am withheld an explanation of why I was denied a loan on the foundation that I, as a layperson, wouldn't be able to understand the reason anyway—and further prevented from inquiring about the reasons why my loan was denied in the first place—I will have suffered an epistemic injustice. This epistemic injustice would put me at risk of suffering a hermeneutic harm, as I will have been prevented from accessing the explanations required for me to make sense of my experiences. Importantly, what is unjust about cases of epistemic injustice, such as this, is that the harm suffered is as a result of a wrong—a power imbalance that marginalizes certain groups and privileges others, leading to negative identity and structural prejudice which allows some groups to contribute to knowledge practices and others not.

Below, we will outline three descriptions of hermeneutic harm that may be exacerbated by AI. We then link these harms to particular forms of epistemic injustice that we find in the literature more generally.

4.1 Obscured self-understanding

Obscured self-understanding plays out in cases where AI-systems mediate important parts of an individual's identity by reinforcing harmful social stereotypes. If an agent is presented with information about their social group which is wrong, misleading, or discriminatory, they may rightly object to this. However, if the system making these suggestions continually reinforces negative stereotypes and discrimination which cannot be challenged, we run the risk of hermeneutic harm. For example, Automatic Gender Recognition (AGR) systems, used in physical access control, data analytics, advertising, and to analyse social media usage, utilises research that actively excludes and erases trans and non-gender conforming identities [26]. When "relied upon to understand (amongst other things) gendered dynamics in social media" [26], AGR reproduces gender binaries that are based on physiological characteristics, leading to underrepresentation and discrimination against those whose gender identities depart from these categorisations.

Reinforcing these understandings of gender identity may lead to the misclassification of and discrimination against those whose gender identities do not fit within this (Western) binary, resulting in hermeneutic harm as denial or erasure of one's identity will impact one's ability to make sense of one's experiences. This limited sense-making capacity is also indicative of a hermeneutic injustice which occurs as people in these communities are prevented from contributing to the knowledge systems that these classification models are trained on.

4.2 Asymmetrical Testimonial Obstruction

Misrecognition, in some cases, comes with the opportunity to protest or defend oneself from being incorrectly recognized. Decisions made on the basis of AI recommendations that offer no recourse or opportunity to rectify a misrecognition can contribute to hermeneutic harm. For example, a decision support system that incorrectly classifies a student as having overstayed their visa (when in fact they are legally in the country) would prevent the student from lodging appropriate testimony against the decision. There is therefore a testimonial obstruction delivered on the basis of a power asymmetry between human and machine. If the human agent who makes use of such a system does not fully understand why the system made the recommendation, yet acted on the basis of that recommendation, there is no proper target for the student to make a complaint. Thus, the student's expectation that they might contest the inaccurate misrecognition with appropriate testimony is frustrated. They expect somebody to give an account or explanation of what happens, but no such explanation is forthcoming, and thus the student cannot make sense of the situation.

Hermeneutic harm in the case of this type of misrecognition comes about as a result of pre-emptive testimonial injustice. In the case of governments and bureaucratic systems implementing these systems, there is a clear power dynamic that undermines one's ability to easily contest classification errors. This is exacerbated through the implementation of decision support systems, as they are awarded even greater epistemic privilege. Thus, when these systems pre-emptively misrecognize human agents, they are at a greater risk of being prevented from offering testimony to contest this misrecognition, and the human agent would in this case have suffered an epistemic injustice.

4.3 Bureaucratic Displacement

Bureaucratic displacement involves agents that interact with bureaucratic systems that make use of AI in the service of some positive project, but the AI in question does not in fact contribute to that goal, causing a disruption in the sense-making processes of human agents engaging with the bureaucratic system. A real-world example of this is the Dutch tax authority's use of an algorithmic system to detect child benefits fraud [27]. However, the system ended up unfairly discriminating against people on the basis of prejudicial 'evidence' (such as the persons' gender, or whether they are divorced). Although the tax authorities insist that there was human oversight into all cases deemed 'high risk', this would often mean that those who were deemed high risk would be subject to increased scrutiny by tax authorities. For example, it was found that switching one's gender, or having an adult son, could increase the associated risk score, which would result in further interviews with the tax authorities. When reflecting on the fact that her having an adult son pushed her score up, one woman tellingly remarked "What does he have to do with this?" [27].

It is here that we can locate bureaucratic displacement as a form of hermeneutic harm: the tax authorities, who are meant to work in the interests of the public, made use of an opaque algorithmic system that mislabeled citizens and caused them severe harm. In this case, an agent's sense-making is frustrated as they are unable to understand why government authorities are implementing measures that cause them harm, based on discriminatory evidence (such as gender or divorce) that should have no relevant impact

on their risk score. This hermeneutic harm is further indicative of an epistemic injustice, specifically contributory injustice. Contributory injustice is present in this case as willfully ignorant officials were able to benefit from the efficiency of implementing these systems and benefit financially from the allowances paid back by those wrongfully targeted as fraudulent by these systems. This would encourage the support and maintenance of the tax authority's use of algorithmic fraud detection even though the evidence produced was based on biased data. This evidence was further privileged over end user complaints and difficulties in sense making, contributing to epistemic injustice.

Conclusion

In this extended abstract we have sketched novel hermeneutic harms as a potential avenue to identify sites of epistemic injustice. First, we introduced the ideas of epistemic injustice and hermeneutic harm. Second, we argued that epistemic injustice is a useful framework to mobilize when evaluating the potential impacts of AI-systems. Third, using the idea of hermeneutic harm, we identified three hermeneutic harms that may be triggered by AI systems, and offered an analysis of the accompanying epistemic injustice(s) that might exist at the root of those harms. Further work could involve identifying other forms of hermeneutic harm and linking them to different kinds of epistemic injustice. Given the constraints of the extended abstract, we could not go further into our more positive contribution, but we hope that this brief survey and modest proposal inspires new ways of tackling the ethical issues thrown up by AI.

Reference List

[1] A. P. Rebera, L. Lauwaert, and A.-K. Oimann, "Hidden Risks: Artificial Intelligence and Hermeneutic Harm," *Minds & Machines*, vol. 35, no. 3, Jul. 2025, doi: 10.1007/s11023-025-09733-0.

[2] M. Fricker, *Epistemic injustice: power and the ethics of knowing*. Oxford: Oxford University Press, 2007.

[3] H. Carel and I. J. Kidd, "Epistemic injustice in healthcare: a philosophial analysis," *Medicine, Health Care, and Philosophy*, vol. 17, no. 4, pp. 529–540, 2014, doi: 10.1007/s11019-014-9560-2.

[4] C. Hookway, "Some Varieties of Epistemic Injustice: Reflections on Fricker," *Episteme*, vol. 7, no. 2, pp. 151–163, 2010, doi: 10.3366/epi.2010.0005.

[5] J.-Y. Lee, "Anticipatory Epistemic Injustice," *Tandf: Social Epistemology*, vol. 35, no. 6, pp. 564–576, 2021, doi: 10.1080/02691728.2021.1924306.

[6] J. Medina, "Varieties of Hermeneutical Injustice 1," in *The Routledge Handbook of Epistemic Injustice*, Routledge, 2017.

[7] J. Wanderer, "Varieties of Testimonial Injustice," in *The Routledge Handbook of Epistemic Injustice*, Routledge, 2017.

[8] K. Dotson, "A cautionary tale: on limiting epistemic oppression," *Frontiers: A Journal of Women's Studies*, vol. 33, no. 1, p. 24+, Jan. 2012, Accessed: Oct. 01, 2024. [Online]. Available: https://link-gale-

com.eux.idm.oclc.org/apps/doc/A288539334/LitRC?u=ed_itw&sid=summon&xid=ed828f4b

[9] M. Fricker, "Evolving Concepts of Epistemic Injustice," in *The Routledge Handbook of Epistemic Injustice*, Routledge, 2017.

[10] G. Pohlhaus, "Relational Knowing and Epistemic Injustice: Toward a Theory of 'Willful Hermeneutical Ignorance,'" *Hypatia*, vol. 27, no. 4, pp. 715–735, 2012, Accessed: Oct. 02, 2024. [Online]. Available: http://www.jstor.org.eux.idm.oclc.org/stable/23352291

[11] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta GA USA: ACM, Jan. 2019, pp. 59–68. doi: 10.1145/3287560.3287598.

[12] L. Sartori and A. Theodorou, "A sociotechnical perspective for the future of AI: narratives, inequalities, and human control," *Ethics and Information Technology*, vol. 24, no. 1, pp. 102–104, 2022, doi: 10.1007/s10676-022-09624-3.

[13] L. Weidinger *et al.*, "Sociotechnical Safety Evaluation of Generative AI Systems," Oct. 31, 2023, *arXiv*: arXiv:2310.11986. Accessed: Nov. 18, 2024. [Online]. Available: http://arxiv.org/abs/2310.11986

[14] O. Kudina and I. Van De Poel, "A sociotechnical system perspective on AI," *Minds & Machines*, vol. 34, no. 3, pp. 21, s11023-024-09680–2, Jun. 2024, doi: 10.1007/s11023-024-09680-2.

[15] R. Alvarado, "AI as an Epistemic Technology," *Sci Eng Ethics*, vol. 29, no. 5, Oct. 2023, doi: 10.1007/s11948-023-00451-3.

[16] S. Milano and C. Prunkl, "Algorithmic profiling as a source of hermeneutical injustice," *Philos Stud*, vol. 182, no. 1, pp. 185–203, Jan. 2025, doi: 10.1007/s11098-023-02095-2.

[17] J. Kay, A. Kasirzadeh, and S. Mohamed, "Epistemic Injustice in Generative AI," Aug. 21, 2024, *arXiv*: arXiv:2408.11441. doi: 10.48550/arXiv.2408.11441.

[18] S. Glaberson, "The Epistemic Injustice of Algorithmic Family Policing," *UC Irvine Law Review (Forthcoming)*, pp. 404–456, 2022.

[19] G. Pozzi, "Automated opioid risk scores: a case for machine learning-induced epistemic injustice in healthcare," *Ethics Inf Technol*, vol. 25, no. 1, Mar. 2023, doi: 10.1007/s10676-023-09676-z.

[20] J. Symons and R. Alvarado, "Epistemic injustice and data science technologies," *Synthese*, vol. 200, no. 2, Apr. 2022, doi: 10.1007/s11229-022-03631-z.

[21] G. Hull, "Dirty data labeled dirt cheap: epistemic injustice in machine learning systems," *Ethics Inf Technol*, vol. 25, no. 3, Sep. 2023, doi: 10.1007/s10676-023-09712-y.

[22] A. L. Hoffmann, "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse," *Information, Communication & Society*, vol. 22, no. 7, pp. 900–915, Jun. 2019, doi: 10.1080/1369118x.2019.1573912.

[23] A. Birhane, "Algorithmic injustice: a relational ethics approach," *Patterns*, vol. 2, no. 1, pp. 1–9, 2021, doi: 10.1016/j.patter.2021.100205.

[24] A. Kasirzadeh, "Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy," in *Proceedings of the 2022 AAAI/ACM Conference on*

AI, Ethics, and Society, Oxford United Kingdom: ACM, Jul. 2022, pp. 349–356. doi: 10.1145/3514094.3534188.

[25] C. L. Park, "Making sense of the meaning literature: An integrative review of meaning making and its effects on adjustment to stressful life events.," *Psychological Bulletin*, vol. 136, no. 2, pp. 257–301, Mar. 2010, doi: 10.1037/a0018301.

[26] O. Keyes, "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, p. 88:1-88:22, Nov. 2018, doi: 10.1145/3274357.

[27] M. Burges, E. Schot, and G. Geiger, "This Algorithm Could Ruin Your Life," *Wired*, Mar. 06, 2023. Accessed: Aug. 08, 2025. [Online]. Available: https://www.wired.com/story/welfare-algorithms-discrimination/