Agency Laundering and Information Technologies

Alan Rubel (arubel@wisc.edu), University of Wisconsin-Madison (corresponding author)
Clinton Castro (clinton.g.m.castro@gmail.com), Florida International University
Adam Pham (adamkpham@gmail.com), University of Wisconsin-Madison

Abstract:

When agents insert technological systems into their decision-making processes, they can obscure moral responsibility for the results. This can give rise to a distinct moral wrong, which we call "agency laundering." At root, agency laundering involves 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. We argue that the concept of agency laundering helps in understanding important moral problems in a number of recent cases involving automated, or algorithmic, decision-systems. We apply our conception of agency laundering to a series of examples, including Facebook's automated advertising suggestions, Uber's driver interfaces, algorithmic evaluation of K-12 teachers, and risk assessment in criminal sentencing. We distinguish agency laundering from several other critiques of information technology, including the so-called "responsibility gap," "bias laundering," and masking.

1 Introduction

There have been numerous examples of automated decision systems going wrong in consequential ways.[1] In 2018 an Uber automated driving system failed to recognize a bicyclist, whom it struck and killed (Levin and Wong 2018). In 2012, the Target corporation received international attention when, based on predictive analytics and an automated advertising system, it sent fliers targeting women seeking prenatal products to a minor before she had revealed her pregnancy to one of her parents (Duhigg 2012).[2] In 2017, the news organization ProPublica was able to use Facebook's automated system to make an ad buy targeting users with anti-Semitic affiliations (Angwin and Varner 2017). The system even suggested additional racist categories to make the ad purchase more effective. In the criminal justice system, risk assessment algorithms are used in making important decisions, but have very different results depending on the race and ethnicity of defendants (Angwin and Larson

[1] We presented a paper discussing the idea of agency laundering at the 2019 iConference (March 31 – April 4, 2019, Washington, D.C.) which appears in the conference proceedings (Rubel et al. 2019). The account of agency laundering in this paper is expanded and has changed substantially from the account in those proceedings. Most importantly, the conference paper did not tie agency laundering to an account of responsibility, and instead invoked concepts of de facto and de jure power. The conference paper did not include many of the examples used here (e.g., Democratic Chair, Uber, *HISD*). The conference paper did discuss the Facebook and *Loomis* examples, but the treatment has changed in accord with the changes to the account of laundering. We would like to thank audiences at Eindhoven Technological University, Delft Technological University, Twente Technological University, the Zicklin Center for Normative Business Ethics at the Warton School of Business, University of Pennsylvania, and the Privacy Law Scholars Conference. In particular, we would like to thank Richard Warner, Filippo Santoni de Sio, Sven Nyholm, Owen King, Philip Jansen, and three anonymous reviewers for their careful consideration and enormously helpful comments.

[2] Note that a number of commentators believe the story makes too close a connection between predictive analytics and pregnancy-related advertising. There are reasons to send such advertising to people who are not pregnant, the advertising may have been based on a criteria unrelated to pregnancy, and others. (Harford 2014)

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

2016). A common element in these stories is that the technology itself plays an important role. The existence and use of technological systems are a key part of the explanation of the events. Whether (and how) the technologies are relevant in assessing moral responsibility is considerably more complex.

Much of the literature on ethics in big data and automated decision systems examines how such systems harm data subjects (Eubanks 2018; O'Neil 2016), reflect and engender discrimination (Noble 2018; Barocas and Selbst 2016; Angwin and Larson 2016; Sweeney 2013; Citron 2008), and lack transparency (Pasquale 2015). Some commentators link wrongs in such systems to failures of respect for the agency and autonomy of data subjects (Mittelstadt et al. 2016). This paper departs from the literature in a key way. Whereas most of this literature focuses on those who are subject to information systems (as sources of data, subjects of algorithmic decision-systems, users of social media, etc.), this paper considers the moral agency of those who deploy information technologies (as collectors of big data, users of algorithmic decision-systems, developers of social media sites, and so on).

We will argue that a type of moral wrong that can arise in using automated decision tools is "agency laundering." At root, agency laundering involves 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.. Laundering is not unique to information technologies. However, we argue that the concept of agency laundering helps understand important moral problems in a number of recent cases involving automated, or algorithmic, decision-systems. The moral concerns are not merely that values are instantiated within automated systems. Instead, intermingling moral wrongs with morally permissibly processes undermines a fundamental facet of responsibility itself.

We begin, in section 2, with an account of responsibility to ground our arguments. In section 3, we develop our account of agency laundering and explain its moral salience. In sections 4-7 we offer several case studies that allow us to apply and further explain our conception. One is Facebook's targeted advertising system and its response to complaints that it allows users to make racist ad purchases. This is a clear case of agency laundering. Next, we consider Uber's use of algorithmic systems in its driver-management apps and show how we can distinguish cases of agency laundering from non-agency-laundering in structurally similar cases. We then turn to public-sector uses, showing how school districts can launder agency in teacher-evaluation cases, and how courts can avoid agency laundering by clarifying responsibility for decision systems. In section 8 we explain how agency laundering is distinct from other concepts, especially the "responsibility gap" (Matthias 2004).

2 Agency and Responsibility

Our argument turns on the concept of responsibility. For a person to launder her agency requires that she be a moral agent in the first place, and being a moral agent requires that one be in some sense morally responsible. In this section, we first distinguish several facets of responsibility and how they relate to one another (2.1). This helps structure our understanding of agency laundering in section 3. Then we offer a substantive account of an agent's responsibility (2.2). This will ground our understanding of the moral wrongs associated with agency laundering.

2.1 The Structure of Responsibility

In *Punishment and Responsibility*, HLA Hart describes a ship captain who gets drunk, wrecks his ship, is convicted of criminal negligence, and whose employer is held financially liable for the loss of life

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

and property (Hart 1968, 211). Hart's allegory and the distinctions he offers are useful in grounding our account.[3]

To begin, a person might be responsible in virtue of a role. In Hart's example, a person is responsible for a ship's safety in virtue the fact that she is the captain. A person's role requires her to anticipate events in some domain and to take actions to avoid bad outcomes in that domain.[4] A ship captain should anticipate bad weather and obstacles and plot course accordingly. Parents should anticipate their children's needs and plan ways to address them. Financial advisors should anticipate client needs and economic forecasts and guide clients' actions suitably. Although one's well-defined social roles (ship captain, parent, financial trustor) may give rise to specific responsibilities, the idea of role responsibility is broad enough to encompass general obligations one has as a moral agent. So, for example, adults have a responsibility to operate heavy machinery carefully, regardless of their specialized social roles; community members have a responsibility to pay applicable taxes; and people engaged in commerce have a responsibility to bargain in good faith.[5]

Second is *causal* responsibility, or the link between an agent's action (or omission, or disposition[6]) and an event that results from it. Chris Kutz calls this *explanatory* responsibility, as causation generally explains an event (Kutz 2004, 549). Any explanation of the ship wreck that ignores the captain and the captain's drinking would be inadequate.

Causal responsibility in this sense does not entail moral responsibility. That is because of the third facet, capacity responsibility, which relates to whether an agent has the requisite capacities to be responsible for an outcome. One may lack capacity responsibility due to pathology or pre-reflective, non-deliberative action. In Hart's example, it is possible that a ship captain's drinking was due to extreme, clinical anxiety, in which case her intoxication is something she caused, but for which she lacked the required capacity to be responsible.[7] Alternatively, one may lack capacity due to lack of access to relevant information. That is, an agent must be in a position to access certain facts about her actions and their significance in order to be retrospectively morally responsible for them.

We can sum up the structure so far as follows. For a person to be (retrospectively) morally responsible—which is to say *morally liable*—for some event or outcome, she must have some role responsibility (either a specific duty that attaches to a social role or a general duty as a moral agent) and she must be causally responsible for the outcome (which is to say an action of hers is a key part of the explanation of the outcome). Moreover, she must have capacity responsibility. That is, her action must not be the result of some pathology or pre-reflective

[3] Our account of the structure of responsibility follows closely those articulated by Nicole Vincent (2011) and Chris Kutz (2004). Both Vincent and Kutz recast Hart's ship captain case to distinguish various facets of responsibility.

[4] Antony Duff calls this "prospective" responsibility. (Duff 1998) Here we should note that we are only discussing morally justifiable roles, where the holders of role responsibility are themselves moral agents. Hence, being assigned a role within a criminal organization, or being assigned a role when one lacks the capacity to act morally, cannot confer role responsibility in the required sense.

[5] See (Vincent 2011; Williams 2008; Goodin 1986, 1987)

[6] For the sake of simplicity, we will refer to 'actions' in discussing responsibility. However, our account extends to omissions and dispositions. Note, too, that causal responsibility is complicated in over-determination cases. But those cases don't affect our analysis here.

[7] This, of course, may not absolve the captain completely. See Fischer and Ravizza 2000, 49–51 for an explanation of "tracing" responsibility to prior actions.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

action, and she must in some sense have access to relevant information.[8] For the remainder of the paper we use "moral responsibility" and "moral liability" interchangeably, and they will refer to this conjunction of role responsibility, causal responsibility, and capacity responsibility.

2.2 The Content of Responsibility

With these distinctions in mind, we can turn to the *content* of moral responsibility. In other words, once we have determined that an actor has some kind of role responsibility, is causally responsible for an outcome, and has the requisite capacity to be responsible, there is a further question about what this responsibility means. There are two key features of the view we endorse here. First, moral responsibility is fundamentally relational and grounded in social roles. Second, being morally responsible for some action means that one is accountable for, and should be able to provide an account of her reasons for, that action.

The view that moral responsibility is fundamentally relational owes a great deal to Peter Strawson's seminal article, "Freedom and Resentment" (Strawson 1962). Holding a person responsible by forming reactive attitudes about her (e.g., appreciation, admiration, disdain) is a feature of interpersonal relationships in which one regards the other as a participant. We might resent the captain for getting drunk and steering her ship onto the rocks, or we might admire her for her skill in guiding the ship to safety during a storm. However, we do not form such reactive attitudes towards entities that are not participants in relationships; resentment and admiration are not reasonable reactions to the actions of infants or machines. If an autopilot algorithm successfully steers the ship to safety, it would be appropriate to be impressed, baffled, or happy, but not to feel respect and admiration for the algorithm itself.

Despite these important insights, precisely what (if anything) justifies reactive attitudes is a further question. As Marina Oshana points out, the mere fact (if it is) that people are committed to the appropriateness of their reactive attitudes toward (some) people for (some of) their actions cannot suffice to explain why those reactions are appropriate. We do not call a person morally responsible just because others *regard her* as responsible. Rather, "we call a person an appropriate subject of reactive attitudes because the person *is* [morally] responsible" (Oshana 1997, 75 (emphasis added)).

While keeping in mind the important social function of responsibility attributions, our view aligns with the constellation of views for which an agent's moral responsibility turns on whether she is answerable or accountable for her actions. Angela Smith, for example, argues that for an agent to be morally responsible for something is for the agent to be "open, in principle, to demands for justification regarding that thing" (Smith 2012, 577–78).[9] And blame is in effect a

[8] There remain some controversial issues, including for example Frankfurt-style cases in which one may be responsible or not regardless of whether she does or does not know how her actions will be causally effective. But the issues in those cases turn on the link between causal responsibility and the ability to do otherwise. That does not affect our arguments.

[9] Within this group of views, there is substantial debate about whether person X is responsible for Y in virtue of Y being *attributable* to X, of X being *answerable* for Y, or of X being *accountable* for Y. Scanlon's view focuses on attributability (Scanlon 2008). Shoemaker distinguishes between attributability, answerability, and accountability (Shoemaker 2011). Smith (like Shoemaker) distinguishes a thing being attributable to a person and that person being responsible for it; however, she views accountability as a species of answerability. What is important for our purposes is that each of the views in this constellation recognizes that the content of responsibility claims is that responsible agents are those for whom it is appropriate, or for whom it ought to be the case, that they *provide an account* of their intentions, interests, and reasons for an action.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

demand that the agent "justify herself." Oshana's view is related. She articulates an accountability view according to which a person is responsible if, and only if, "it ought to be the case that the person account for her behavior." Giving such an account requires a person to provide a statement of her "beliefs or intentions" for her actions. "Thus," Oshana explains, "'X is accountable for Y' can be unpacked as 'It is appropriate that X explain her intentions in doing (or being) Y'" (Oshana 1997, 77).[10]

The key insight of the accountability views is that they identify not only who is morally responsible but what that responsibility involves. Specifically, it is justifiable to ask the responsible agent to account for their actions, omissions, or dispositions. She should be able to explain her intentions, reasons, and actions in terms that other relationship participants can understand.

3 Agency Laundering

With our discussion of responsibility in mind, we can return to the paper's central argument. Using an automated process to make decisions can allow a person to distance herself from morally suspect actions by attributing the decision to the system, thereby laundering her agency. Put slightly differently, invoking the complexity or automated nature of a decision system to explain an outcome allows a party to imply that the action is something for which she is not morally responsible.

Compare money laundering.[11] Where one has such large amounts of illicit cash that spending it or placing it into legitimate financial instruments would be suspicious, one can launder it by mingling it with other, legitimate streams of income so that the illicit cash appears legal. For example, one might add the illegal cash to money received in a legal, cash-dependent business.[12] The bad thing (income from illicit source) is hidden by the existence of some other, similar phenomenon. To be clear, we are not making an argument by analogy; decisions are not like cash. Rather, the point is that it is possible to obscure the source of responsibility for actions and make them appear unsuspicious by mingling them with other actions.

Consider a minor example ("Chair"). Suppose that Cheese State University vests department chairs with control over curriculum. A chair and several members of her department would like to get rid of phlogiston studies ("P-studies") because they think it is unimportant. The chair could do this unilaterally by removing courses, reassigning instructors, and altering degree requirements, but wants to avoid the wrath of the department phlogistologists. She therefore delegates curriculum decisions to a committee of people who she knows want to eliminate P-studies. When P-partisans complain, the chair responds that it was the committee's decision, though she knew from the beginning what that decision would be. By impaneling a committee to ensure the results she wanted, the chair obscures her role in the decision. The committee appears to be the relevant power, though it remained the chair.

---

[10] Fischer and Ravizza provide an accountability view that bridges Strawson's attention to the social function of holding others responsibility by way of reactive attitudes and accountability views' attention to reasons. Specifically, they maintain that an agent is responsible if she is an apt target of reactive attitudes. More important here, though, is that being morally responsible for actions requires that agents exercise "guidance control." That requires that agents be at least weakly reasons-responsive, which is to say that where the agent has access to strong reasons in favor or against an action, she will act in accordance with those reasons. It also requires that the source of actions be the agent, which is to say that the reason-responsiveness is internal to the agent (Fischer and Ravizza 2000, 31–41).

[11] 18 U.S. Code § 1956 - Laundering of monetary instruments

[12] Other aspects of money laundering are about concealing identities of agents, for example by routing illicit money through shell corporations and bank accounts in permissive jurisdictions.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

There are several features of Chair to address initially. First is that the chair had legitimate institutional authority to make the decision, and if she had moved to eliminate P studies unilaterally it would have happened. Her institutional authority is a form of role responsibility for her department's curriculum. She has the responsibility to anticipate educational needs, department resources, student demand, scholarly trends, and so forth, and to ensure that her department's offerings adequately address them. And her de facto power to alter the curriculum is a form of causal responsibility; when the curriculum changes, the chair's actions are an essential part of the explanation why.

Second, although the chair has power to make the decision, she draws in a separate body by giving the committee some degree of causal responsibility. Because the curriculum change would not occur without the committee's work, the committee is an essential part of the explanation for the curriculum change. It is not the only cause, as it is mixed with the chair's actions. Third, when the chair forms the committee, she implies it is neutral, would weigh evidence fairly, and might act in a way that the chair doesn't anticipate. But that's a ruse—ex hypothesi the chair knows that the committee will act just as she wishes.

Fourth, the chair's actions obscure her causal responsibility with respect to the curriculum. The chair is able to obscure the fact that she orchestrated the result by making the committee partially causally responsible (i.e., a key part of the explanation) for the result. Fifth, although the chair appears to fulfill her responsibility in shepherding the curriculum, her appointment of the committee obscures her designs to eliminate P-studies.

The following is a definition of agency laundering that incorporates these features of Chair. An agent (*a*) launders her agency where:

(1) *a* is morally responsible with respect to some domain *X*, AND
(2) *a* ensures that *b* (some process, person, or entity) has some causal responsibility with respect to *X*, AND
(3) *a* ascribes (implicitly or explicitly) morally relevant qualities to *b*'s actions (e.g., relevance, neutrality, reliability), AND
(4) In virtue of (2) and (3), *a* obscures the scope of her causal responsibility with respect to *X*, AND
(5) In virtue of (4), *a* fails to adequately account for events within *X* for which she is morally responsible.

This definition only gets us so far. It sets out the structure of agency laundering, which tracks and incorporates the structure of moral responsibility from section 2.1. However, it does not explain the moral problem of agency laundering itself (if there is one). That's our next task.

There are several ways in which the chair may have acted wrongly. One possibility is that it is unjustifiable to eliminate phlogistology in any case. But let's leave that aside, and assume that it's permissible to eliminate it based on its substance and the context. More important is that the chair's ascription of morally relevant qualities to the committee is misleading, and she has therefore deceived people about the process involved. Regardless of whether getting rid of P-studies is justifiable, the chair's obscuring her reasons and intentions in impaneling the committee do not appear justifiable. Others with whom the chair has a relationship have a claim to understand such an important facet of their professional lives.

A still deeper moral problem is that the chair's action allows her to avoid the core demand of responsibility, which is to provide an account. Regardless of whether she is meeting her role

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

responsibilities with respect to the curriculum, she is forestalling others' ability to demand an account for her actions within a domain of their legitimate concern.[13] This is the defining feature of agency laundering, and it turns on the substantive account of responsibility in section 2.2. There we explained that responsibility is first about social relations. We hold others responsible for their actions in part by forming reactive attitudes, and such reactive attitudes are key in understanding responsibility. However, our view is that moral responsibility is also a matter of whether agents are open to demands to justify their actions and whether it is appropriate for others to demand an account of their reasons and intentions.

Now we come full circle. Agency laundering involves a kind of misdirection (as in (2)-(4)). But, crucially, the misdirection undermines others' ability to demand reasons for an agent's actions. In other words, the *laundering* part of agency laundering cuts straight to the heart of *what responsibility is* by undermining the ability of others to ask the agent to provide an account.

Department members will be unable to ask the chair for her reasons and intentions in eliminating phlogistography, because the chair's actions look like formation of a committee that (apparently) deliberated about and then eliminated the subfield. Department members would reasonably believe that all the chair has to provide is an account of delegation to the committee. But an account that focused on the committee would *not* be an account of the chair's actual reasons and intentions, which are about engineering an outcome, not initiating a process to weigh things.[14]

It is worth explaining the role of condition (5) a bit further. What matters about (5) is that it distinguishes cases like Chair from structurally similar cases of delegation. Consider a variation in which the chair thinks P-studies should be eliminated, and she knows that there are so few P-sympathists that any full committee will have a majority of P-eliminationists. Nonetheless, she delegates the curriculum decision to a committee because of her commitment to inclusive, democratic department governance. As in Chair, (1)-(3) obtain. And (4) plausibly obtains, as the chair's causal responsibility in forming the committee may obscure her causal role in deciding to

[13] Two other accounts addressing causal and moral responsibility in the computing context are worth noting here. First, Daniel Dennett (1997) posits that machines may be credited with (i.e., responsible for) some tasks (e.g., Deep Blue beating Kasparov) but cannot be responsible for others (e.g., murdering Kasparov). We would argue that this difference tracks the causal/moral responsibility distinction, though that is not Dennett's claim.

Helen Nissenbaum (1994) argues that the increased use of computing systems poses a threat to accountability, based on four key barriers. These include the problem of many hands, the existence of bugs that cause computing failures, the ability to use computers as scapegoats, and the separation of system ownership from legal liability for problems. In doing so she notes that distributed causal responsibility can function to obscure responsibility and blameworthiness (p. 74). Our view of laundering can apply to each of the barriers she discusses, but does not depend on any of them. Consider the example of "blaming the computer," or pointing to the computer as the sole source of causal responsibility. That considered by itself would not seem to be a case of laundering, but instead just a straightforward denial of responsibility. If instead, it included a process by which a party ensures the computer has causal responsibility, ascribes morally relevant qualities to the computer's actions, obscures the party's causal responsibility, and in so doing fails to adequately account for events for which the party is morally responsible, it could be laundering. In other words, merely blaming something else does not rise to laundering. Laundering is, we take it, more insidious in that it forestalls others' abilities to demand an account of actions within domains of their legitimate concern.

Thanks to an anonymous reviewer for pointing us to these articles.

[14] Note that agency laundering does not require that one infringe one's substantive role responsibilities (except to the extent that one's role responsibility includes being transparent about one's causal responsibility). In Chair, for example, it's plausible that the chair was fulfilling her role responsibilities with respect to the department's curriculum. We return to this point in section 4.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

review P-studies. But the key difference is that the committee formation in "Democratic Chair" is not a sham, constructed so that the chair can avoid having to account for her actions regarding the curriculum. Just as in the original example, department members will reasonably believe that the action for which the chair should provide an account is the formation of the committee. But in democratic chair, that *is* the only action for which she should provide an account.

So, that's the account. Let's turn to some cases. These will help us understand how predictive, automated decision systems can launder agency.

4 Case 1: Facebook and Anti-Semitic Advertising

In 2017, ProPublica published a report detailing an investigation into Facebook targeted advertising practices (Angwin and Varner 2017). Using Facebook's automated system, the ProPublica team found a user-generated category called "Jew hater" with over 2200 members. While two thousand Facebook users choosing to identify as "Jew hater" in their profiles seems like a lot, Facebook's platform helpfully informed the ProPublica team that it was too small an audience for an effective ad buy. To help ProPublica find a larger audience (and hence have a better ad purchase), Facebook suggested a number of additional categories. For example, it suggested including the category "Second Amendment," presumably because of some overlap in users' choices of interests in their profiles. ProPublica used the platform to select other profiles displaying anti-Semitic categories, and Facebook approved ProPublica's ad with minor changes.

Facebook's platform also allows clients to target ads by excluding profiles by age, geographic, and race and ethnic categories. For example, advertisers can target users in specific places and income ranges while excluding people with specific "ethnic affinities." Many of these affiliations are generated automatically, based on content users and their friends have liked or shared (Angwin, Scheiber, and Tobin 2017). In some cases it is not the category that creates a problem, but the purpose of the ad. Targeting an ad by age makes sense in some contexts (life insurance, toys), but is discriminatory in others (job recruitment).[15]

When ProPublica revealed the anti-Semitic categories and other news outlets reported similarly odious categories (Oremus and Carey 2017), Facebook responded by explaining that algorithms had created the categories based on user responses to target fields (e.g., answers to questions about education and hobbies). It also pledged to address the issue. But Facebook was loath to claim it had responsibility. Chief Operating Officer Sheryl Sandberg claimed in a public response that "[w]e never intended or anticipated this functionality being used this way" (Sandberg 2017). That is no doubt true, though Facebook both wishes to sell advertising and employ as little labor as possible to monitor how that advertising functions.

Is it agency laundering? An agent (Facebook) launders its agency where:

(1) Facebook has moral responsibility with respect to targeted advertising, AND
(2) Facebook ensures that its algorithmic advertising process has some causal responsibility with respect to targeted advertising on its platform, AND
(3) Facebook ascribes morally relevant qualities to its algorithmic advertising process's actions, AND
(4) In virtue of (2) and (3), Facebook obscures the scope of its causal responsibility with respect to targeted advertising on its platform, AND

[15] Note that Facebook has recently taken measures aimed at reducing discriminatory advertising (Levin 2019).

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

(5) In virtue of (4) Facebook fails to adequately account for events within a domain for which it is morally responsible: specifically, the way in which its advertising platform helps target advertising to racists.

Each of these conditions appears to obtain. Certainly, Facebook has causal responsibility with respect to targeted advertising on its platform [Redacted]. A more difficult question is whether Facebook has role, or prospective, responsibility. The clearest sense in which they have role responsibility is that they have de jure authority over their platform, and they have a *general* responsibility to be good members of the broad community of people who use the platform. More specifically, they have (in our view) a *specific* responsibility to ensure that their platform does not facilitate racists to easily find an audience to whom they can advertise.

The claims that Facebook has such specific moral responsibilities will no doubt be controversial. Others may argue that Facebook has a moral responsibility to be a mere conduit of communication among members.[16] That is unconvincing for a couple of reasons. For one, this case is about advertising. Any claims about how Facebook should structure information between end users tells us nothing about Facebook's responsibility vis a vis advertisers. Moreover, Facebook already acts as if it has responsibilities with respect to both content and advertising. It has community standards, by which it judges and removes content, and it restricts certain kinds of advertising.[17] In any case, agency laundering only requires that Facebook have general responsibilities within this domain.

Facebook's categories are derived in part by automated systems. It takes a hands-off approach, letting users generate profile information, letting an algorithm pick out characteristics from user profiles, letting advertisers peruse those categories, and letting an algorithm suggest compatible categories to build better ad target groups. Thus, Facebook ensures that an algorithmic process has causal responsibility (i.e., is a key part of the explanation) for what ads appear to whom on Facebook's platform. That's condition (2).

Facebook's business model includes allowing advertisers to target groups of people narrowly and effectively. It does this in a way that avoids the labor costs associated with human approval of ad targets or human oversight of ad categories and purchases. In so doing, Facebook implies that its algorithmically-generated categories and suggestions are relevant to advertisers (otherwise, advertisers would have no reason to purchase ads). And the fact that one can place ads based on those categories without oversight implies that Facebook believes (at least implicitly) that whatever ads served to whatever audience are appropriate. These are morally relevant qualities, as per our third condition. The algorithms' causal responsibility and implication that they are appropriate obscure the scope of Facebook's causal responsibility (condition (4)).

Finally, in automating its advertising process, Facebook is able to claim that it "never intended or anticipated this functionality being used this way." It effectively distances itself from the fact that a system for which it is (causally and morally) responsible allows noxious (and in the case of discriminatory categories, illegal) advertising. That is, the causal responsibility of the algorithm's suggestions deflects from Facebook's causal responsibility in creating a platform that uses the

[16] Note that this is a possible moral claim that one might make about Facebook and other media organizations. This is a distinct question from what kinds of legal rights and obligations information intermediaries have in light of (inter alia) the United States Communications Decency Act (see 47 U.S.C. section 230), the European Union's eCommerce Directive, articles 12-16 (Council Directive 2000/31, 2000 O.J. (L 178) 1 (EC)), and the EU's General Data Protection Regulation (GDPR) (Commission Regulation 2016/679, 2016 O.J. (L 119) 1 (EU)). See, e.g., Keller (2018)

[17] See, e.g., (Facebook 2018)

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

algorithm, minimizes the labor that would be required to better monitor advertising categories, and profits from the automated system. Its attribution of morally salient characteristics (relevance, usefulness) presupposes that its optimization is consistent with Facebook's responsibilities, though it was not.

Here is where understanding Facebook's actions as agency laundering is a difference-maker. Conditions 1-4 describe several important moral features. But the crux of laundering is condition five. The fact that Facebook is morally responsible with respect to targeted advertising means that it is appropriate to demand that Facebook provide an account of its intentions and reasons in facilitating racists in easily finding an audience to whom they can advertise. Facebook has inserted an automated procedure into its advertisement purchasing procedure, and it suggests that the *algorithms* are the natural object to scrutinize rather than Facebook's reasons and intentions with respect to building a system that deploys them and lets them run with minimal supervision. In doing so, Facebook undermines the central feature of responsibility by deflecting demands for an account of Facebook's reasons, intentions, and actions in helping racists target advertise. Hence, the automated process is a mechanism by which Facebook launders its agency.

There are several potential rejoinders to our argument here. One might disagree about what Facebook's responsibilities are. One might argue instead that it is advertisers and users who bear responsibility for populating Facebook's categories with racist characteristics. Certainly, it is true that users populating categories with anti-Semitic and other racist ads bear responsibility for those actions, and any advertiser targeting ads based on such categories bears responsibility for doing so. But, as in Chair, that others have acted wrongly does not tell us much about Facebook's responsibility. One might further argue that Facebook has not laundered its agency because it has agreed to address the problem. But the fact that Facebook has indicated an intention to address these problems demonstrates that it is a problem within Facebook's control.

A related objection concerns the degree, or the severity, of Facebook's failure to fulfill its responsibilities (assuming that it has some). Perhaps Facebook knew of problems in how its algorithms functioned to allow malignant actions. But perhaps instead it was merely negligent.[18] This is no doubt an area others will reasonably dispute. It does not matter for our analysis of laundering, though. Facebook laundered *whatever degree of agency it had*. Moreover, it can launder its agency even if it meets its substantive role responsibilities. That the advertising platform afforded the opportunity to target advertising in a racist way is something for which detailed explanation of intentions, reasons, and actions is warranted.

Another potential objection is that it may well be that no particular Facebook contractor or developer acted with discriminatory intent, alleviating any potential moral responsibility any of them might have for the outcome (see Binns 2017). However, Facebook's role responsibility is not

[18] One can frame this as a question of capacity responsibility. That is, if Facebook did not have epistemic access to the relevant information about the possibility of misuse, it would not have the necessary capacity to be morally responsibility. Note here that epistemic access is not limited to actual knowledge, but the ability to garner it under reasonable conditions. Hence, Facebook's moral responsibility will turn on the degree to which it could reasonably have known about potential for misuse. And that would define its degree of agency laundering.

One further complicating issue is mitigation. Facebook or another social media company might use its suggestion system to better understand relations among (for example) racists or purveyors of disinformation to promote anti-racist or epistemically sound information. The degree to which that would mitigate or deepen laundering is a question beyond what we can cover here. Thanks to an anonymous reviewer for making this point.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

reducible to any particular individual developer within Facebook. Rather, the company's responsibility is better understood as widely distributed across its contractors, employees, and other stakeholders. Moreover, Facebook is a complex system, and the consequences of its operations over time are impossible to predict. Coeckelbergh and Wackers (2007) argue that organizations deploying such complex, vulnerable systems have obligations to manage their operations not only legally, but with a certain positive 'imagination' regarding systemic crises or other harms. In other words, it unjustifiable to simply let such complex systems run their course and cause harm.

There is a further, related question about whether the conception of responsibility we have outlined here is properly attributable to collectives. There is significant philosophical debate about collective responsibility, and we cannot do it justice here. But we can note two things. First, the accounts of responsibility we outline in section 2 need not be limited to individual wills. Certainly, we do have reactive attitudes towards collections of people, and those targets may be apt. Further, it seems plausible to attribute reasons to groups, in which case it seems plausible that such a group may be responsible in the sense that it ought to be the case that the collective be accountable. Second, even if it is the case that a collective's responsibility is reducible to the responsibility of its individual members, this would imply that those individuals have laundered their agency. In any case, whatever responsibility there is, Facebook's reliance on algorithms to do work and to explain its failures is (on the conception outlined here) an instance of agency laundering.

5 Case 2: Uber and driver management

Another private-sector example shows how our concept of agency laundering can distinguish between structurally similar cases. The ride-hailing company Uber has received substantial social, regulatory, and academic criticism based on its AI-driven, algorithmic systems. Uber uses such systems to map routes, track passengers, monitor drivers, anticipate demand, steer driver behavior, and (at one point) identify and deceive regulators. Many of these uses have been criticized elsewhere on the grounds that they are deceptive, unfair, opaque, or even illegal.[19] Our task here, though, is to consider whether any are instances of agency laundering and, if so, whether analyzing them as agency laundering sheds light on moral concerns with Uber's practices.

Consider how Uber uses algorithmic systems to keep its drivers working. One way is by providing reminders of individual drivers' goals. For example, the Uber app might display a message that the driver is very close to her goal of earning $50 for her shift, which may induce her to take more riders. Similarly, Uber at times sends drivers their next ride requests before they have delivered their current rider. This creates a kind "queue effect," much like video platforms that keep people watching by immediately starting the next episode of a series (Scheiber 2017). A number of critics—including drivers—object to these practices on the grounds that they rely on non-rational mechanisms or are manipulative (Scheiber 2017; Calo and Rosenblat 2017).

A second way that Uber gets drivers to keep working involves the prospect of "surge pricing." Uber drivers can increase their per-hour earnings by driving during high-demand/low-supply periods. When there are lots of passengers seeking rides and relatively few drivers working, Uber will charge higher (surge) prices and drivers thus earn more. Uber's driver app will often prompt drivers to work at times that Uber anticipates will be high-demand. So, it might say that (e.g.) New Year's Eve will probably have surge pricing (Rosenblat 2018, 128–32). However, such prompts do not guarantee

[19] One tool, named "Greyball," was developed to surreptitiously ban users who Uber believed was violating the company's terms of service. Uber eventually used Greyball to surreptitiously ban people Uber believed to be government regulators investigating whether Uber was operating illegally. See (Isaac 2017).

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

surge pricing, and drivers do not know when they accept a ride whether it will be surge-priced. In some cases, the Uber app estimates surge pricing, but fares during that period are normal, either because demand does not materialize or because enough drivers are working to offset the demand (Rosenblat 2018, 98–100).

These two cases are structurally similar: app-based notices that prompt drivers to work. But only the surge-pricing appears to be a case of agency laundering. Begin by running both through our understanding of laundering.

(1) Uber has moral responsibility with respect to fielding drivers, AND
(2) Uber ensures that its algorithm has some causal responsibility with respect to fielding drivers, AND
(3) Uber ascribes morally relevant qualities to its app-based prompts to drivers, AND
(4) In virtue of (2) and (3), Uber obscures the scope of its causal responsibility with respect to fielding drivers, AND
(5) In virtue of (4), Uber fails to adequately account for events within a domain for which it is morally responsible: specifically, the way in which its interface induces driving.

Conditions (1) and (2) are clear enough. Although Uber claims to be a technology company merely connecting riders and drivers through a platform, it nonetheless plays a large role in getting people to both drive and ride. It enters into contractual relationships with drivers and riders, maintains standards for drivers and equipment, subjects drivers to background checks, adjudicates disputes, and so forth. And there is no question that its algorithms are a key part of the explanation of which drivers are driving when. Uber ascribes morally relevant features to the algorithms: that they reflect drivers' own goals, that they are reliable, that they are based on a neutral assessment of facts on the ground (condition (3)).

The differences in the cases concern conditions (4) and (5).

Begin with the case of goal-reminders and queueing effects. It is difficult to see how Uber obscures its causal responsibility in incentivizing driving when it uses these tactics to spur drivers into taking more rides. Certainly, Uber is drawing on (or even exploiting) behavioral psychology, and behavioral psychology is an essential part of the explanation of drivers' decisions to drive. But that fact, and the fact that Uber has set up a system in which algorithms instantiate such strategies, aren't obscured.

Now consider condition 5. Here, too, it is difficult to see how Uber fails to adequately account for events within a domain for which it is morally responsible. As a provider of ride-hailing services, Uber has an interest in keeping enough drivers on the road, and it is using a straightforward tactic to promote this interest. Further, Uber has been clear about the practice. In a recent New York Times article, a spokesperson for Uber describes goal-reminding and queueing as ways to incentivize driving (Scheiber, 2017). As far as we can tell, Uber does not launder its agency when it uses goal-reminders and queueing.

One plausible counterargument here is that, at least in extreme cases, using such tools undermines drivers' wills so much that it obscures the scope of Uber's causal responsibility. Perhaps the interface is sufficiently gamified that users have hallmarks of addiction, or perhaps the quality of drivers' wills is so degraded that decisions to drive do not count as drivers' own. In that case, Uber's causal responsibility would be far greater than it appears and any adequate account of Uber's responsibility would include an explanation of how it circumvents drivers' wills. That possibility is worth both empirical and philosophical examination. Nonetheless, at least weaker forms of nudging seem well within the range of responsible employer behavior and not cases of agency laundering.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

Contrast the goal-reminders and queuing with the surge pricing case. Uber uses machine learning techniques to predict high-demand/low-supply times, and uses those predictions to prompt drivers to work. This *does* seem to obscure Uber's causal responsibility in fielding drivers, per condition (4). The judgment that surge pricing is likely to occur appears to be an inference about how the world outside Uber is operating, and Uber is merely reacting to it. Indeed, in comments pertaining to the phenomenon of surge pricing, Travis Kalanick (Uber's cofounder and former CEO) said, "We are not setting the price. The market is setting the price" (Clark 2015). In fact, Uber is causally responsible for setting up a system in which there are pay differentials (where driving at surge times is more attractive to drivers) and then using those facts to induce driving. Surge periods are not a natural feature about the world that Uber measures, but a period defined and deployed by Uber.

More important, though, is condition 5. By predicting surge pricing and signaling the likelihood of surge pricing to drivers, Uber simultaneously exploits surge pricing and makes it less likely. In other words, by using surge pricing as an inducement to drivers seeking a better wage, Uber helps ensure that supply more closely matches demand. Thus, it creates for drivers reasonable expectations of better pay and fails to meet them. Then, when the prices are not offered, Uber tells the drivers it is the market that is making the decision, not Uber. But this is a failure to account for the situation a driver finds herself in when, for example, she has driven to Times Square on New Year's Eve under Uber's advice that there will be a surge and then finds that the surge has disappeared. As Alex Rosenblat observes, "When drivers follow this advice and find that they have been dispatched to pick up a passenger for a nonpremium-priced ride, meaning that surge pricing has disappeared, they feel tricked."[20] It is not the market that reached out to the driver to quell the surge. It is not the market that decided how Uber's payment system works. Rather, it is Uber that set up a system where a driver who responds to its enticements may not get surge rates if the campaign to get drivers to an area has worked. And this is the action within its domain of moral responsibility for which Uber owes an account. Uber's claim that it is simply the market's doing is an inadequate account, satisfying condition 5. Thus, Uber launders its agency.

So, our conception of agency laundering is sensitive enough to distinguish between different uses of algorithmic systems to influence driver behavior. The next question is whether analyzing each in terms of agency laundering adds something of value beyond simply analyzing Uber's responsibilities to its employees. We believe that it does. The laundering analysis emphasizes the fact that use of tools (committees, bureaucracies, technologies) may be a way to simultaneously violate duties and undermine accountability. This is a way to show that laundering adds something. By calling it laundering, we can make clear what is happening. But more importantly, our argument picks out a discrete moral infirmity, viz., eroding others' ability to demand Uber provide an account of its reasons and intentions. In Uber's case, use of a tool to both predict surge pricing and induce drivers looks similar to use of other prompts. One might be tempted to think of it as a case of nudging (or perhaps of manipulation). But that would miss the fact that by tying the process to a prediction about facts on the ground, Uber can deflect attention from its own responsibility for creating a situation in which it simultaneously predicts surge pricing and makes it less likely.

6 Case 3: IVAs, Teachers, and Laundering

Facebook, Uber, and other large technology firms receive substantial attention. It would be a mistake, however, to think that agency laundering is primarily the province of the private sector. The

[20] *Uberland*, p. 129.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

depth and importance of agency laundering may be even greater in public agencies. To demonstrate, we will examine a case involving the Houston Independent School District (HISD).

In 2012, HISD engaged in a project to improve quality of teaching by firing its most ineffective teachers. This was done by using Individual Value-Added measurements (IVAs), which purport to measure an individual teacher's contribution to student achievement by comparing the results of annual standardized tests (Isenberg and Hock 2012). More specifically, HISD used a proprietary IVA, EVAAS (developed by the company SAS), to evaluate teachers. In the first three years using EVAAS, HISD "exited" between 20% and 25% of the teachers rated ineffective. The teachers sued alleging that their due process rights were violated (Houston Federation of Teachers, Local 2415 v. Houston Independent School District 2017).

The district court agreed with the teachers' claim that their due process rights were violated. EVAAS is proprietary, which prevents teachers from auditing their scores to see if they were produced properly. Further, there was no mechanism in place to detect or correct any clerical or coding errors that may have affected teachers' scores (Houston Federation of Teachers, Local 2415 v. Houston Independent School District 2017). Even further, were any mistakes detected, they would not be corrected. HISD explains why in response to a "frequently asked question":

> Once completed, any re-analysis [of one's EVAAS score] can only occur at the system level. What this means is that if we change information for one teacher, we would have to run the analysis for the entire district, which has two effects: one, this would be very costly for the district, as the analysis itself would have to be paid for again; and two, this re-analysis has the potential to change all other teachers' reports (Houston Independent School District 2015 (emphasis in original)).

So, EVAAS is in practice not auditable for two interrelated reasons. First, all of its scores are so deeply interconnected with other scores that the only way to recalculate a score is to recalculate them all. Second, recalculating all of the scores is a complex task and thus very costly.

Despite its shortcomings, HISD defended its use of EVAAS on the grounds that it reliably measures student progress.[21] But this claim, even if true, is largely irrelevant to the question of whether the use of EVAAS is justified. Student progress and the contributions an individual teacher makes to student progress are distinct quantities. To measure one is not to measure the other. Indeed, the American Statistical Association (ASA) issued a statement in 2014 pointing out that most studies conclude that teachers have only a marginal effect on the test scores that IVAs (such as EVAAS) take as inputs (American Statistical Society 2014, p. 2). The ASA further concludes that IVAs have large standard errors, and that these errors make rankings of teachers unstable, "even under the best conditions" (American Statistical Society 2014, p. 7). So, even if EVAAS reliably measures student progress, this is a poor proxy for teacher effectiveness.

Has HISD laundered its responsibility for firing teachers? We think so.

(1) HISD is morally responsible with respect to hiring, firing, and promoting teachers, AND
(2) HISD ensures that EVAAS has some causal responsibility in making those determinations, AND
(3) HISD ascribes morally relevant qualities to EVAAS, AND

[21] ("Defendant's Original Answer and Defenses, Houston Federation of Teachers, Local 2415 v. Houston Independent School District (S.D. Tex.)," n.d., 9 citing (Sanders et al. 2009))

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

(4) In virtue of (2) and (3), HISD obscures the scope of its causal responsibility with respect to hiring, firing, and promoting teachers, AND
(5) In virtue of (4), HISD fails to adequately account for events within a domain for which it is morally responsible: specifically, the "exiting" of teachers deemed ineffective through EVAAS.

In virtue of HISD's role as an employer, the first condition is met. When HISD implements EVAAS to aide in personnel decisions it meets the second condition. HISD meets the third condition implicitly by using EVAAS for high-stakes decisions. It meets the third condition explicitly by invoking EVAAS's reliability in measuring student progress as a reason in favor of using EVAAS. The fourth condition is met when HISD repeatedly refers to one good thing that EVAAS does (measure student progress) to obscure the fact that HISD is implementing a system in which teachers are fired based on measures for which the teachers are not responsible (recall the statement from the ASA above). Finally, the fifth is met because teachers who are fired on account of their EVAAS scores are given a faulty accounting of why they were fired. They are told they are being fired for being ineffective, when, given EVAAS's flaws, this is likely not the case. Hence, understanding HISD's actions as laundering shows us that there is something going on beyond lack of transparency; the mechanism of evaluation positively misdirects those who would seek reasons for how teachers are treated. It forestalls teachers' ability to demand an account for HISD's actions within a domain of their legitimate concern

Note that HISD uses EVAAS while fulfilling its public function of managing an education system. That means that the public has a collective stake in how the system functions and has an interest in the actions HISD undertakes. Hence, the fact that HISD's laundering makes accountability all the more difficult matters in a way that accountability of private firms does not; it suggests that use of EVAAS must conform to standards of public reason, rather than aligning only to the isolated wishes of HISD (See Binns 2018).

7 Case 4: Laundering in Criminal Sentencing

So far, we have described agency laundering in both private sector and public sector cases. And in the Uber case we saw how use of algorithmic decision systems will not be agency laundering where the agent does not obscure their causal responsibility for outcomes. Our final case demonstrates how a public entity can avoid agency laundering by making clear their moral responsibility for an outcome.

Eric Loomis pleaded guilty to crimes related to a drive-by shooting. In considering Loomis's sentence, the circuit court ordered a presentence investigation report ("PSI"). The PSI incorporated the results from the COMPAS risk assessment tool. COMPAS is a proprietary information system that combines a wide array of information (about, e.g., friends, family, work, education, drug and alcohol abuse, criminal record, housing stability) from a variety of sources (e.g., state records, questionnaire responses, assessments from within the criminal justice system) to generate risk profiles of people charged with crimes. COMPAS classified Loomis as "high risk" for both recidivism and violent recidivism. Loomis received a sentence in the maximum range. In sentencing Loomis, the judge referenced the COMPAS report even though Northpointe (the developer of COMPAS) is clear that the tool is not designed to be used for sentencing.

Loomis and COMPAS have been the subjects of significant criticism. However, we think that the Wisconsin Supreme Court's opinion in the case shows how actors can avoid agency laundering in

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

deploying algorithmic systems. Hence, the case shows that our understanding of agency laundering is not so broad as to be meaningless [Redacted].

A trial court launders its agency where:

(1) The trial court has moral responsibility with respect to sentencing, AND
(2) The trial court ensures that COMPAS has some causal responsibility with respect to sentencing, AND
(3) The trial court ascribes morally relevant qualities to COMPAS, AND
(4) In virtue of (2) and (3), the trial court obscures the scope of its causal responsibility with respect to sentencing, AND
(5) in virtue of (4), the trial court fails to adequately account for decisions pertaining to sentencing, specifically the decision to sentence Loomis in the maximum range.

The trial court certainly has moral responsibility with respect to sentencing. But did the trial court ensure that COMPAS had some causal responsibility with respect to sentencing? The judge referenced Loomis's risk scores and they plausibly had an effect on sentencing. This, though, was only one of the factors the judge described. He also considered important the conduct Loomis admitted as part of the read-in charges and Loomis's conduct while under prior supervision. Let's interpret this as giving some degree of causal responsibility to COMPAS. It is, after all, at least plausible that the COMPAS score is a key part of the explanation for Loomis's sentence. The court's use of COMPAS implies that it is useful, reliable, and fair, which are morally relevant qualities, per condition (3).

The question of agency laundering in *Loomis* turns on conditions (4) and (5). Although the judge in the case referenced the COMPAS assessment in his decision, he also indicated that his own judgment (based on Loomis's conduct and history) led him to a similar conclusion. There is some possibility that the judge was confabulating by ascribing his own reasons to the outcome COMPAS reached. If that's true, it would not be that the use of COMPAS obscures the trial court's causal responsibility. Rather, it would be that the court's description of its reasons obscures the scope of COMPAS's causal responsibility.

The key issue, though, is whether the trial court fails to adequately account for decisions pertaining to sentencing. Consider the following from the Wisconsin Supreme Court's decision.

> We determine that because the circuit court explained that its consideration of the COMPAS risk scores was supported by other independent factors, its use was not determinative in deciding whether Loomis could be supervised safely and effectively in the community. Therefore, the circuit court did not erroneously exercise its discretion. (Loomis, ¶9)

The passage makes clear that tools like COMPAS cannot be used alone, and use of such scores has to be supported by other factors that are independent of the tool. Similarly, the court required that courts weigh all relevant factors in order to sentence an individual defendant (Loomis, ¶74), and it prohibited trial courts from using scores to determine whether to incarcerate a person or not, to determine the length and severity of sentence, and to determine aggravating or mitigating factors in sentencing (Loomis, ¶¶88-98). And the court required that any PSI that uses a COMPAS report carry a number of warnings about the limitations of such reports.

The supreme court's *Loomis* opinion places responsibility squarely on the trial court in using tools like COMPAS. It prohibits trial courts from relying completely on the COMPAS algorithm, and it requires

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

trial courts to use other factors to support any use of risk assessment algorithms. Hence, the court forecloses the ability of trial courts to use algorithms as a way to distance themselves from responsibility. Thus, *Loomis* addresses condition (5), and it is not a case of agency laundering.[22]

Deploying tools like COMPAS could certainly be a means by which courts (and others in the criminal justice system) can launder their agency. However, the *Loomis* decision is tailored precisely to avoid that. Hence, it appears to be to be a good test case for our view. It is a use of an algorithmic system where one does not launder their agency. As a result, it can demonstrate how other actors may fail in their moral responsibilities, even where their actions superficially resemble the *Loomis* court's. The court did not forestall others' ability to demand an account for its actions within a domain of their legitimate concern; but a different court (or different actor within a criminal justice system) might do so by failing to provide its own reasons for decisions.

8 Related concepts and concerns

8.1 The Responsibility Gap

Agency laundering can help shed light on some other concepts related to moral issues in technology. One of these is the "responsibility gap." In a 2004 article, Andreas Matthias argued that in some cases a technological system may be sufficiently sophisticated that no person or persons are responsible for the outcomes it causes (Matthias 2004). The idea is that machine-learning systems may be so opaque to human developers and users that it is impossible to predict how those systems will behave. Where such systems cause harm, it may be (on Matthias's view) a mistake to attribute responsibility to the developer, the owner, or any other person. The rules by which machine learning systems act "are not fixed during the production process, but can be changed during the operation of the machine, *by the machine itself*" (Matthias 2004, 177 (emphasis in original)). These actions do not mesh with traditional accounts of responsibility "because nobody has enough *control* over the machine's actions to be able to assume the responsibility for them" (Matthias 2004, 177). He provides a number of examples. One is an elevator system that, having used an AI system to adapt to use patterns over time, leaves an executive stranded and late for a meeting. Another is a machine learning system to diagnose lung cancer, but which has a high false-positive rate (and causes emotional and financial stress to people diagnosed). Yet another is an AI children's toy that, in learning to navigate a new home environment, injures a child.

There has been a great deal of discussion of the responsibility gap in the years since Matthias's article was first published. Here, we want to illustrate how agency laundering is distinct from responsibility gaps and how it can explain where responsibility fits in the gaps.

Note first that Matthias's conception of the responsibility gap focuses on an automated system's causal responsibility for some outcome. In the toy case, Matthias posits that the responsibility gap pertains to the action of knocking over and injuring the child. Our account of agency laundering, however, considers a wider range of actions. Imagine that the toy manufacturer developed, marketed, and sold the toy without fully testing its ability to knock over and injure a toddler. The manufacturer would seem to have causal and role responsibility with respect to whether its toys

[22] Note that *Loomis* demonstrates another way one can launder even while fulfilling one's substantive role responsibilities. Imagine that the trial court had deliberated about its decision, but did not explain its reasoning for the sentence. Suppose instead it merely wrote that it agreed with the COMPAS report's assessment with no further comment. That would obscure the scope of the court's causal responsibility and would fail to provide an adequate account of the decision. But in that case, the court would not have violated some other substantive role responsibility.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

injure children (condition (1)). It would also ensure that the toy has causal responsibility for whether it injures children (as Matthias describes the case, the child's injury is explicable only by describing how the toy operates) (condition (2)). By selling the toy, the manufacturer attributes morally relevant qualities to the toy (age appropriateness, safety) (condition (3)). It would also be difficult to provide an adequate account of a toy manufacturer's distribution of a toy that has the affordances (size, weight, mobility, unpredictability) to knock over a small child (condition (5)).

The question, then, is whether the manufacturer obscures the scope of its causal responsibility with respect to the injury (condition (4)). Nothing in the example (either Matthias's version or ours) suggests that it does. However, if the manufacturer were to *posit* a responsibility gap (for example, by saying that the machine learning process was opaque, and hence the manufacturer could not anticipate injury), that would fulfill condition (4) and be an instance of agency laundering. In other words, invoking the idea of a responsibility gap is a mechanism by which people may launder their agency.

There are other possibilities as well. One might set up a system expressly to avoid being held to account. Such a scenario would appear to be a form of preemptive laundering, and it would be advanced by whatever responsibility gap it creates . A different possibility is that one creates a gap between one's actions and outcomes for good reasons, but in so doing ensures that there will be a responsibility gap. Suppose, for example, that an agency responsible for assessing how likely persons accused of crime are to re-offend. In order to address a known problem of arbitrary assessments by human decision-makers, it deploys a system similar to COMPAS (while acknowledging biases similar to those in COMPAS). This would look like a case similar to democratic chair in section 3. The agency's use of the system would not be a means to avoid having to account for some other action. Rather, it is the decision to deploy the system that requires an account, and that decision is not obscured.[23]

We leave open whether there are genuine cases of responsibility gaps—that's a topic others have addressed more thoroughly than we can do here. But our analysis of agency laundering requires thinking about role and causal responsibilities of people who deploy technologies like those Matthias contemplates. That forces one to consider a wider range of actions than the operations of an AI system, and can help distinguish genuine responsibility gaps from responsibility obfuscation and agency laundering.

8.2 Bias Laundering, Masking, and Humans in the Loop

At a 2016 conference sponsored by the Society for the Advancement of Socio-Economics, Maciej Ceglowski stated that "machine learning [is] an ingenious way of disclaiming responsibility for anything. Machine learning is like money laundering for bias"(Ceglowski 2016).[24] Although Ceglowski does not spell out what laundering is or why it matters morally, there do seem to be some points of similarity and difference worth noting. What Ceglowki's comment picks out is the ability to obscure something important and deflect disapproval. So, if an algorithm (e.g., for predictive policing) is built on criminal justice data which is itself based on over-policing black communities, the algorithm may be mathematically neutral tool that reflects biases that already exist. The tool's neutrality can appear neutral *tout court* to the extent that one fails to examine the bias in the underlying data sources.

[23] Thanks to a reviewer here for pointing out these possibilities and noting their similarities to "Chair" and "Democratic Chair."

[24] Thanks to [REDACTED] for pointing us to this talk.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

However, our understanding of agency laundering is a general account of laundering, and it is broad enough that it encompasses bias laundering. Agency laundering can obscure many different kinds of wrongs, and limiting the concept to bias laundering would fail to capture them. Likewise, there is no need to link laundering tightly with machine learning or algorithmic decision-making. As we've explained, any process or socio-technical system can be a mechanism for laundering. More importantly, our account explains just how laundering is related to responsibility, both structurally and morally. Finally, in our view the thing that is laundered is typically agency, and that is typically the appropriate target of analysis.[25] One might at times act as an agent, yet launder responsibility, but in any case *these* are the things laundered. Thus, while we agree that machine learning can be a means of disclaiming responsibility, just what it means to 'disclaim responsibility' and just what it is that one is responsible for are difficult questions to answer. This paper is an attempt to do just that.

A similar concept is "masking," or the intentional use of algorithmic systems to obfuscate discrimination (Barocas and Selbst 2016, 692–93, 712–14). Barocas and Selbst describe masking as a way of using data mining to return discriminatory results, but hiding discriminatory intent behind an information systems. Certainly, masking could be part of laundering. However, the other elements of laundering—relevant role responsibilities, ascription of morally relevant qualities, tension with fundamental aspect of responsibility (viz., accountability)—are not necessarily elements of masking.

There is another important issue related to socio-technical systems, responsibility, and moral liability. The distinction between systems with humans in- and out-of-the loop are well-established. Systems employing humans in the loop include things like automated cars that provide for human override, autonomous weapons systems that require humans to approve strikes, and content moderation in which humans help teach algorithms what content is objectionable and make decisions in cases for which automated systems are not yet adept. Control in such systems is itself a complicated concept, and there is an active area of scholarship surrounding whether (and if so, how) there can be meaningful human control even for systems that leave humans out of particular decision loops (Santoni de Sio and van den Hoven 2018).

Of particular importance for our project is that having humans in the loop may itself obscure causal and moral responsibility. Ben Wagner (2019) notes that there are many purportedly automated systems that rely on humans to take an active role, fix mistakes, or replace system decisions. However, he argues that the actual human role may be compromised by the design of the system. For example, there may be insufficient time to make decisions, they may grow weary or inured to a process, or they may lack sufficient training and experience to make good decisions. He outlines a number of criteria important in determining whether systems are "quasi-autonomous," such that humans in the loop "have responsibility but little agency" (or, in our usage, humans have causal but not capacity responsibility, and hence cannot be morally liable).

Madeleine Elish (2019) considers similar scenarios in which human actors have a causal role within socio-technical systems (including AI). She argues that responsibility for outcomes may be misattributed to human actors within such systems, creating a kind of "moral crumple zone" that protects the system from attributions of responsibility. In our conception, humans' causal responsibility could obscure the causal responsibility of a technical system (of course it cannot

[25] We appreciate an anonymous reviewer raising the question of whether there are cases where one maintains agency but launders accountability instead. Our sense is that any such case would involve minimizing one's agency. In other words, accountability is the thing that is avoided, and one avoids it by laundering the degree to which one is (morally) responsible, which is in turn a function of a person's agency in a process. Likewise, money laundering is a way to forestall accountability, and it is the laundering of some other thing (viz., money) that helps avoid the accountability.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

obscure the *moral* responsibility of a technical system, for such a system does not have capacity responsibility).

The systems that Wagner and Elish envision are ones that could potentially launder agency (though not necessarily). Suppose, for example, an autonomous vehicle has a human in the loop, but the human has too little time to respond when needed and causes an accident. That would seem to fulfill the conditions (1) and (2): some entity with moral responsibility has ensured a human has causal responsibility. The questions are whether placing a human in the loop attributes morally relevant qualities to the human's actions (efficacy, perhaps), whether doing so obscures the causal responsibility of the larger system, and thereby fails to adequately account for the moral responsibility of the larger entity. What is key for our view, though, is that the mechanism for laundering need not be technological at all; that is, humans in the loop can be a means of laundering just as well as automation itself.

8.3 Concerns

One potential objection to our conception of agency laundering is that it is merely a metaphor and as such does not add a great deal to our ability to analyze and evaluate the relationship between information technologies and responsibility.[26] There are a couple of reasons to think otherwise. Using the concept of laundering takes its cue from the idea of money laundering, which is of course metaphorical. Crooks don't literally wash money. Rather, they obscure its sources by mixing it with money from legitimate sources. Hence, whatever actions work to obscure the source of illicit funds serve to launder. Laundering is way of obscuring the source of morally weighty states of affairs by mixing actions with technologies, procedures, or bureaucracies. Part of the value of using the laundering metaphor (for both money and agency) is that it plays a "descriptive role in helping a lay person understand" what the underlying phenomenon is.[27] That is, a metaphor can help capture the gist of a concept, and in this case give people an intuitive grasp of the underlying concerns before following the entire argument.

Note, too, that the concept of agency laundering can help us both to make judgments in difficult cases and to more fully explicate antecedent moral wrongs. So, for example, in the Facebook case, it is unclear just what the moral wrong is in using an automated targeted advertising system that bad actors can exploit. It is plausible that Facebook did not act wrongly in developing and using such a system. However, the advertising platform is still within a domain for which Facebook has moral responsibility, and its conflation of its actions with an automated system's actions undermines the foundation of responsibility, viz., providing an account.

Similarly, the concept of agency laundering can explain why Uber acts wrongly in some cases (surge pricing) but not in others (goal prompts, queuing).[28] Both actions are within Uber's domain of responsibility, and both are actions where there is an open question about whether Uber infringes its substantive role responsibility. Our account of agency laundering can help evaluate what, if anything, Uber does wrong. The account may be of particular use in public-facing cases, where organizations have a remit to serve the public and derive legitimacy from public trust and support. In cases like HISD and *Loomis*, the possibility of socio-technical systems forestalling persons' abilities to demand

[26] Thanks to an anonymous reviewer for pressing us on this point.
[27] Thanks to an anonymous reviewer for this language and description. In this signaling respect, our use of a metaphor here works similarly to the "crumple zone" metaphor in Elish (2019), discussed in section 8.2.
[28] Note that there may be other, non-laundering moral wrongs involved in goal prompts and queuing, as discussed in section 5.

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

an account of organization actions within a domain of legitimate concern is particularly important. Drawing on the metaphor of laundering here helps capture content of the concept.

A further advantage of our account is that it may help in understanding what kinds of rights to explanation people have in the context of automated or algorithmic decision systems, for example in the GDPR.[29] Such a right (if there is one) is generally discussed as an individual right in the face of adverse decisions (Wachter et al 2017; Selbst and Powles 2017; Kaminski 2019). But agency laundering is a problem not just for individuals whose interests have been affected. It is also a general problem, and our arguments about laundering and forestalling others' abilities to demand an account within areas of their legitimate interest extend further. After all, how a massive social media company helps target ads, how an international employer of drivers with millions of users and drivers induces use, how a school district evaluates and fires teachers, and how a criminal justice system wields its power are areas of general legitimate interest, regardless of whether a particular individual has a claim to an explanation of a discrete event.

9 Conclusion

Our goals in this paper were to, first, explain a type of wrong that arises when agents obscure responsibility for their actions. We have outlined this type of wrong and called it "agency laundering." Second, was to draw on several cases to help specify our account of agency laundering. We have argued that some of these (Facebook advertising, Uber's surge-pricing prompts, HISD's use of EVAAS) involve laundering and two (other Uber prompts, use of COMPAS in the *Loomis* case) do not. Third, we have argued that analyzing these cases in terms of agency laundering both helps understand the cases and adds something morally. Lastly, we have distinguished agency laundering from other relevant concepts.

We have not given the final word on agency laundering here. One further question concerns the degree to which laundering must be intentional. In other words, can a person who uses a tool to make decisions launder her agency inadvertently? This appears compatible with our definition of agency, though the moral importance of such laundering warrants further consideration. Another question concerns how widely the concept of agency laundering applies. A number of people with whom we've discussed this project have asked whether large-scale social processes (e.g., political events and movements) can serve as mechanisms of laundering. Perhaps so, though that would involve sorting through complex issues of causal responsibility and conceptual questions of capacity responsibility.

References

American Statistical Society. 2014. "ASA Statement on Using Value-Added Models for Educational Assessment." http://www.amstat.org/policy/pdfs/asa_vam_statement.pdf.

Angwin, Julia, and Jeff Larson. 2016. "Machine Bias." Text/html. ProPublica. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[29] Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119).

This is the authors' accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

Angwin, Julia, Noam Scheiber, and Ariana Tobin. 2017. “Dozens of Companies Are Using Facebook to Exclude Older….” *ProPublica*, December 20, 2017. https://www.propublica.org/article/facebook-ads-age-discrimination-targeting.

Angwin, Julia, Madeleine Varner, and Ariana Tobin. 2017. “Facebook Enabled Advertisers to Reach ‘Jew Haters.’” Text/html. ProPublica. September 14, 2017. https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters.

Barocas, Solon, and Andrew D. Selbst. 2016. “Big Data’s Disparate Impact Essay.” *California Law Review* 104: 671–732.

Binns, Reuben. 2017. “Fairness in Machine Learning: Lessons from Political Philosophy.” ArXiv:1712.03586 [Cs], December 10, 2017. http://arxiv.org/abs/1712.03586

———. 2018. “Algorithmic Accountability and Public Reason.” *Philosophy & Technology* 31(4): 543–56. https://doi.org/10.1007/s13347-017-0263-5.

Calo, Ryan, and Alex Rosenblat. 2017. “The Taking Economy: Uber, Information, and Power.” *Columbia Law Review* 117 (6): 1623–90.

Ceglowski, Maciej. 2016. “The Moral Economy of Tech.” Presented at the Moral Economies, Economic Moralities,), Society for the Advancement of Socio-Economics, Berkeley, CA, June 24. https://sase.org/event/2016-berkeley/.

Citron, Danielle Keats. 2008. “Technological Due Process.” *Washington University Law Review* 85 (6): 1249–1313.

Clark, Liat. 2015. “Uber Denies Researchers’ ‘phantom Cars’ Map Claim.” *Wired UK*, July 28, 2015. https://www.wired.co.uk/article/uber-cars-always-in-real-time.

Coeckelbergh, Mark, and Ger Wackers. 2007. “Imagination, Distributed Responsibility and Vulnerable Technological Systems: The Case of Snorre A.” Science & Engineering Ethics 13, no. 2 (June 2007): 235–48

“Defendant’s Original Answer and Defenses, Houston Federation of Teachers, Local 2415 v. Houston Independent School District (S.D. Tex.).” n.d.

Dennett, D. C. 1997. "When HAL Kills, Who's to Blame? Computer Ethics," in *HAL's Legacy: 2001's Computer as Dream and Reality*, in D. G. Stork (ed.), 351-365. Cambridge, MA: MIT Press.

Duff, R.A. 1998. “Responsibility.” *Routledge Encyclopedia of Philosophy*. https://doi.org/10.4324/9780415249126-L085-1.

Duhigg, Charles. 2012. “How Companies Learn Your Secrets.” *The New York Times*, February 16, 2012, sec. Magazine. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Elish, Madeleine Clare. 2019. "Moral crumple zones: Cautionary tales in human-robot interaction." *Engaging Science, Technology, and Society* 5: 40-60.

Eubanks, Virginia. 2018. *Automating Inequality : How High-Tech Tools Profile, Police, and Punish the Poor*. First edition. New York, NY: St. Martin’s Press.

Facebook. 2018. “Community Standards Enforcement.” May 2018. https://transparency.facebook.com/community-standards-enforcement.

Fischer, John Martin, and Mark S.J. Ravizza. 2000. *Responsibility and Control : A Theory of Moral Responsibility*. Cambridge, U.K.; New York: Cambridge University Press.

Goodin, Robert E. 1986. “Responsibilities.” *Philosophical Quarterly* 36 (142): 50–56.

———. 1987. “Apportioning Responsibilities.” *Law & Philosophy* 6 (August): 167–85. https://doi.org/10.1007/BF00145427.

Harford, Tim. 2014. “Big Data: Are We Making a Big Mistake?” Financial Times. March 28, 2014. https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0.

Hart, H. L. A. 1968. *Punishment and Responsibility; Essays in the Philosophy of Law*. New York: Oxford University Press.

Hill, Thomas Jr. 1984. “Autonomy and Benevolent Lies.” *Journal of Value Inquiry* 18: 251–97.

Houston Federation of Teachers, Local 2415 v. Houston Independent School District. 2017, 251 F.Supp.3d 1168. S.D. Tex.

This is the authors’ accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

Houston Independent School District. 2015. “EVAAS/Value-Added Frequently Asked Questions.” http://static.battelleforkids.org/documents/HISD/EVAAS-Value-Added-FAQs-Final-2015-02-02.pdf.

Isaac, Mike. 2017. “How Uber Deceives the Authorities Worldwide.” *The New York Times*, December 22, 2017, sec. Technology. https://www.nytimes.com/2017/03/03/technology/uber-greyball-program-evade-authorities.html.

Isenberg, Eric, and Heinrich Hock. 2012. *Measuring School and Teacher Value Added in DC, 2011-2012 School Year. Final Report*. Mathematica Policy Research, Inc. https://eric.ed.gov/?id=ED565712.

Kaminski, Margot E. 2019. "The Right to Explanation Explained." *Berkeley Technology Law Journal* 34(1): 189-218.

Keller, Daphne. 2018. "The Right Tools: Europe’s Intermediary Liability Laws and the EU 2016 General Data Protection Regulation," *Berkeley Technology Law Journal* 33(1): 287-364.

Kutz, Christopher. 2004. “Responsibility.” In *The Oxford Handbook of Jurisprudence and Philosophy of Law*, edited by Jules Coleman, Scott Shapiro, and Kenneth Einar Himma, 548–87.

Levin, Sam. 2019. “Facebook cracks down on discriminatory ads after years of backlash.” *The Guardian*, March 19, 2019. Sec. Technology. https://www.theguardian.com/technology/2019/mar/19/facebook-advertising-discrimination-lawsuit-aclu-race-gender

Levin, Sam, and Julia Carrie Wong. 2018. “Self-Driving Uber Kills Arizona Woman in First Fatal Crash Involving Pedestrian.” *The Guardian*, March 19, 2018, sec. Technology. https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

Matthias, Andreas. 2004. “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata.” *Ethics and Information Technology* 6 (3): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. “The Ethics of Algorithms: Mapping the Debate.” *Big Data & Society* 3 (2). https://doi.org/10.1177/2053951716679679.

Nissenbaum, Helen. 1994. "Computing and Accountability." *Communications of the Association for Computing Machinery* 37(1): 72-80.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression : How Search Engines Reinforce Racism*. New York: New York University Press.

O’Neil, Cathy, author. 2016. *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Oremus, Will, and Bill Carey. 2017. “Facebook Let Advertisers Target ‘Jew-Haters.’ It Doesn’t End There.” Slate Magazine. September 14, 2017. https://slate.com/technology/2017/09/facebook-let-advertisers-target-jew-haters-it-doesnt-end-there.html.

Oshana, Marina A. L. 1997. “Ascriptions of Responsibility.” *American Philosophical Quarterly* 34 (1): 71–83.

Pasquale, Frank author. 2015. *The Black Box Society : The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Rosenblat, Alex. 2018. *Uberland : How Algorithms Are Rewriting the Rules of Work*. Oakland, California: University of California Press.

Sandberg, Sheryl. 2017. “Last Week We Temporarily Disabled Some of Our Ads Tools.” Facebook. September 20, 2017. https://www.facebook.com/sheryl/posts/10159255449515177.

Sanders, William L., S. Paul Wright, June C. Rivers, and Jill G. Leandro. 2009. “Addressing Common Concerns About Value-Added Modeling.” SAS. https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/value-added-modeling-107101.pdf.

This is the authors’ accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.

Santoni De Sio, Filippo and Jeroen van den Hoven. 2018. “Meaningful Human Control Over Autonomous Systems: A Philosophical Account.” *Frontiers In Robotics and AI* 5(15) https://doi.org/10.3389/frobt.2018.00015

Scanlon, Thomas. 2008. *Moral Dimensions : Permissibility, Meaning, Blame*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Scheiber, Noam. 2017. “How Uber Uses Psychological Tricks to Push Its Drivers’ Buttons.” *The New York Times*, April 2, 2017, sec. Technology. https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html.

Selbst, Andrew and Julia Powles. 2017. "Meaningful Information and the Right to Explanation." *International Data Privacy Law* 7(4): 233-242.

Shoemaker, David. 2011. “Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility.” *Ethics* 121 (3): 602–32.

Smith, Angela M. 2012. “Attributability, Answerability, and Accountability: In Defense of a Unified Account.” *Ethics* 112 (3): 575–89.

Strawson, Peter. 1962. “Freedom and Resentment.” *Proceedings of the British Academy* 48: 1–25.

Sweeney, Latanya. 2013. “Discrimination in Online Ad Delivery.” *ArXiv:1301.6822 [Cs]*, January. http://arxiv.org/abs/1301.6822.

Vincent, Nicole A. 2011. “A Structured Taxonomy of Responsibility Concepts.” In *Moral Responsibility: Beyond Free Will and Determinism*, edited by Nicole A. Vincent, Ibo van de Poel, and Jeroen van den Hoven, 15–35. Library of Ethics and Applied Philosophy. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4_2.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 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.

Wagner, Ben. 2019. "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems." *Policy & Internet* 11(1): 104-122.

Williams, Garrath. 2008. “Responsibility as a Virtue.” *Ethical Theory and Moral Practice* 11 (4): 455–70.

This is the authors’ accepted manuscript of an article forthcoming in *Ethical Theory & Moral Practice*. The final authenticated version is available online at https://doi.org/10.1007/s10677-019-10030-w.