Acta Cogitata: An Undergraduate Journal in Philosophy

Volume 11 Article 3

"The Algorithm Decides": Unintentional Agency Laundering & Explanation

Carson Johnston
University of Guelph

"The Algorithm Decides": Unintentional Agency Laundering & Explanation

Carson Johnston, University of Guelph

Abstract

In this paper I explore a situation under-explored by AI researchers where those who deploy decision-making algorithms unintentionally launder their moral agency to algorithms through anthropomorphic ascriptions of their underlying architecture. Often, this kind of agency laundering occurs rather innocently, by attempting to render an otherwise opaque system transparent through simplified and analogous explanations intended to enhance the decision subject's understanding. Consequently, when unintentional agency laundering happens, the decision subject's agency to seek recourse for adverse outcomes is undermined in the process of laundering the data controller's moral agency to a non-agent. This paper explores this situation as it pertains to traditional philosophical accounts of responsibility, explanation, and knowledge and engages in recent literature from AI ethics. The paper proposes that explanation can be a mechanism that closes responsibility gaps in AI. However, only if explanations do not invoke unintentional agency laundering.

From the work of Stephen Darwall, it can be granted that "[w]hen we hold someone accountable, we implicitly address a claim or demand on his conduct" (Darwall, 2013, p.22). We do so implicitly or explicitly, through specific moral emotions termed "reactive attitudes" by Strawson (Strawson, 1968; Darwall, 2013, p.22). These are emotions such as "resent" or "remorse" that allow someone to hold another accountable or to hold themselves accountable. Moreover, from Darwall we can grant that "when we hold someone morally responsible, we commit ourselves also to the idea that he has the very same competence and authority to hold himself (and us) responsible also" (Darwall, 2013, p.22).

Darwall and Strawson's work on accountability has gone largely unappreciated in the Ethics of Artificial Intelligence (AI). In this paper, I apply some of their central claims to a situation currently under-explored by AI researchers where those who deploy decision-making algorithms sometimes unintentionally launder their moral agency to algorithms through anthropomorphic ascriptions of their underlying architecture. Often, this kind of agency laundering occurs rather innocently, by attempting to render an otherwise opaque system transparent through simplified and analogous explanations intended to enhance the decision subject's understanding. Consequently, when unintentional agency laundering happens, the decision subject's ability to seek recourse for adverse outcomes is undermined in the process of laundering the data controller's moral agency to a non-agent. This situation might occur as follows.

Francis is a data controller, and he deploys AlgorithmAlpha that produces an adverse outcome for the decision subject Cindy. In virtue of having a right to explanation, Cindy demands an explanation for *why* the adverse outcome occurred from Francis. Francis gives Cindy an explanation of Type A. Explanations of Type A often inadvertently anthropomorphize algorithms, so, Francis' explanation accidentally anthropomorphizes AlgorithmAlpha. Because AlgorithmAlpha is anthropomorphized, AlgorithmAlpha could be regarded as the individual with whom Cindy directs her reactive attitudes. This would make AlgorithmAlpha accountable to provide a de facto explanation to Cindy if she contests Francis' original explanation and assumes AlgorithmAlpha has humanlike agency. A de facto explanation would be one that provides a justification for the initial explanation. However, AlgorithmAlpha lacks the same competence as Cindy to be able to hold 'himself' accountable to her (Darwall, 2013, p.22). Consequently, Cindy is unable to contest the decision reached by AlgorithmAlpha now that her demands are directed toward 'him'.

The aim of this paper is to unpack and understand this situation. The paper has four sections. In Section One, I introduce the concept of agency laundering as well as some of its core commitments to distinguish between two forms, (1) intentional and (2) unintentional. In Section Two, I elaborate how unintentional agency laundering arises practically. Section Three accounts for the dangers of unintentional agency laundering. Finally, in Section Four, I explain how the dangers of unintentional agency laundering negatively impact a decision subject's temporally extended agency, thus expanding the current literature on rights to explanation. Taken together, these support the following thesis: unintentional agency laundering via anthropomorphizing an algorithm falsely holds a non-agent accountable thereby undermining a decision subject's ability to contest adverse outcomes produced by the algorithm.

Section One: Agency Laundering

In the past decade, decision-making algorithms have been fruitfully deployed in high-stakes decision-making processes such as predicting rates of recidivism, credit scoring, advertising, and medical diagnosis because they often add a level of accuracy that surpasses that of a human (Angwin & Larson, 2016; Adomavicius & Yang, 2022; Binns, 2018). However, it is common in these processes for those affected by algorithmic decisions to have an indirect relationship with the people who build and/or deploy the algorithms. That said, when an adverse outcome occurs (for example, predicted high rates of recidivism, denial of a loan, unusual medical diagnosis, etc.) these individuals (henceforth, *decision subjects*) are reasonably owed justifications for *why* such adverse outcomes were produced (Binns, 2018; Goodman & Flaxman, 2017). Although the content and formulation of a justification may differ, recent regulations like the European Union's "General Data Protection Regulation" and Canada's "Directive on Automated Decision-Making" suggest that a 'right to demand an explanation' implies that there is an agent morally responsible to give a justification for the outcome that resulted.

An important paper by Rubel, Castro, and Pham (2019) presents a unique contribution to this discussion. Much of the literature at the time of publication concerns the agency and ethics pertaining to decision subjects. Rubel, Castro, and Pham (2019) however, centered their paper on those who design, control, or deploy the algorithms. For the rest of the paper, I will refer to these individuals generally as *data controllers.*[1]

Rubel, Castro, and Pham (2019) define agency laundering as "obfuscating one's moral responsibility by enlisting a technology or process to take some action and letting it forestall others from demanding an account for bad outcomes that result" (Rubel, Castro, & Pham, 2019, p.1018).[2] Rubel, Castro, and Pham (2019) ground their conception of moral responsibility in the influential work of Strawson, understanding that for someone to be responsible at all, there must be someone to hold them responsible and one does this by forming appropriate reactive attitudes towards the morally responsible agent (Rubel, et al., 2019, p.1020). Their views on responsibility can be extended outwards to Darwall's work on second-personal ethics that recognizes a certain cognitive capability (second-personal competence) needed to hold someone responsible in a Strawsonian-like moral relationship (Isern-mas & Gomila, 2020). Rubel, Castro, and Pham (2019) would grant that decision-making algorithms are not entities with which second-personal competence naturally applies and so would not be the subject of one's reactive attitudes (Rubel et al., 2019, p.1020). Furthermore, they suggest that insofar as data controllers are responsible, by way of decision subjects forming appropriate reactive attitudes towards them, it is a facet of this responsibility that data controllers also ought to grant decision subjects an account of their reasons (Rubel et al., 2019, p.1020). Consistent with Rubel, Castro, and Pham's (2019) definition of agency laundering, the goal of agency laundering is to obscure or misdirect the source of responsibility that would result in any demands or accounts to be directed towards someone else (Rubel et al. 2019, p.1022).

Rubel, Castro, and Pham (2019) rely on the views of Angela Smith (2012) and Marina Oshana (1997) to ground their understanding of accountability. For Angela Smith, to be accountable is to be open to give a justification (Smith, 2012, p.577-78; Rubel et al., 2019, p.1021). For Marina Oshana, to be accountable is for it to be appropriate for that individual to explain their reasons (Oshana 1997, p.77; Rubel et al., 2019, p.1021). Shoemaker (2011) takes accountability one step further by tying it back to the relationship that makes one accountable from which he understands accountability as embedded in responsibility. Specifically, Shoemaker (2011) refers to this as accountability-responsibility. For Shoemaker, someone is accountability-responsible when "one is susceptible for being held to account just in case one has the capacity to recognize and appreciate the demands defining the various relationships as

[1] I use "data controller" throughout this paper to signify any human individual or group that would be morally responsible for the outcome of a decision algorithm and therefore, accountable to deliver an explanation to a decision subject or a group of decision subjects. This could include those who design, review, deploy, etc.
[2] The authors note that the concept of agency laundering is not unique to the domain of artificial intelligence and may be applicable to many different domains.

reason-giving" (Shoemaker, 2011, p.631). Regardless of which accountability view one subscribes to, accountability follows from moral responsibility as understood in a Strawsonian-like moral relationship and pre-supposes a certain cognitive capacity of those involved in the moral relationship.

In understanding these basic components at play in the definition of agency laundering, two defining features come to light. The first relates to the notion of *agency* whereby the acting agent intentionally misdirects their involvement in the decision-making process to some other person, process, or entity. The second relates to the notion of *laundering* and "cuts straight to the heart of *what responsibility is* by undermining the ability of others to ask the agent to provide an account" (Rubel et al., 2019, p.1024). For instance, in the case of a data controller deploying a decision-making algorithm, they may launder their agency by "invoking the complexity or automated nature of a decision system to explain an outcome" (Rubel et al., 2019, p.1021). Doing so then allows the decision subjects to assume that the data controller is not morally responsible for the algorithm's outcome, forestalling the decision subject from contesting adverse outcomes.

Rubel, Castro, and Pham (2019) would consider this an example of agency laundering whereby a responsibility gap is invoked as a mechanism for the data controller to intentionally distance himself from the outcome produced by an algorithm or misdirect a decision subjects' reactive attitudes towards a false-responsible agent (Rubel et al., 2019, p.1035). It could be the case that there *is* a responsibility gap, but Rubel, Castro, and Pham (2019) leave open the possibility of genuine responsibility gaps. In citing Matthias (2004), they illuminate that some automated decision-making processes are so complex that it is nearly impossible to determine *who* is responsible for adverse outcomes (Rubel et al., 2019, p.1034).[3] However, Rubel, Castro, and Pham (2019) note that the kind of responsibility involved in a responsibility gap pertains to causal responsibility and cannot fully explain away the possibility of other forms pertaining to a capable and accountable agent (Rubel, et al., 2019, p.1035).

Unfortunately, Rubel, Castro, and Pham (2019) are unable to establish whether agency laundering must always be intentional or if it is possible for agency laundering to be unintentional (Rubel et al. 2019, p.1039). In what follows, I present a novel contribution to this discussion by arguing for the possibility of unintentional agency laundering. The next section discusses how this kind of agency laundering takes shape through a specific form of explanation, *Everyday Explanation*, as it pertains to the scenario involving the data controller (Francis) and the decision subject (Cindy) from the introduction.

[3] I elaborate on the problems of opacity in Section Four, and further touch on responsibility gaps in the conclusion of the paper.

Section Two: How Explanation Invokes Unintentional Agency Laundering

The ease with which humans offer, and judge explanations is extraordinary (Selbst & Barocas, 2018). But, insofar as explanation is a natural, ubiquitous phenomena, it is neither distinct in form nor easily exercised in certain contexts (Selbst & Barocas, 2018; Miller, 2019; Pacer & Lombrozo, 2017). There are many kinds of explanation, methods for evaluating the quality of an explanation, and reasons for giving one kind over another that permeate the domains of philosophy, psychology, and cognitive science (Miller, 2019). While I believe that AI researchers have much to learn from the study of explanation in these domains, this paper is concerned with a single kind of explanation: Everyday Explanation (EE) because it can give rise to unintentional agency laundering. EEs are "explanations of why particular facts (events, properties, decisions, etc.) occurred, rather than explanations of more general relationships, such as those seen in scientific explanation" (Miller, 2019, p.3). As per the introductory scenario, EEs would be a Type A explanation.

EEs are intuitively appealing because they adhere to an empirically based and often-shared stipulation that explanations offering minimal complexities ought to be preferred (Selbst & Barocas, 2018; Pacer & Lombrozo, 2017; Zerilli, Knott, Maclaurin, Gavaghan, 2019). In the domain of AI, explanatory limitations often pertain to the black-box nature or sheer complexity of the underlying architecture (Burrell, 2016; Selbst & Barocas, 2018; Zednik, 2021). Thus, in situations where decision subjects have limited relevant knowledge, simplified explanations like EEs can be used.

Often, EEs will rely on folk-theoretic concepts and belief-desire psychology (Lombrozo & Carey, 2006; Miller, 2019; Zerilli et al., 2019). That is, they typically make use of 'everyday' terms that illuminate the ways in which *we believe we think* (Miller, 2019). In this way, EEs invoke a simplicity that should be understood like Ockham's razor, where unnecessary entities (for example, explanation about an algorithm's architecture) are excluded (Pacer & Lombrozo, 2017). Alternatively, one may consider this along the lines of Daniel Dennett's "Intentional Stance" (1996) in which any considerations of design and implementation are eschewed from the explanation (Zerilli et al., 2019). Thus, when explaining the causes of an algorithm's outcome to a person or group lacking the right background knowledge, or when the mechanisms by which an algorithm produces its outcome are unknown, these simple-layman's explanations can be used (Zerilli et al., 2019). In the situation involving Francis and Cindy, Francis need only give Cindy that which he thinks is necessary for her understanding of a certain fact or decision. I will explain why this might be problematic despite its intuitive appeal for explaining complex information technologies.

Particularly, the specific language used to substitute complexities in EEs is a slippery slope because it can give rise to unintentional agency laundering when anthropomorphic ascriptions and qualities of agency are assigned to a technological process or algorithm such as, with words like *learned, chose, decided, etc.* The following passage uses words such as "decides" and "guides" in such a way that incidentally places the algorithm in contact with decision subjects such as medical professionals and job applicants *as if it were an acting agent*.

> "…when you apply for a loan, algorithms increasingly make mortgage approval decisions. If you apply for a job, resume-screening algorithms decide whom to invite for an interview… In medicine, we're moving towards personalized medicine. Two people with the same symptoms might not get the same treatment… Algorithms guide doctors on those decisions." (Knowledge at Wharton, 2019).

When explanations like this are given, the data controller or decision subject might misinterpret the agential responsibility of the algorithm. This makes it possible for a decision subject to accidentally view the algorithm as having human-like agency over the decision it produces and misplaces the data controller's moral responsibility. However, prior to a full account of this kind of error via anthropomorphism, it is first important to grasp the different ways in which a technology or algorithm may be anthropomorphized.

Anthropomorphism & Evolution

Anthropomorphizing non-human entities is a natural constituent of our psychology and has been evolutionarily adaptive as it cues us to potential dangers and makes us receptive to respond appropriately (Boyer, 2007). It also aids in fostering trust, as Guthrie (1993) explains, humans have a natural tendency to anthropomorphize even the simplest of cues. "That is, we tend to interpret even very faint cues in terms of human traits; we see faces in the clouds and human bodies in trees and mountains" (Guthrie, as cited in Boyer 2007, p.144). When applied to technology, the ways in which we design technology, how we interact with it, and how we describe it, are all avenues through which this technology might be anthropomorphized.[4]

As such, the first way humans might anthropomorphize technology is through various design considerations. These concern variables such as names, personalities, voices, scripts, body, and movement. Scorici, Schultz, and Seele (2022) conceptualize this as "humanwashing" that lets "observers believe in unrealistic robot capacities (e.g., artificial general intelligence) or distract observers from the true capacities that a robot may perform (e.g., military use cases)" (Scorici, Schultz, & Seele, 2022, p.2). A chatbot might be anthropomorphized on all except body and movement whereas a children's toy robot may be anthropomorphized on all dimensions.

[4] It should be noted that anthropomorphizing technology is not limited to these three avenues. There might be other possible ways to anthropomorphize technology and reasons for doing so that go beyond the scope of this paper.

The second way in which a technology may be anthropomorphized is through the actual interaction we have with the technology. For instance, interfacing with the technology or system in such a way that we attribute to it the human characteristic of "mindedness" (Hull, 2022). We do this when we engage in conversation or dialogue with a chatbot, play with a toy robot, or talk to our smart home devices.

The third avenue for anthropomorphism then pertains to how we explain the technology ex post by substituting technical complexities concerning the design of the technology with folk-theoretic concepts and belief-desire psychology in the form of 'Everyday Explanations'. We attribute a sort of "mindedness" to an algorithm or technology by equating our own mental states with that which we elect 'it' to have in an explanation. As Bender suggests, "[t]he terminology used with large language models, like "learning" or even "neural nets," creates a false analogy to the human brain" (Bender as cited in Hull, 2022). It is this avenue that can invoke unintentional agency laundering. This occurs in direct consequence of drawing false analogies between human attributes and those the algorithm has. It should be emphasized that the ways in which an algorithm decides can be fundamentally different than that of humans, despite our intuitive tendencies to the contrary (Burrell, 2016).

In fact, in some cases, it is precisely because algorithms process information differently from humans that they are such powerful and extraordinary decision-making tools. Adversely, the downfall of confounding our understanding via substituting technical complexities with analogous and everyday terms, is that it prevents us from giving a correct technical explanation. Though, even if it were not our goal to give a correct technical explanation, this kind of explanation—everyday explanation—muddles our understanding of the agential responsibility owed to an algorithm. Thus, if what we are truly targeting in demands for explanations or in the explanations themselves is accountability, then what is needed is an understanding of the structure of moral relationships, which can *only* be between humans and *not* between decision subjects and algorithms.

The Structure of Moral Relationships & Accountability

The structure of moral relationships can be understood from a morality-as-accountability standpoint. A morality-as-accountability standpoint views a moral relationship in a Strawsonian sense (De Kenessey & Darwall, 2014). As Darwall puts forth, these relationships are bipolar in nature between the obligor and the obligee (Darwall, 2013). Or, perhaps between the accountable person(s) and affected person(s) in which the affected person holds the other accountable through forming reactive attitudes towards that individual. Recalling Francis and Cindy, Francis takes the role of the accountable and Cindy that of the affected. The accountable individual then has the responsibility to legitimately hold themselves accountable. A necessary feature in this is that both individuals must view the other as a participant in a relationship and by doing so, accept some additional propositions about the other in the sense that they possess the capability or capacity to appreciate the demands of this relationship (Isern-mas & Gomila, 2020; Shoemaker, 2011). As such, it is a key insight of any accountability view that members of the moral community "identify not only who is morally responsible but what that responsibility involves" (Rubel et. al. 2019, p.1021).

Yet, an anthropomorphized algorithm like AlgorithmAlpha can eschew the source of accountability in such a way that a decision subject might view it as accountable for delivering a de facto explanation (the reasons for an adverse outcome demanded by a decision subject if not initially satisfied). As a result, Cindy's reactive attitudes become directed towards AlgorithmAlpha (a false-responsible agent) as opposed to Francis (a true-responsible agent) owing the de facto explanation. In the next section, I elaborate on the dangers of agency laundering when accountability-responsibility is granted to a false-responsible agent.

Section Three: The Dangers of Unintentional Agency Laundering

So far it is understood that in cases like that of Francis and Cindy, unintentional agency laundering occurs when an algorithm is granted accountability-responsibility for adverse outcomes. Recalling Shoemaker's definition of accountability-responsibility, someone is accountability-responsible when "one is susceptible for being held to account just in case one has the capacity to recognize and appreciate the demands defining the various relationships as reason-giving" (Shoemaker 2011, p.144). However, only an algorithm post-anthropomorphism would *appear* to be accountability responsible. An anthropomorphized algorithm does not *actually* possess the proper capacity to be a moral agent in the sense of having second-personal competence. Rather, the algorithm is only falsely granted this capacity through anthropomorphic ascriptions of the underlying technology as discussed in Section Two.

Thus, it is evident that anthropomorphized algorithms would not adhere to the qualifications of accountability-responsibility. However, when unintentional agency laundering occurs, accountability-responsibility *is* taken away from the true-responsible agent (the human) and directed to the false-responsible agent (the algorithm) making the Strawsonian moral relationship one between the decision subject and the algorithm. Unintentional agency laundering is then the mechanism by which an algorithm is regarded as a real, agential participant in this bipolar partnership with the decision subject.

The resulting danger is what Kurt Gray and colleagues call an "algorithmic outrage deficit." In a recent study, Gray et al. (2022) found that moral outrage towards humans and algorithms was asymmetrical. It seemed to be the case that participants experienced less moral outrage about discrimination when this discrimination was perpetrated by an algorithm versus a human or corporation. This was typically on account that participants viewed the algorithms as less motivated by prejudice given that they lack a "mind" and therefore, are not equally subject to human bias.

However, they also manipulated the anthropomorphism assigned to the algorithms. They assigned participants to a high or low anthropomorphism condition. In the high condition, they used phrases such as "just like a human being", "chooses based on opinions it has formed" and ascriptions such as has "tastes", "feelings", "thinks", "likes", and so on (Gray et al., 2022). Gray et al. (2022) found increased outrage towards algorithms under the high-anthropomorphism condition in part with beliefs that these algorithms were more motivated by prejudice. However, overall results yielded those participants were less outraged at a highly anthropomorphized algorithm than a human. Extrapolated from this study is the possibility that an algorithmic outrage deficit negatively affects the agency typically attributed to decision subjects in virtue of having a right to explanation. In the final section I elaborate how an "algorithm outrage deficit" might affect a decision subject's ability to carry out their right to explanation.

Section Four: Temporally Extended Agency & Avoiding and Algorithmic Outrage Deficit

Characteristically, most adults exercise *temporally extended agency* (Venkatasubramanian & Alfano, 2020). Temporally extended agency adheres to the assumption that society is organized in such a way that humans will engage in planning to consecutively achieve short-term goals in the fulfilment of long-term goals in the grand scheme of one's overall desired life trajectory (Venkatasubramanian & Alfano, 2020, p.284). Thus, when something occurs that prevents us from actualizing plans and goals, we may desire recourse.

Venkatasubramanian and Alfano (2020) define algorithmic recourse as "the systematic process of reversing unfavorable decisions by algorithms and bureaucracies across a range of counterfactual scenarios" (Venkatasubramanian & Alfano, 2020, p.284). In their view, recourse is a modally robust good (Venkatasubramanian & Alfano, 2020, p.285). As such, when an algorithm produces an adverse outcome, a decision subject can benefit from this good by choosing to enact their right to explanation.

However, this does not mean that recourse is dependent on enacting a right to explanation. Rather, recourse depends on if a decision subject was provided an explanation of certain quality and rigour. For instance, receiving an explanation that does not launder the data controller's agency, so that the decision subject can contest the explanation. In this way, denied recourse should not be considered as undermining the decision subject's ability to enact their temporally extended agency in the same sense that unintentional agency laundering undermines this. Instead, not receiving recourse but receiving a proper explanation about why the adverse outcome occurred that is justifiably *not* reversed, still allows the decision subject to gain knowledge of *how* they can avoid this bad outcome in the future, thereby appreciating what it means to have temporally extended agency.[5]

That said, genuinely achieving recourse after a demand has been made to a data controller is complicated by the opacity of a decision-making system. Burrell (2016) identifies three such kinds of opacity. Only the third kind of opacity identified by Burrell is relevant for this paper. Opacity of this form stems from the technical complexity of the system in question and "the demands of human scale reasoning and styles of semantic interpretation" (Burrell, 2016, p.2). For decision subjects, this kind of opacity can be interpreted as a lack of understanding or knowledge of the systems' epistemically relevant elements (EREs; Zednik, 2021). These should be taken as the components, properties, objects, etc. of a system that an agent needs to know or understand to reach their optimal level of transparency. By optimal level of transparency, I mean that which would be sufficient for the decision subject to contest an adverse outcome.

This is needed for the data controller to overcome opacity and avoid the dangers of unintentional agency laundering. Thus, as it pertains to a decision subject's right to an explanation it is not merely enough that they be given an explanation generally. Rather, they morally ought to be given an explanation that adheres to certain standards of quality and rigour. Evidently, this is not an easy thing to do considering existing explanatory limitations pertaining to the complexity of algorithmic decision-making. However, it should be the task of AI researchers and philosophers going forward to formulate possible standards to hold explanations given to decision subjects. Especially, considering that algorithms are deployed for high stakes decisions that presently have a disparate impact on marginalized groups.

Conclusion:

This brings us full circle to role of responsibility gaps in explanation and the ways in which explanation can be a mechanism for closing the responsibility gap. Through explanation, a data controller like Francis can render an otherwise opaque system transparent by focusing solely on what is epistemically relevant for the decision subject. However, *only if* the explanation adheres to a certain quality. In future work, I hope to explore possible standards for explanations in this context. Precisely, because unintentional agency laundering can arise when the language used in an Everyday Explanation anthropomorphizes a decision-making algorithm and falsely grants the algorithm accountability-responsibility. Thus, causing a decision subject to misinterpret the agential responsibility an algorithm has in producing adverse outcomes. The resulting danger is an algorithmic outrage deficit that undermines a decision subjects' ability to contest adverse outcomes as it relates to a decision subject's temporally extended agency. This shows that a right to explanation does not just demand that an explanation be given. Rather, that

[5] It is important here to note that inequalities of outcome will always exist. Thus, for those that do exist, it is necessary that they be justified, and we determine their justification through properly explaining the reasons why it occurred.

a *good* explanation must be given. Specifically, this would be one that does not unintentionally launder a data controller's moral agency as to avoid the dangers of unintentionally agency laundering.

References

Adomavicius, G., & Yang, M. (2022). Integrating Behavioral, Economic, and Technical Insights to Understand and Address Algorithmic Bias: A Human-Centric Perspective. ACM Transactions on Management Information Systems, 13(3), 1–27. https://doi.org/10.1145/3519420

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Machine bias. ProPublica. Retrieved December 12, 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. Journal of Experimental Psychology. General. https://doi.org/10.1037/xge0001250

Binns, R. (2018). Algorithmic Accountability and Public Reason. Philosophy & Technology, 31(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Boyer, P. (2007). *Religion Explained*. Basic Books.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1-12. https://doi.org/10.1177/2053951715622512

Darwall, S., & Dill, B. (2014). Moral Psychology as Accountability. In Moral Psychology and Human Agency. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198717812.003.0003

Darwall, S. (2013). 1 Morality's Distinctiveness. In Morality, Authority, and Law. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199662586.003.0001

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

Hull, G. (2022). Just because the computer talks to you about its feelings doesn't make it sentient. New APPS: Art, Politics, Philosophy, Science. https://www.newappsblog.com/2022/06/just-because-the-computer-talks-to-you-about-its-feelings-doesnt-make-it-sentient.html

Isern-Mas, C., & Gomila, A. (2020). Naturalizing Darwall's Second Person Standpoint. Integrative Physiological and Behavioral Science, 54(4), 785–804. https://doi.org/10.1007/s12124-020-09547-y

Knowledge at Wharton. (2019). Who Made That Decision: You or an Algorithm? Knowledge at Wharton. https://knowledge.wharton.upenn.edu/article/algorithms-decision-making/

Lipton, Z. C. (2016). The Mythos of Model Interpretability. https://doi.org/10.48550/arxiv.1606.03490

Liquin, E. G., Metz, S. E., & Lombrozo, T. (2020). Science demands explanation, religion tolerates mystery. Cognition, 204, 104398–104398. https://doi.org/10.1016/j.cognition.2020.104398

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. Cognition, 99(2), 167–204. https://doi.org/10.1016/j.cognition.2004.12.009

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

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

Oshana MAL (1997) Ascriptions of responsibility. Am Philos Q 34(1):71–83

Pacer, M., & Lombrozo, T. (2017). Ockham's Razor Cuts to the Root: Simplicity in Causal Explanation. Journal of Experimental Psychology. General, 146(12), 1761–1780. https://doi.org/10.1037/xge0000318

Rubel, A., Castro, C., & Pham, A. (2019). Agency Laundering and Information Technologies. Ethical Theory and Moral Practice, 22(4), 1017–1041. https://doi.org/10.1007/s10677-019-10030-w

Scorici, G., Schultz, M. D., & Seele, P. (2022). Anthropomorphization and beyond: conceptualizing humanwashing of AI-enabled machines. *AI & Society*. https://doi.org/10.1007/s00146-022-01492-1

Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. Ethics, 121(3), 602–632. https://doi.org/10.1086/659003

Selbst, A., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. Fordham Law Review, 87(3), 1085–.

Smith, A. M. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics*, *122*(3), 575–589. https://doi.org/10.1086/664752

Summers, J. S., & Sinnott-Armstrong, W. (2015). Scrupulous agents. Philosophical Psychology, 28(7), 947–966. https://doi.org/10.1080/09515089.2014.949005

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 284–293. https://doi.org/10.1145/3351095.3372876

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. Philosophy & Technology, 34(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? Philosophy & Technology, 32(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6