Minds and Machines (2025) 35:33
https://doi.org/10.1007/s11023-025-09733-0

Hidden Risks: Artificial Intelligence and Hermeneutic Harm

Andrew P. Rebera[1,2] · **Lode Lauwaert**[2] · **Ann-Katrien Oimann**[1,2]

Received: 23 November 2024 / Accepted: 7 July 2025

Abstract
The AI Ethics literature has identified many forms of harm caused, perpetuated or exacerbated by artificial intelligence (AI). One, however, has been overlooked. In this paper we argue that the increasing use of AI heightens the risk of 'hermeneutic harm', which occurs when people are unable to make sense of, or come to terms with, unexpected, unwelcome, or harmful events they experience. We develop several examples to support our argument that AI increases the risk of hermeneutic harm. Importantly, our argument makes no assumption of flawed design, biased training data, or misuse: hermeneutic harm can occur regardless. Explainable AI (XAI) could plausibly reduce the risk of hermeneutic harm in some cases. Thus, one respect in which this paper advances the field is that it shows the need to further broaden XAI's understanding of the social function of explanation. Yet XAI cannot fully mitigate the risk of hermeneutic harm, which (as our choice of examples shows) would persist even if all 'black-box' problems of system opacity were to be solved. The paper thus highlights an important but underexplored risk posed by AI systems.

Keywords Artificial Intelligence · AI Ethics · Explainable AI · Hermeneutic Harm · LAWS · Sense-making · Transparency

✉ Andrew P. Rebera
rebera.andrew@gmail.com

Lode Lauwaert
lode.lauwaert@kuleuven.be

Ann-Katrien Oimann
ann-katrien.oimann@military.be; ann-katrien.oimann@kuleuven.be

[1] Department of Behavioural Sciences, Royal Military Academy, Brussels, Belgium

[2] Institute of Philosophy, KU Leuven, Leuven, Belgium

Published online: 24 July 2025

1 Introduction

The AI Ethics literature has identified many forms of harm caused, perpetuated or exacerbated by AI: discrimination, unfairness, and bias to name only a few. One, however, has been overlooked. We argue that the increasing use of AI heightens the risk of *hermeneutic harm*, which occurs when people are unable to make sense of, or come to terms with, unexpected—and especially unwelcome or harmful—events they experience.

> **Hermeneutic harm**. Emotional and psychological pain caused by a prolonged inability to make sense of an event (or events) in one's life.

Hermeneutic harm is neither a new problem nor one associated exclusively with AI. Significant life events, such as bereavement, injury, violence and betrayal can leave one questioning—and struggling to understand—how or why things turned out as they did, how to go on now that so much has changed (or seems to have), or even who one really is. These questions often resist simple answer and may, in their resistance, call into question fundamental assumptions about our lives. Answers, in the form of an explanation of why something happened, or an account of who is responsible, may serve as useful resources in coming to terms with events; but sense-making is distinct from explanation or attribution of responsibility. These may sometimes—perhaps often—be prerequisites of sense-making; but they offer no guarantee, and other questions and uncertainties may remain. Prolonged exposure to such uncertainty can be harmful: hermeneutically harmful.

In this paper we show how the proliferation of AI systems increases the risk of hermeneutic harm. Following Rebera (2024), we model hermeneutic harm as *secondary* harm, i.e. as harm arising from the attempt to deal with a prior (primary) harm. (For comparison: if you are mugged, that is a primary harm; if you are traumatised by the mugging, that is a secondary harm.) Importantly, an event need not be *objectively* harmful to provoke hermeneutic harm; indeed, it is not strictly necessary that an event be experienced as harmful at all: events experienced as unfamiliar, unexpected, jarring or unsettling may all provoke processes of sense-making. This shows the potential extent of the problem. AI systems can cause hermeneutic harm even if they do not (objectively) cause primary harm and even if they function entirely as intended; we single out no particular techniques of AI (transformer-based large language models (LLMs), Reinforcement Learning from Human Feedback, neural networks, etc.), because all can give rise to hermeneutic harm; moreover, our argument makes no assumption of flawed design, biased training data, or misuse: hermeneutic harm can occur regardless.[1]

In Sect. 2 we explain what is involved in 'making sense' of apparently harmful events and how hermeneutic harm can arise when this fails. In Sects. 3 to 6 we

[1] Similarly, while we discuss 'explainable AI' (XAI) below (i.e. techniques designed to render the decision-making of AI systems transparent and understandable to humans), we do not adjudicate between different techniques since, with respect to addressing hermeneutic harm, the problem is of the kind of explanation that can be provided, not of how, from a technical XAI point of view, that explanation is produced.

present examples of hermeneutic harm due to AI systems. Hermeneutic harm is not unavoidable in any of the cases, but the likelihood of a solution varies. One promising solution is 'explainable AI' (XAI). We show (primarily in Sects. 3–5) both how XAI could mitigate some risk of hermeneutic harm and why it cannot fully do so. In Sect. 6, we use two examples of serious hermeneutic harm to illustrate its distinction from the problem of value-alignment in AI (to which it is related). We conclude (Sect. 7) that, given the nature of AI systems, and the purposes for which they are often deployed, their proliferation increases the risk of hermeneutic harm.

2 Making Sense of a Harm

We all have expectations of how things are likely to go: for us personally, for others, and in the world more broadly. Expectations can be foundational to our worldview or more peripheral, highly certain or less so, widely shared or idiosyncratic. Expectations concern matters of fact and of value (of how things *will* be and how they *should* be); they are grounded in experience, evidence, conviction, trust and faith (even if these are delusional or misplaced); they are dispositional, rather than occurrent (typically showing up only when confounded); they reveal implicit understandings of how things—people and other agents included—typically behave: how they react to different situations, how they interact, how they ought to treat each other, what they may and may not do, and so on.

The literature on meaning- and sense-making classes these kinds of expectations as 'global meaning', understood as "core schemas through which people interpret their experiences of the world" (Park, 2010, p. 258). People make sense of events by comparing their understanding of what happened—known as 'situational meaning'—against their prior expectations of what would happen (i.e. global meaning). Things make sense when *experience fits expectation*: when what happened is in line with what was expected; when global and situational meaning are in tune.[2] Unexpected or jarring events show up discrepancies between global and situational meaning. Recalcitrant experiences may confound expectations and provoke processes of sense-making aimed at resolving the tensions (these processes need not be conscious or deliberate).[3] *Assimilation* is a process of adjusting situational meaning to better fit global meaning, i.e. revising one's understanding of what happened to better fit what one expected. *Accommodation* involves the reverse adjustment, i.e. retrofitting one's expectations to better accommodate one's account of what happened (Park, 2010, p. 260). Sense-making thus involves a recursive process of reconfiguring and reconciling one's understanding of events and one's expectations and worldview more broadly.

Sense-making is, therefore, distinct from the process of establishing an explanation of an event. Discovering an explanation may be an important part of a sense-

[2] Or at least when what happens is not so out of line with what was expected as to draw conscious attention to itself.

[3] The processes of sense-making discussed above are thought to be commonplace (Neimeyer, 2016; Park, 2010) but are not presumed to be universal (Davis et al., 2010).

making process, but part of the challenge of sense-making is to fit the available explanations into one's broader understanding of an event's significance in one's life. A distinction is typically drawn between sense-making as a search for the *significance* of an event in one's life (e.g. coming to see a traumatic event as having provoked personal growth) and sense-making as a search for *comprehension*, i.e. an understanding of what happened, why, who was involved, and so on (Taylor, 1983). For present purposes we are concerned with the latter. Comprehension is, in this context, to be understood in terms of a *coherent* account of what happened. If an account is internally consistent and fits with global meaning, then its truth is of secondary importance as far as sense-making is concerned.[4] (A wildly untrue account is more likely to be inconsistent with global meaning and with the beliefs of people with whom one interacts.) Successful sense-making as comprehension involves settling on a plausible and durable account of what happened that leaves little or no tension between situational and global meaning.

Though harm is unavoidable in life, particular harms are normally unexpected and therefore call for some degree of sense-making. Some harms are attributable to bad luck (e.g. being in the wrong place at the wrong time). In these cases, 'bad luck' serves, in effect, as a label that can be attached to events to fit them neatly into a plausible account of what happened and why. In one sense, an appeal to bad luck is no explanation at all. But, in relation to sense-making, an explanation is of use to the extent that it supports one in resolving tensions between situational and global meaning: to label an event 'bad luck'—or even 'inexplicable'—might well support this.

When a harm is caused by another agent (call these 'agent-harms'), we look for explanations in terms of agency and intention. To make sense of agent-harms requires understanding the motives of those who cause them. Understanding the motives that led someone to harm us may not lessen the hurt we feel, but not knowing what happened, or not understanding why, may *exacerbate* the harm. For example, Stretesky et al. (2010) report the frustration felt by co-victims of unsolved homicides when the justice system limited their access to information about events surrounding the murder of their loved-ones. Similarly, Allais (2008) discusses the experience of Babalwa Mhlauli, whose father was abused, tortured and murdered during Apartheid in South Africa. In her testimony to the South African Truth and Reconciliation Commission, Mhlauli said "We do want to forgive, but we don't know whom to forgive" (cited at Allais, 2008, p. 40). In both cases, the co-victims' ability to make sense of what happened was obstructed; and in both cases, this obstruction plausibly constituted additional harm. In the first case, the lack of information extended co-victims' bereavement and grieving (Stretesky et al., 2010, p. 887); in the second, Mhlauli was denied the closure forgiveness may have allowed because the fitting subjects of her forgiveness remained unknown. These are cases where a lack of information, or the lack of an adequate explanation, obstructs the possibility of sense-making.

From these observations, the concepts of hermeneutic harm and epistemic injustice are clearly related. Hermeneutic harm arises from a prolonged inability to make sense of events in one's life; the variety of epistemic injustice known as *hermeneutic*

[4] Taylor (1983, p. 1168) insists on the importance of *illusion*—including bias, exaggeration, focus on positives rather than negatives, etc.—to sense-making.

injustice arises when a person or group lacks the conceptual or interpretive resources to make sense of their experience (Fricker, 2009). But it should be noted that victims of hermeneutic harm do not necessarily lack the conceptual resources necessary to name, describe, or otherwise make sense of their situations; or if they do lack those resources, those resources may be available elsewhere in their community. For example, a lack of technical understanding may leave a person struggling to make sense of a harmful decision from an AI system, but there are likely experts who could, in principle, explain. Such a person is thus arguably not victim to 'semantically produced' hermeneutic injustice (Medina, 2017). More plausibly, such cases involve 'distributive' epistemic injustice, where someone receives less of an epistemic good—in this case information—than they are due (Fricker, 2017, p. 53); insofar as this imbalance arises from the design or deployment of such AI systems, we might also speak of 'systemic', 'structural', or 'institutional' epistemic injustices (Medina, 2017; Pohlhaus Jr, 2017); we might further, in some cases, argue that a form of hermeneutic injustice arises or is exacerbated because victims lack the resources to put a 'perspective' on events, a way of thinking about them that captures their character in an illuminating way (Sliwa, 2023).

There is scope for theorising hermeneutic harm as a form of epistemic injustice—and it may indeed be fruitful to explore this. We will not, however, explore that avenue here. For one thing, to suffer hermeneutic harm does not necessarily undermine one's status or capacity to function in one's epistemic community (as per Fricker's original statement of the problem). More fundamentally, in our view hermeneutic harm is plausibly not a form of hermeneutic injustice because it arises, at least in part, *in response to* ongoing epistemic injustice. That is, hermeneutic harm arises from the prolonged inability to resolve situations of epistemic injustice; it may be, for instance, the prolonged inability to terminate a process of what Sliwa (2023) calls 'hermeneutical inquiry'.

Hermeneutic harms need not be especially serious. Countless minor primary harms are simply brushed off and so, if such acts cause hermeneutic harm at all, it is minor indeed. Yet with highly traumatic primary harms the secondary harms could be severe. In the cases to be discussed below, we argue that primary harms caused by AI systems can give rise to secondary harms that are worthy of more consideration than they have so far received in the AI and AI Ethics literatures. In Sects. 5 and 6 we present cases in which the confounding of normative expectations about how we and our loved ones should be treated leads to hermeneutic harm. But we begin now, in Sects. 3 and 4, with cases caused by the confounding of epistemic expectations about transparency.

3 Transparency Expectations I: Access To Explanations

Some harms caused by AI systems are due to the 'black box' problem, namely that certain AI systems—especially those based on advanced machine learning (ML) models such as deep neural networks (DNNs)—are *opaque*: extremely difficult to fully interpret or explain. Whereas traditional software executes clearly defined, rule-based instructions, advanced ML models identify patterns in complex datasets based

on distributed, non-linear transformations across potentially billions of parameters. This difference is central to the high performance that ML models offer. But the price of uncovering relationships in this way is a lack of transparency as to how it is achieved on any occasion (Ali et al., 2023; Barredo Arrieta et al., 2020; Gunning et al., 2019).

Consider, then, the following case.

(1) A very highly qualified applicant is summarily rejected for a position at a multinational. Surprised, dismayed, and a little angry, he requests an explanation from the HR team. He is told that the decision was based on screening and validation of his CV and supporting documents by an AI hiring system. Requesting further clarification, he is told that the AI's model bases decisions on a massive number of weighted datapoints; but which datapoints, how they are weighted, and what their evaluation involves, no one can say.[5]

This scenario involves no assumption of bias or unfairness in the hiring system's training data, design, or decision. Still, the applicant is treated shabbily, and hermeneutic harm might follow. He might reasonably feel he has a basic right to know the grounds on which the decision was taken.[6] Moreover, in a straightforwardly procedural sense, system opacity obstructs access to information or explanations that people plausibly need in order to make sense of the AI's decisions. Without a satisfactory account of his rejection, the applicant is hard-pressed to make full sense of his experience. While he might let the issue drop, putting it down as 'just one of those things'[7], he may reasonably expect that, given the quality of his application, there should be a good reason behind the decision. But if no reason is provided, so that he cannot tell whether it is a good one or not, he will be hard-pressed to reconcile his expectations (of being a strong candidate) with his experience (summary rejection). This is, straightforwardly, a potential source of hermeneutic harm.

When risks of hermeneutic harm are attributable, in significant part, to failure to explain AI systems' decisions, an obvious preventive move is to provide explanations. AI systems are not all irremediably opaque. 'Explainable AI' (XAI) researchers distinguish black-box ML models from 'white-box' models, whose decision-processes can be interpreted directly (e.g. because input-output relations are relatively simple), and 'grey-box' models which, though less transparent than white-box models, can be interpreted with reasonable accuracy (Ali et al., 2023, pp. 2–3). Once interpreted by developers, white- and grey-box models can, in principle, be explained to (non-expert) users. So, in cases such as (1), where decisions are relatively momentous, XAI techniques enable the development of systems based on white- or grey-box models or which, if based on black-box models, incorporate additional tools to facili-

[5] Between 2014 and 2015 Amazon used an AI CV screening tool which was found to systematically discriminate against female applicants (Crawford, 2021; Dastin, 2018). The possibility of similar issues in healthcare, insurance, social service provision, criminal justice, and other domains raises many concerns (O'Neil, 2017).

[6] We explore how confounding of such 'normative expectations' may cause hermeneutic harm in Sect. 6.

[7] Cf. Section 2 on categorising events as 'bad luck' or 'inexplicable'.

tate interpretation and/or explanation (Holland Michel, 2020; Wachter et al., 2018). If successfully implemented, these measures allow systems or their operators to provide explanations that support the sense-making of people affected by those systems.[8] Of course, not just *any* explanation will do, as the next section shows.

4 Transparency Expectations II: Incomprehensibility

In scenario (1), hermeneutic harm arises from the unresolved mismatch between the applicant's expectation of there being good reasons why his very strong application was rejected and his experience that no reasons (good or otherwise) are given. An explanation would support his sense-making, but none is provided. Consider scenario (2).

(2) Aiming at expansion, a small business owner applies for a bank loan. She submits the required paperwork, including an updated business plan, very positive financial reports and forecasts, and letters of commitment from suppliers, distributors, and so on. A few days later, her application is denied. Confused, she telephones the bank and is informed that the application was evaluated and rejected by an AI system; the decision cannot be explained over the phone, but the bank agrees to send an AI Decision Report. The Report arrives by email: a huge JSON file, which the business owner is unsure how to open, let alone interpret. In contact with the bank again, she is invited to a meeting at which the bank's IT team reviews the Decision Report and identifies many features evaluated by the ML model, as well as certain of its weightings and thresholds. They cannot, however, give precise answers to questions about the impact on the decision of specific aspects of the business owner's status or application. She leaves the meeting with no clear understanding of the reasons for the decision.

In providing the AI Decision Report, the bank in (2) outperforms the multinational in (1). Yet the Report was uninterpretable to the business owner. And if she can't make sense of the report, she can't make sense of the decision; and if she can't make sense of the decision, she is at risk of hermeneutic harm.

The Decision Report fails as an explanation because is not directed, in either form or content, to the appropriate stakeholder group. The requirements of a developer are not those of an end-user; and among end-users, the needs of a job or loan applicant are not necessarily those of, say, a scientist, healthcare professional, judge, or intelligence analyst (Gunning et al., 2019; Langer et al., 2021). The Decision Report's content (what data the model uses, which parameters are prioritised, and how they are interrelated, etc.) enables the bank to verify that the decision met their quality standards; its form suits their AI developers and IT team. But neither the form nor the content helps the business owner. For her, as for other people who are not AI specialists, it may be better to employ natural language, narrative, explanation by

[8] An explanation might facilitate sense-making yet be unsatisfactory in other respects (clearly explained discriminatory decisions are still discriminatory).

example, visualisations, and so forth (Barredo Arrieta et al., 2020; Borrego-Díaz & Galán-Páez, 2022; Lipton, 2018).

XAI research recognises the need to match styles of presentation to the intended audience. In an important contribution, Miller (2019) emphasises the importance of respecting explanatory practices as people experience them—in particular that good explanations are *contrastive*, i.e. that "people do not ask why event *P* happened, but rather why event *P* happened *instead of* some event *Q*" (Miller, 2019, p. 3). Wachter et al. (2018) stress that people often seek to understand decisions in order to then act upon them (e.g. to contest them). Miller and Wachter et al. both emphasise the role of explanation as a support to processes of understanding that enable people to move on from decisions that affect them (whether by challenging them or by reflecting on how, in a practical sense, they could have gone differently).[9] Explanations of these kinds are well-suited to mitigating risks of hermeneutic harm as posed in (1) and (2).

So, the bank in (2) could have taken a step towards minimising the risk of hermeneutic harm to the business owner by better tailoring the AI Decision Report to her needs. But we should note the considerable difficulties involved. There is generally reckoned to be a trade-off between the interpretability and accuracy of ML models (Ali et al., 2023; Barredo Arrieta et al., 2020; Gunning et al., 2019; Langer et al., 2021). According to a recent review article:

> Are the XAI techniques available today sufficient to resolve all the explainability concerns, even if several tools are used? The answer is **NO**. […] Transparent models cannot handle sophisticated real-world applications. Many applications need the modelling complexity provided by black-box systems. (Ali et al., 2023, p. 40)

Explainability tools can (at least currently) provide only an approximation of the decision-making functions in black-box systems (Mittelstadt et al., 2019); and while end-users do not necessarily need comprehensive explanations of decisions, even the limited information they do require—for instance, whether the system makes use of some factor that is intuitively relevant (and if so how)—may be hard to come by (for discussion see Lam, 2022). Moreover, further complications are introduced when AI systems produce outputs in interaction with humans or other AI or complex socio-technical systems (Borrego-Díaz & Galán-Páez, 2022; Miller, 2019).

XAI aims to overcome these challenges. But whether this can at present be reliably achieved for relatively interpretable systems is questionable, as is whether it can be achieved *at all* for black-box systems. Plausibly, situations like (1) and (2) *can* be resolved through XAI. But XAI has not been sufficiently alive to the need to mitigate the risk of hermeneutic harm. This is problematic because providing an explanation *in the wrong way* might exacerbate the problem—for instance by demonstrating to people how very little sense they can make of certain decisions affecting them (Keil,

[9] The Decision Report offers what are known as 'data-' and 'model-explainability', but not 'post-hoc explainability', i.e. it lacks the kind of contrastive, counterfactual, and reasons-based strategies highlighted by Miller, Wachter et al. and others (Ali et al., 2023, p. 10; Barredo Arrieta et al., 2020; Borrego-Díaz & Galán-Páez, 2022; Buijsman, 2022; Langer et al., 2021; Lipton, 2018; Miller, 2019; Wachter et al., 2018, 2018). It is, in this sense, an 'interpretation' rather than an 'explanation' (Ali et al., 2023, p. 2).

2006, p. 229). Explaining badly might make things worse, as the following case suggests.

5 Normative Expectations I: Social Intelligence

We have seen above how hermeneutic harm can result from a sort of epistemic failure: something apparently problematic happens and sense-making is obstructed because relevant parties are unable to satisfactorily explain or understand why. In (2) the explanation was not clear enough to support sense-making; in (1) the explanation was hardly forthcoming at all. We have suggested that XAI—for all its other social, ethical, or scientific benefits (Zednik & Boelsen, 2022)—is unlikely to fully resolve these epistemic difficulties in all cases and is therefore unlikely (for now at least) to fully minimise the risk of hermeneutic harm. But what if—as is likely in at least some cases—XAI *could* provide epistemically suitable explanations? Even here, we suggest, hermeneutic harm may occur. Consider the following variant on (2).

(3) Disappointed at having been denied a loan despite an apparently strong application, a small business owner telephones the bank for an explanation. The bank manager invites the business owner for a meeting. The next day, the bank manager welcomes the business owner into her office and explains clearly, honestly, and with specific reference to the fine details of the application and supporting documents, why the loan request was rejected. She responds firmly but respectfully to all the business owner's objections and questions. She explores other credit options with the business owner and advises on the most promising given her and the business's current financial status.

The business owner enters the interaction in (3) with nascent suspicion of having been treated unfairly. But in many continuations of the story, she leaves the interaction feeling quite otherwise, even though her expectations were not all met (she was disillusioned not to receive the loan, and she still isn't getting it). The exchange with the bank manager enables her to make sense of the rejection; and notice that she might make sense of it even if she disagrees with the criteria by which the application was rejected (and, possibly, even if she thinks the criteria objectively unfair). A positive experience with the bank manager helps to mitigate the risk of hermeneutic harm.

This shows that satisfaction of *all* expectations is no prerequisite of sense-making. What matters is that *enough* of the most important expectations—which include normative expectations about how we ought to be treated by others—are met. It also reminds us of the important social function that providing explanations plays. Whereas *epistemic* functions aim to answer questions such as *what happened*, *how*, and *why*, *social* functions have to do with building or maintaining interpersonal relationships, creating shared meaning (Miller, 2019), and with supporting actions or societal practices (e.g. excusing or blaming others, or responding to decisions that affect one) (Mittelstadt et al., 2019; Wachter et al., 2018). In (3), the bank manager's words, demeanour and behaviour—and even the setting: the photos on her

desk or the way her colleagues speak to her[10]—convey to the business owner that she is respected, that the bank and the bank manager consider her as more than just some algorithmic profile or faceless account in the system. The explanation the bank manager provides—understood as not only the content delivered but as the delivery too—serves to acknowledge the value of the business owner. This is *social* in the sense that it is a recognition of various of the expectations that we all have of each other in our interactions. But more than this, it is a mark of respect: we can call it a *normative* requirement on explanation.[11] (Moreover, failing to show respect in this way risks undermining both the epistemic and (other) social functions of explanation.) By giving an explanation in the right way, the explainer shows and conveys that the recipient is valued and respected. By giving an explanation badly, or by not giving one at all, the opposite is suggested. This is plausibly at least part of the reason for the risk of hermeneutic harm in (1) and (2).

Humans are, in general, brought up and socialised to satisfy the myriad unspoken normative expectations which we each have of each other. AI systems do not, as far as anyone can tell, have the kind of social intelligence this requires (Coeckelbergh, 2020, p. 2063). AI systems, such as those in (1) and (2), are not sensitive to intangible and unremarked qualities of interpersonal exchanges; they lack responsiveness to social and interpersonal cues that, though enormously complex on one level, are utterly mundane and unremarkable to humans. That said, it is not impossible that AI systems could be developed which are able to meet some of our normative expectations. Indeed, the problem of hermeneutic harm suggests that this should be a research avenue in XAI. For example, where the bank manager provides satisfying answers to the business owner's questions, this could plausibly be replicated by contrastive, counterfactual approaches in XAI, as discussed earlier. Miller (2019), de Graaf and Malle (2017), Hilton (1990) and others suggest the importance of providing explanations in conversational language, in terms of folk concepts and folk theory of mind; and this too could be achieved, perhaps with LLMs.[12]

Nonetheless, given the complexity and diversity of human social interaction, it remains unlikely that explanations delivered by AI systems will satisfy all our normative requirements (and even if the *appearance* is achieved, a problematic lack of sincerity is plausible). Coeckelbergh (2020, pp. 2063–2064) suggests that, since ultimately only humans can deliver satisfactory explanations, XAI will succeed only if it renders systems interpretable to developers who can then deliver explanations to

[10] These aspects of the interaction give the business owner the sense that the bank manager is the kind of person one would want—and ought to be able—to deal with on important matters like bank loans; they convey the sense that the bank cares enough about its customers that it employs bank managers like this one.

[11] This is important because it provides a (partial, additional) explanation of why we have a reasonable expectation to understand AI systems' decisions that affect us. For Coeckelbergh (2020), the right to know is a corollary of the responsibility of the system's developers or deployers. But hermeneutic harm suggests an additional reason: we need to understand decisions that affect us because, if the consequences are serious enough, and if we cannot make sense of them, we risk suffering hermeneutic harm.

[12] Moreover, the goal of XAI is inextricably bound up with the goal of trustworthy and responsible AI (Ali et al., 2023; Barredo Arrieta et al., 2020)—and a key success factor in (3) is that the business owner comes to trust the bank manager sufficiently to accept her (the bank manager's) decision even if she (the business owner) doesn't agree with it.

end-users. Similarly, Maclure (2021, p. 13) argues that to meet all the social requirements of explanation will "require maintaining the procedures and protocols that were already in place before the introduction of machine learning tools." Both observations are plausible. But, as Maclure (2021, p. 14) sees, if *this* is what explainability, in a full sense, requires, enforcing it would undermine one of the main attractions of deploying AI systems, namely their capacity to function independently. She holds (Maclure, 2021, p. 15) that "when the rights, opportunities and wellbeing of citizens are at play", it is implausible that the requirements of explainability ought to be given up or watered down. If that is right, then hermeneutic harm makes demands on explainability which it is unlikely that AI systems can fully meet. XAI may satisfy our epistemic requirements on explanation, and some of the social ones, but it cannot, for now, fully satisfy our normative requirements, and therefore cannot meet our expectations of how we, as recipients of explanations, should be treated. This increases the risk of hermeneutic harm.

6 Normative Expectations II: Fundamental Disagreement

We turn now to cases of hermeneutic harm that arise when normative expectations about how we ought to be treated are confounded.

(4) A self-driving car, equipped with the latest AI technology, navigates morning traffic in the city. A pedestrian suddenly steps onto a crossing and the car brakes sharply. However, it calculates with high probability that, given the speed and proximity of traffic behind it, to stop before the crossing will cause a multi-vehicle pile-up and so the safest option overall is not to stop, but to hit the pedestrian. The pedestrian is severely injured.

(5) During fierce urban fighting, AI modules in an airborne lethal autonomous weapon system (LAWS) make a low probability identification of a senior member of the enemy's military leadership. The target is located amid what the LAWS identifies, with high probability, as a large crowd of civilians. Analysing the likely outcomes of potential scenarios, the LAWS determines that given the target's strategic value, and despite low confidence in its identification and high confidence in the identification of the non-combatants, an attack is warranted. It unleashes a barrage of precision-guided munitions. Hundreds are killed. The target is not among them.

These scenarios are fictional but far from fantastical. In Arizona in 2018, Elaine Herzberg became the first pedestrian to be killed in an accident with a self-driving car (a test vehicle from Uber struck her as she pushed a bicycle across a road).[13] In 2021 the UN Panel of Experts on Libya (2021, para. 63) reported that LAWS had been used

[13] For a summary of the incident see https://en.wikipedia.org/wiki/Death_of_Elaine_Herzberg. Accidents involving autonomous vehicles occurred earlier, including fatal ones, but Herzberg was the first pedestrian killed.

in March 2020 against the forces of General Khalifa Haftar; and AI-based decision support systems are being used for targeting in current and recent conflicts.[14]

In scenarios (4) and (5), an AI system acts with agency and autonomy. Self-driving cars, LAWS—and a variety of other technologies—can initiate (or end) courses of action independently of human control, purposely achieving real-world outcomes. These AI systems function as agents. They are not, of course, *full moral* agents, but they nevertheless act with sufficient independence—for instance in the selection of courses of action—that their agency poses difficult questions about responsibility and answerability.[15] Without overstating the case, we may nonetheless speak here of *agent-harms* caused by AI systems, even if their agency is limited.[16]

As discussed in Sect. 2, making sense of agent-harms requires understanding the motives and intentions of the agents in question. In (4) and (5), agent-harms result from actions that would, under straightforward descriptions ('driving into a pedestrian on a crossing', 'firing missiles into a crowd of non-combatants') be considered impermissible. More pertinently, the agent-harms result from actions that violate deep-seated norms and expectations. We all have normative expectations that neither we nor our loved ones (nor anyone else for that matter) should be deliberately harmed; more particularly, we have (non-occurrent) normative expectations of not being run down on a crossing and that our loved ones will not be massacred in the street. But these normative expectations are directly confounded in (4) and (5).

Yet there are conditions under which an agent can perform an otherwise impermissible action, which confounds normative expectations, but not be (fully) culpable. To run someone down by accident is rightly considered a less serious offence than to do so intentionally; and to kill non-combatants knowingly but unintentionally is considered acceptable under some circumstances (e.g. by proponents of the Doctrine of Double Effect). These considerations might be applied to (4) and (5). Although the harms are plainly not accidental—since, in both scenarios, the prospect of harm to the victims is explicitly weighed and traded-off against alternative outcomes in the AI systems' determinations of the optimal courses of action—a case could be made that they are unintended.[17]

[14] For examples see: *Advanced Targeting and Lethality Aided System* (ATLAS) from the US Army (Schwarz, 2021; Zhang et al., 2020), *Maven Smart System* from Palantir, and the *Gospel* and *Lavender* systems allegedly used by the IDF in Gaza. For a mapping of recently reported cases, see Nadibaidze et al. (2024).

[15] Thus they give rise to cases characterised in the literature as 'responsibility gaps', where the AI system is not morally responsible for harms caused (since it is not a moral agent) and yet the humans who deployed the AI are (allegedly) not morally responsible either (since the AI selected the harmful course of action independently of their knowledge or control). For influential statements of the problem see e.g., Matthias (2004) and Sparrow (2007). For responses see, e.g.Hatherall and Sethi (2024); Hindriks and Veluwenkamp (2023); Lauwaert (2021); Nyholm, 2018a); Tigard (2021). On the relation between responsibility gaps and hermeneutic harm see Rebera (2024).

[16] On difficulties in attributing moral agency in this context see Coeckelbergh (2009, pp. 182–184).

[17] If the LAWS doesn't intend the civilian deaths but merely foresees them, might it meet the standards for Double Effect? Possibly. But before so concluding, one would have to say *far* more about the nature of its 'intentions' and other mental-state-analogues. We speak, throughout the paper, of AI systems as *intending*, *deciding*, *acting*, etc. We do not intend this as an anthropomorphic attribution of mental states to artificial entities (cf. Coghlan, 2024), but merely adopt the *intentional stance* (Dennett, 1998), since it is helpful to describe systems *as if* they have mental states.

Yet even if the actions of the car or the LAWS can be justified all things considered, there may remain a fundamental clash between our normative expectations of how things *should* go and our understanding of how they *actually* go. Some people have a very strong Kantian, deontological aversion to sacrificing a person or group for the greater good. On this view, people are always and everywhere to be regarded as ends-in-themselves and never as mere means to an end. Moreover, empirical work suggests that while strong majorities of people respond favourably to the 'utilitarian' suggestion that autonomous vehicle design should aim to minimise overall harm, by sacrificing passengers for the sake of greater numbers of pedestrians, their stated purchase intentions deviate from this: they would rather buy a self-driving car that prioritises their own or their loved ones' safety (Bonnefon et al., 2016).[18] That this is suggestive of moral ambivalence is beside the point. In relation to hermeneutic harm, the question is not (or not *primarily*) what expectations people *ought* to have, but just what expectations they have.[19] Whatever the source of the deep-seated normative expectations that one's loved ones should not be instrumentally exploited, scenarios like (4) and (5) present the possibility of a fundamental mismatch of these people's normative expectations and actual experience (i.e. of global and situational meaning).

It is important to be clear that the problem is not that the person in (4) cannot understand why the car chose to hit them, or that those who grieve the dead in (5) do not understand why the LAWS attacked.[20] In scenarios like (4), even if you knew why the car hit you, you might still (normatively) expect not to be subject to such decisions (*mutatis mutandis* in (5)). This is why it is a *fundamental* clash of normative expectations—and it can be a source of hermeneutic harm.

A solution is not impossible. We could, for instance, adjust our expectations when dealing with AI systems or agents (though, in many cases—though presumably not all—one may wonder why *we* should adapt to AI, rather than AI be designed to fit with our practices). It may then be possible to develop AI systems that cannot confound certain norms, e.g. that are hard-coded to be more Kantian, or more utilitarian, or which enable the user to calibrate the system to align with their values (Nyholm, 2018b). But even here, it is not clear that *all* clashes of expectation can be avoided, for there is no uniformity of ethical approach within, let alone across, communities (If a utilitarian calibrates his self-driving car to minimise overall harm, with the result

[18] There are reasons to treat the empirical work with a degree of scepticism (Nyholm, 2018b), not least that in one of the studies reported by Bonnefon et al. (2016), utilitarian intuitions stood up even when participants were instructed to imagine themselves and loved ones in the car (p. 1574).

[19] Of course, society should not simply accept everyone's expectations unquestioningly (for some may be unreasonable). Thus, some people may have to adjust their expectations of how they are to be treated. This rather delicate matter is what is involved in 'accommodation' (see Sect. 2). So, the point is that sometimes one *pragmatically* ought to accommodate events (in support of one's own wellbeing) and sometimes one *morally* ought to accommodate events (because, e.g., one's prior expectations were in some sense unfair).

[20] There are many cases in which an explanation of someone's behaviour suffices to resolve the tension between experience and expectation. If someone suddenly leaves the table during dinner—confounding the expectation that diners remain seated—an explanation (say, they received a message that their father was taken ill) might suffice to bridge the gap in understanding. Note also that since the tension that might be caused by your co-diner leaving the table is likely extremely short-lived, this would not meet our standards for hermeneutic harm as defined at the beginning of the paper.

that a Kantian is sacrificed for the greater good, the Kantian's expectations are confounded regardless.)

These cases evoke the problem of *value alignment*, i.e. of aligning technology (especially AI) with the values that we, as a society, consider important. This is a significant ongoing challenge to the AI community (Christian, 2020; Russell, 2023), complicated by the presence of technological, metaphysical, normative, political, and other aspects (Gabriel, 2020). The AI systems in (4) and (5) are shown to be misaligned with the values of the (co-)victims by the fact that they perform their respective harmful actions in a calculated manner, rather than by accident (which clashes with deep-rooted expectations that people should not be used as means to others' ends). In these cases, hermeneutic harm results, in some degree, from misalignment.[21]

Yet for present purposes it is important to focus on the correct misalignment. Value misalignment is the misalignment of human and AI values; but fundamentally, hermeneutic harm arises from the misalignment of expectation and experience, of global and situational meaning. Value alignment can give rise to the latter misalignment, but the two should not be confused (to do so is to confuse primary with secondary harm): the relationship between them is not one of identity, but causation. The challenge of AI alignment is therefore also the challenge of preventing misalignments of values and principles that, potentially, increase the risk of hermeneutic harm; and thus the proliferation of misaligned AI systems is also the proliferation of risks of hermeneutic harm.

7 Conclusion

Let's clarify what has and has not been affirmed. First, we have not claimed that AI systems *inevitably* cause hermeneutic harm: our claim is that AI systems *can* cause hermeneutic harm (and to this end we have presented scenarios in which they do). Second, we have not suggested that hermeneutic harm is *new*: our claim is that hermeneutic harm, though well known in other literatures, has been neglected in the AI Ethics and XAI literatures. We assert, additionally, that the growing use of AI systems has the potential to increase the incidence of hermeneutic harm. AI need not be faulty to cause hermeneutic harm; and as our choice of examples shows, hermeneutic harm does not result only from poorly designed AI systems or systems trained on biased or otherwise problematic datasets.

Cases (1) and (2) involve problems of epistemic transparency. XAI could plausibly reduce the risk of hermeneutic harm in many such cases, for, as shown by the example of Babalwa Mhlauli (discussed in Sect. 2), not knowing or not understanding why some event occurred can be a fundamental obstacle to one's efforts to come to terms with it. Supplying satisfactory explanations of AI decision-making could be pivotal in assisting sense-making on some occasions. But XAI research has not recognised hermeneutic harm, and the advances that it has made in addressing social functions of explanation do not fully address the risks. So, one respect in which this paper advances the field is that it shows the need to further broaden XAI's under-

[21] Plausibly there is no hermeneutical harm without some degree of value misalignment.

standing of the social function of explanation. This, it must be acknowledged, comes with its own risks. Too great an emphasis on making explanations psychologically fitting risks making it easier for bad actors to deceive and manipulate end-users (Baron, 2023; Lipton, 2018, pp. 21–22).[22] Moreover, in some cases, not only an explanation, but an attribution of responsibility is required to support sense-making. Here, issues of hermeneutic harm approach those of so-called 'responsibility gaps' (where harms are caused but not all responsibility for them can be satisfactorily attributed). Clarifying the structures of responsibility and answerability around AI decision-making would doubtless provide some mitigation of the risks of hermeneutic harm.[23] But as Rebera (2024) argues, hermeneutic harm can arise quite independently of whether responsibility gaps are real or not.[24] So while addressing problems associated with responsibility gaps would also mitigate *some* risk of hermeneutic harm, it can be no more than a partial response.

Other responses to mitigate the risk or impact of hermeneutic harm may be called for. One possibility, mentioned above, is that people may have to adjust their expectations of how they are to be treated. As discussed in Sect. 2, sense-making sometimes involves a process of 'accommodating' what happened by retroactively adjusting one's expectations. Plausibly, new technologies sometime motivate alterations in *everyone's* expectations. Another possibility is to design specific 'harm mitigation' measures, along the lines proposed by McMahon et al. (2019) in relation to harms arising from exploitation of people's digital data. Adapting to the case in hand, these measures could include setting up (semi-)formal bodies to provide support to victims of hermeneutic harm. This support might be therapeutic, or perhaps even legal or financial; but whatever form it takes, it is motivated by solidarity (McMahon et al., 2019, p. 164) and recognition of need, regardless of whether how the harms arose. But these are only indications of possible responses to the problem of hermeneutic harm. Our goal in this paper has been to argue that AI ethics should pay more attention to the risk of hermeneutic harm. Solutions musts be considered in further work.

Sections 5 and 6 introduced cases where hermeneutic harm is provoked by the confounding of normative expectations and principles. In these cases, the risk of hermeneutic harm persists even if all transparency problems are solved (i.e. even if it is possible to provide reasonably accurate, comprehensible explanations of the AI systems' decisions). Indeed, in a certain sense, the confounding of normative expectations is a more fundamental risk-factor for hermeneutic harm than the confounding of epistemic ones. Hermeneutic harm *can* result even if epistemic expectations are fully met, but it *cannot* occur if normative expectations are fully met.

The appeal of AI systems is that they function as effective, powerful, autonomous, non-human agents across a wide variety of potential roles. They will increasingly be

[22] There is a balance to be found between providing accounts that are accurate enough that they are not misleading and providing accounts which are psychologically compelling but inaccurate (which might exacerbate hermeneutic harm).

[23] The literature on responsibility gaps is large. See footnote 15 above.

[24] As with explanation, the sometimes apparently close relation between hermeneutic harm and responsibility gaps should not be overstated. While the kinds of cases that are alleged to give rise to responsibility gaps are cases in which there is a risk of hermeneutic harm, nonetheless the two are distinct. For full discussion of this point see Rebera (2024).

deployed to roles requiring them to make high-stakes decisions, in which their capacity to act against certain of our expectations, norms, and principles is critical to their power and effectiveness, and yet they are unable to recognise or reliably respond to our deepest—and in some sense also our simplest—normative expectations. They only ever relate to us, our values and practices, as *outsiders*: they lack the social intelligence and empathy—the human touch—that would equip them to reliably identify or fully satisfy our normative expectations. For all the value they bring, the proliferation of AI systems increases the risk of hermeneutic harm. This merits attention.

Author Contributions All authors contributed (planning, writing, editing) to all sections.

Funding This work was supported by the Belgian Defence– Royal Higher Institute of Defence (grant number HFM 22 − 03).

Declarations

Ethics Approval and Consent to Participate Not applicable.

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Allais, L. (2008). Wiping the slate clean: The heart of forgiveness. *Philosophy & Public Affairs*, *36*(1), 33–68. https://doi.org/10.1111/j.1088-4963.2008.00123.x

Baron, S. (2023). Explainable AI and causal understanding: Counterfactual approaches considered. *Minds and Machines*, *33*(2), 347–377. https://doi.org/10.1007/s11023-023-09637-x

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Borrego-Díaz, J., & Galán-Páez, J. (2022). Explainable artificial intelligence in data science: From foundational issues towards Socio-technical considerations. *Minds and Machines*, *32*(3), 485–531. https://doi.org/10.1007/s11023-022-09603-z

Buijsman, S. (2022). Defining explanation and explanatory depth in XAI. *Minds and Machines*, *32*(3), 563–584. https://doi.org/10.1007/s11023-022-09607-9

Christian, B. (2020). *The alignment problem: Machine learning and human values* (First edition). W.W. Norton & Company.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *AI & SOCIETY*, *24*(2), 181–189. https://doi.org/10.1007/s00146-009-0208-3

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Coghlan, S. (2024). Anthropomorphizing machines: Reality or popular myth? *Minds and Machines*, *34*(3), 25. https://doi.org/10.1007/s11023-024-09686-w

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Dastin, J. (2018). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Davis, C. G., Wortman, C. B., Lehman, D. R., & Silver, R. C. (2010). Searching for meaning in loss: Are clinical assumptions correct? *Death Studies*, *24*(6), 497–540. https://doi.org/10.1080/07481180050121471

de Graaf, M. M. A., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should Too). *Artificial Intelligence for Human-Robot Interaction AAAI Technical Report*, *FS-17-01*, 19–26.

Dennett, D. C. (1998). *The intentional stance*. MIT Press.

Fricker, M. (2009). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Fricker, M. (2017). Evolving concepts of epistemic injustice. In I. J. Kidd, J. Medina, & G. Pohlhaus Jr. (Eds.), *The Routledge handbook of epistemic injustice* (pp. 53–60). Routledge.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Hatherall, L., & Sethi, N. (2024). Regulating for trustworthy autonomous systems: Exploring stakeholder perspectives on answerability. *Journal of Law and Society*, *51*(4), 586–609. https://doi.org/10.1111/jols.12501

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81. https://doi.org/10.1037/0033-2909.107.1.65

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: From responsibility gaps to control gaps. *Synthese*, *201*(1), 21. https://doi.org/10.1007/s11229-022-04001-5

Holland Michel, A. (2020). *The black box, unlocked: Predictability and understandability in military AI*. United Nations Institute. https://doi.org/10.37559/SecTec/20/AI1. for Disarmament Research.

Keil, F. C. (2006). Explanation and Understanding. *Annual Review of Psychology*, *57*(1), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Lam, N. (2022). Explanations in AI as claims of Tacit knowledge. *Minds and Machines*, *32*(1), 135–158. https://doi.org/10.1007/s11023-021-09588-1

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)?– A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Lauwaert, L. (2021). Artificial intelligence and responsibility. *AI & Society*, *36*(3), 1001–1009. https://doi.org/10.1007/s00146-020-01119-3

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. https://doi.org/10.1145/3236386.3241340

Maclure, J. (2021). AI, explainability and public reason: The argument from the limitations of the human Mind. *Minds and Machines*, *31*(3), 421–438. https://doi.org/10.1007/s11023-021-09570-x

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McMahon, A., Buyx, A., & Prainsack, B. (2019). The Role of Harm Mitigation in the Governance of Data Use in Medicine and Beyond. *Medical Law Review*, fwz016. https://doi.org/10.1093/medlaw/fwz016. Big Data Governance Needs More Collective Responsibility:.

Medina, J. (2017). Varieties of hermeneutical injustice. In I. J. Kidd, J. Medina, & G. Pohlhaus Jr. (Eds.), *The Routledge handbook of epistemic injustice* (pp. 41–52). Routledge.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. https://doi.org/10.1145/3287560.3287574

Nadibaidze, A., Bode, I., & Zhang, Q. (2024). *AI in military decision support systems: A review of developments and debates*. Center for War Studies.

Neimeyer, R. A. (2016). Meaning reconstruction in the wake of loss: Evolution of a research program. *Behaviour Change*, *33*(2), 65–79. https://doi.org/10.1017/bec.2016.4

Nyholm, S. (2018a). Attributing agency to automated systems: Reflections on Human–Robot collaborations and Responsibility-Loci. *Science and Engineering Ethics*, *24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), e12507. https://doi.org/10.1111/phc3.12507

O'Neil, C. (2017). *Weapons of math destruction*. Penguin Books.

Panel of Experts on Libya (2021). *Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011)* (No. S/2021/229; p. 548). United Nations. https://documents.un.org/doc/undoc/gen/n21/037/72/pdf/n2103772.pdf

Park, C. L. (2010). Making sense of the meaning literature: An integrative review of meaning making and its effects on adjustment to stressful life events. *Psychological Bulletin*, *136*(2), 257–301. https://doi.org/10.1037/a0018301

PohlhausJr, G. (2017). Varieties of epistemic injustice. In I. J. Kidd, J. Medina, & G. Pohlhaus Jr. (Eds.), *The Routledge handbook of epistemic injustice* (pp. 13–26). Routledge.

Rebera, A. P. (2024). Reactive attitudes and AI-Agents– Making sense of responsibility and control gaps. *Philosophy & Technology*, *37*(4), 126. https://doi.org/10.1007/s13347-024-00808-x

Russell, S. J. (2023). *Human compatible: Artificial intelligence and the problem of control*. Penguin Books. Reprinted with an afterword.

Schwarz, E. (2021). Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control. *The Philosophical Journal of Conflict and Violence*, *5*(1). https://doi.org/10.22618/TP.PJCV.20215.1.139004

Sliwa, P. (2023). Making sense of things: Moral inquiry as hermeneutical inquiry. *Philosophy and Phenomenological Research*, *phpr.13028*. https://doi.org/10.1111/phpr.13028

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Stretesky, P. B., Shelley, T. O., Hogan, M. J., & Unnithan, N. P. (2010). Sense-making and secondary victimization among unsolved homicide co-victims. *Journal of Criminal Justice*, *38*(5), 880–888. https://doi.org/10.1016/j.jcrimjus.2010.06.003

Taylor, S. E. (1983). Adjustment to threatening events: A theory of cognitive adaptation. *American Psychologist*, *38*, 1161–1171.

Tigard, D. W. (2021). There is no Techno-Responsibility gap. *Philosophy & Technology*, *34*(3), 589–607. https://doi.org/10.1007/s13347-020-00414-7

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887.

Zednik, C., & Boelsen, H. (2022). Scientific exploration and explainable artificial intelligence. *Minds and Machines*, *32*(1), 219–239. https://doi.org/10.1007/s11023-021-09583-6

Zhang, Y., Dai, Z., Zhang, L., Wang, Z., Chen, L., & Zhou, Y. (2020). Application of Artificial Intelligence in Military: From Projects View. *2020 6th International Conference on Big Data and Information Analytics (BigDIA)*, 113–116. https://doi.org/10.1109/BigDIA51454.2020.00026