*Jan 2025*

**This is a draft book chapter for the *Routledge Introduction to Business Ethics: Philosophy, Public Policy, and Management* (eds. Vikram Bhargava and Michael Kates)**

***Kyle van Oosterum (Kyle.vanoosterum@philosophy.ox.ac.uk)***
***University of Oxford***

**--**

## Automated Technologies and Artificial Intelligence (6, 783 words)

In the film *Jurassic Park,* one of the characters points out that the scientists "were so preoccupied with whether or not they could that they didn't stop to think if they *should.*" The line is somewhat clichéd but it is nevertheless extremely relevant to thinking about the ethics of AI and automated technologies. Sometimes it feels like developments in technology inevitably outrun our capacity to regulate them. Part of the problem is that the developers of these technologies hype them up as if they are truly unique. They sometimes claim we have to completely rethink how we use and even think about technology itself. It's great marketing. But like any good marketing it distorts the way we think and we have to resist it. The first step is to change *how* we think about these technologies and philosophical reflection is key to changing *what* we think about them.

This chapter is divided into four sections. In the first section, I critically examine the relationship between responsibility and automated technology. Some philosophers worry that truly autonomous technologies (e.g., self-driving cars) create so-called *responsibility gaps* where it's unclear who can and should be held responsible. Are there genuine gaps in responsibility or can we look to concepts like negligence to close said gaps? How we answer these questions has implications for companies that make autonomous vehicles.

In the second section, I discuss the relationship between explainability and legitimacy. Are persons owed an explanation if a black-box algorithm makes a decision that affects them? If no explanation can be provided, is it morally illegitimate to use these technologies? Are humans really better than algorithms when it comes to explaining their decisions? How we answer these questions has implications for deploying and selling algorithms to assist human decision-makers such as medical diagnosticians.

In the third section, I discuss what is lost when we replace human decision-making with algorithmic decision-making. Is it unfair when humans decide in noisy unpredictable ways? Is noisy decision-making always bad? What is good about human decision-making? How we answer these questions has implications for our run-of-the-mill practices like interviewing candidates for a job.

Finally, I conclude with an overview and brief discussion of large-language models, the most recent and influential use-cases of artificial intelligence. I canvass only some of the interesting questions one could ask and develop a short concern about whether these technologies can be used in paternalistic ways.

## Section 1: Responsibility, Liability and Autonomous Vehicles

### Case Study #1: Uber

On March 19th, 2018, a self-driving vehicle developed by Uber hit and killed a 49-year-old pedestrian named Elaine Herzberg in Tempe, Arizona. The car was driving under the speed limit and showed no sign of slowing down. Herzberg was not crossing a designated crosswalk, instead choosing to use a purely ornamental path pedestrians are forbidden from using which was clearly indicated by road signs. Who should be held responsible for the death of Elaine Herzberg? An instinctive response might be to hold the company responsible. After all, if they hadn't developed the car, Elaine might not have been killed. However, Elaine might not have been killed if she had followed the clear rules not to cross the street. Ideally, the Uber car would have slowed down in time. But how could Uber be held responsible for not foreseeing that their car would hit someone in a no-crossing zone? Moreover, if the car was truly self-driving and acting autonomously of Uber's control, then does that mean we should hold *no one* responsible?

There is a lot to unpack here but there are two broad positions that we can adopt here. The first position defended by some philosophers is that these are cases where no one can be held responsible. The second position claims that we can hold the company responsible. Let's start by taking a closer look at what we mean by responsibility and the closely related concept of liability.

*Responsibility and Liability*: *A Primer*

When philosophers discuss the concept of responsibility, one initial distinction they make is between *forward-looking* and *backward-looking* responsibility. In the backward-looking sense, we say that someone is responsible to the extent that they can be held to account for something that has already happened. This is probably the most intuitive sense of responsibility that jumps out at people when they encounter the *Uber* case study. By contrast, we say that someone is responsible in the forward-looking sense if we are ascribing obligations or duties to them that should constrain their future actions. If we imagine a regular case of a person driving a car, we say that they are responsible for driving carefully and owe such duties of care to their fellow passengers, drivers, and pedestrians.

Another crucial distinction is between *moral* responsibility and *legal* responsibility. Though this is contested terrain, our interest in identifying someone as morally responsible is to determine whether they are *blameworthy* for their action whereas the function of legal responsibility is to determine who is *liable* to be punished or to provide remedy to a victim. Though they often go hand in hand, moral responsibility and legal responsibility can come apart. Someone might be legally responsible, but not morally responsible for rear-ending a car even if the driver in front of them braked for no apparent reason. Whether someone is blameworthy is typically thought to involve certain conditions of 'fittingness', for example, that the target of blame directly intended or reasonably foresaw the consequences of their action. Since those conditions are absent, it is not fitting for us to blame the rear-ending driver.

This might sound unsettling but we need some practical way to ensure those that have suffered harm or loss have some form of redress. This is where the law performs an interesting balancing act. It asks whether it is worse for someone to be liable to pay damages who is not completely responsible or, alternatively, whether it is worse that the victim has no means for compensation. Sometimes, a court or legislature might decide that the latter situation is worse. If so, they will deem that to be a fairer distribution of the burdens of harm. This form of legal responsibility is often referred to as *strict* or *no-fault liability* which requires someone to compensate a person for injury regardless of whether one was careless or negligent. This is contrasted with *fault-based* liability where one is liable only to the extent that the victim can prove that someone wronged them intentionally or acted negligently. Armed with this conceptual primer, let's turn to the two positions I mentioned above.

*Responsibility Gaps, Strict Liability and Negligence*

Some philosophers argue that the more autonomous a technology becomes, the more difficult it is to attribute responsibility to anyone when things go wrong. Here is how the argument goes. It is a *necessary* condition (i.e., required) for being held responsible for action that we could have controlled or foreseen that action. A technology that is truly autonomous will act in unpredictable ways; in other words, it will perform actions that we (the engineers or the operators) could not have foreseen or controlled. If we cannot hold either the engineers or the operators responsible for the autonomous technology, then it seems like a *responsibility gap* has been created (Matthias, 2004). Sometimes, philosophers call this the *responsibility trilemma* (a situation where we have to decide between three equally undesirable alternatives) (Sparrow, 2007) . First, it seems unreasonable to hold the engineers responsible because the technology is autonomous by design. Second, it seems unreasonable to hold the operators/commanders responsible for the same reason. Third, we are left with holding the machine responsible and that too seems implausible. Recall that fittingly blaming someone involves holding an attitude that is directed at someone who intended or foresaw some action. We should add that fitting blame requires that the object of our blame can feel a certain moral emotion, namely, guilt (Véliz, 2021). A guilty person experiences a certain kind of pain for having harmed or hurt another. But that level of moral emotion is plausibly lacking in autonomous vehicles and robots which makes it unreasonable to hold them morally responsible for their conduct. Therefore, it seems like no one is responsible for the tragic event in our *Uber* case study. Is that right?

Here are two ways of responding to the responsibility gap position. Intuitively, we want to map a relationship between causal responsibility (who or what caused some event to occur) and moral responsibility (that whoever caused some event to occur is accordingly blameworthy or praiseworthy). In cases where this mapping is strong, we say that responsibility is *identifiable*. However, this mapping could be weak as in the *Uber* case. In such cases, even if responsibility is not easily identifiable, we might think that responsibility is still *assignable*. That is, we can say about responsibility-ascriptions in these cases not that they are correct or incorrect, but that they could be justified or unjustified. This goes back to our discussion about ensuring there are practical remedies for those who have suffered losses. We should consider whether the greater burden is to assign responsibility to these corporations for these accidents or whether it is worse that victims have no means of compensation. In this case, since the victim was killed, the family members would be owed compensation. Cases like this are not uncommon and they map roughly onto the strict-liability conception. It can be justified to hold someone responsible for damages they have caused regardless of whether they intended the action.

However, there is something very puzzling about attributions of strict liability. Notice that this response implies that the company did *not* act morally wrongly.  Of course, it is an impartially good result that victims get compensation from corporations that harm them. But we are conceding that corporations have not acted wrongly in releasing products that could harm consumers and third parties *because* they aren't identifiably responsible. It is often said that a system of strict liability is useful because it incentivizes these actors to avoid *any* accidents that attributable to their products since they will be on the financial hook. This is an empirical observation though which does not defeat the concern that we are taking responsibility out of the moral picture. Philosophers and legal theorists moved by this concern either want to jettison strict liability from tort law or "fit the round peg of strict liability into the square of negligence liability" (Keating, 2014, p. 293). This latter strategy takes us to the second way of responding to the responsibility gap.

Earlier, we mentioned that negligence is related to a fault-based liability. A person's conduct is negligent if it exposes others to a risk of harm that could have been foreseen and reasonably avoided. If it was the case that Uber could have foreseen such risks of harm and did not adequately or reasonably make attempts to avoid such risks, then we have a case where we can hold someone morally responsible *and* claim they are liable to compensate others. Though there is some controversy about the details of the case, the automobile system was found to *not* consider that people might jaywalk or illegally cross a road.[1] Elaine Herzberg and about 78% of Americans admit that they have jaywalked with nearly 40% surveyed saying that they cross when they think it is safe (Moore, 2014). This seems like something that certainly could have been foreseen. While jaywalking is a crime in Arizona, it seems ignorant not to consider a fairly common form of law-breaking that is relevant to the conduct of driving cars operated by humans or autonomously. In either case, strict liability or negligence-based liability enable us to pinpoint who is or could be held responsible in tricky cases like *Uber*.[2] While it might still be puzzling which model of liability is more appropriate, the question is no longer about responsibility but about the *right way to close them*.

## Section 2: Explainability, Legitimacy and Black-Box Algorithms

### Case Study #2: Medical Diagnostics

In medicine, machine learning algorithms often *outperform* human doctors on a number of diagnostic tests for many diseases. One common concern is that these algorithms produce the correct result but are so complex that we cannot discern the explanation for their diagnosis. Should we prefer algorithms that are more accurate but less likely to be explainable or should we prefer human decision-makers that might be less accurate but capable of giving reasons? Whatever answer we give here will have important implications for whether it is wrong for medical companies to produce and sell so-called *black-box algorithms* to hospitals.

First, I will start by explaining what black-box algorithms are. Secondly, I will turn to the moral importance of what philosophers and computer scientists call *explainability*. Third, I will consider whether a double standard is being applied to algorithms because human beings are often just as opaque in their decision-making.

*On Black-Boxes*

Simply put, machine learning is a way in which patterns in large amounts of statistical data are identified by a machine, hence, why we say that the 'machine learns for itself' (Zerilli, 2022). There are several categories of machine learning which form an *approach* to artificial intelligence, the branch of computer science – with no fully-agreed upon definition – referred to as the science of making computers perform cognitive tasks like thinking, learning and predicting on a comparable or superior level to human beings (Liao, 2020; Zerilli, 2021) First, there is *supervised learning* where a dataset is labelled and trained (i.e., fed data, corrected when it make mistakes) to accomplish a specific task such as identify pictures of dogs. *Unsupervised* learning starts with an unlabeled dataset and attempts to sort the data and identify the pattern on its own. In their own category are *reinforcement learning* algorithms

[1] According to the National Transportation Safety Board's report on the case.

[2] To focus our attention on automated technology, I omitted a detail which is that the Uber car had a human back-up driver. The back-up driver was charged with criminally negligent homicide but Uber did not face any charges. Arguably *both* parties – the back-up driver and Uber – were responsible because of the issue mentioned in the text and the fact that the driver was not paying attention at the time of the collision.

which aim to maximize a reward function (i.e., goal) and are rewarded if they succeed or punished if they fail. This form of machine learning has exploded in relevance as it forms the bedrock of large language models (e.g., ChatGPT) which process and generate text based off a huge corpus of data and learn through human feedback to produce 'helpful' outputs. In this chapter, I will not focus on LLM's but I will briefly hint at some interesting ethical questions that underlie their usage in Section 4. For our purposes, supervised learning algorithms are the ones being used frequently as decision-making tools or replacements for human decision-making altogether in many domains of social life.

One subset of these techniques is called *deep learning* which, inspired by the human brain, is a system made up of many nodes and layers densely connected with one another to predict some output for a particular task. Usually, data moves through these networks in one direction ('feed-forward') and while we are able to correct the learning algorithm when it makes mistakes, the mapping between inputs, outputs and huge numbers of intermediate layers is unique. It is often technically unclear how it has arrived at its decision (Zerilli, 2021). The lack of clarity or transparency about how it is 'reasoning' is what people have in mind when they refer to *black-box* algorithms. And the worry is a fairly intuitive one. If I have been diagnosed with a medical condition, denied parole, denied a bank loan etc., I should be provided with some explanation for why this outcome has occurred. This seems to hold true even if the decision is accurate. With this set-up, let's go deeper and try to make theoretical sense of this intuition.

### *The Right to an Explanation*

Some philosophers argue that we have a right to explanation (Vredenburgh, 2022). What is meant by this is usually that explanations serve an interest that is sufficiently weighty to require others to provide said explanations. Notice that not just any old interest in explanation will do. There has to be something significant about the interest that compels others to satisfy (o not prevent the satisfaction of) the right. According to Kate Vredenburgh (2022, p.213), explanations are important because they serve our interests in *informed self-advocacy*. We have fundamental interests in our interests being considered by others, in being able to navigate systems of rules and in being able to remedy mistakes that such complex systems have caused. Failing to be provided with certain kinds of explanations frustrates these interests. This might consist in not providing someone with the causal explanation for some decision or for the reasons that count in favor of using such a system in the first place. The point is that imposing some system of rules on someone that thwarts their reasonable interests is not morally justifiable or, alternatively, it demands a particularly compelling justification.[3]

We can see how this is relevant to the case of black-box algorithms. Surely, if am denied parole, I am owed an explanation for why this occurred beyond the judge's insistence that the "computer says 'no'". How am I able to navigate an already complicated justice system if I cannot even understand why I am being compelled to remain in prison? How can I check that the algorithm made the right decision if even the judges and computer scientists don't understand what caused that to happen? These two questions suggest that there is a clear *epistemic* requirement for me to able to advocate for my interests. If this interest successfully grounds a right to explanation, then we have the tools to make a judgment about whether it is legitimate or permissible for us to deploy these algorithms in decision-making. If others have a right grounded in important human interests, this acts as a *constraint* on what

[3] Vredenburgh develops the argument in terms of a contractualist moral theory as found in the work of T.M. Scanlon.

we may to do them. Therefore, we ought not use black-box algorithms, other things being equal.

### *Explainability versus Accuracy*

While there is something intuitive about postulating a right to explanation in the case of parole decisions, it is not clear that this account works across all contexts in which algorithmic decision-making occurs.[4] In other words, it may be that in certain domains explainable algorithms are more important than in others. In those other domains, what truly matters is that we get the correct result regardless of how the result was reached. The practice of medical diagnosis and prescription is arguably one such domain.

What is puzzling about knowledge in medicine is that many of the underlying causal mechanisms are not well-understood. This renders many of the decisions made by doctors atheoretical and similarly opaque to black-box algorithms (London, 2019). A well-worn example is that doctors prescribed aspirin as an effective painkiller for a century knowing *that* it worked but not knowing *how* it worked. Surely, from the standpoint of the patient, that some treatment is effective is more important than understanding how it works. But, in this respect, human diagnosticians and algorithms are in roughly the same boat. Moreover, some argue that algorithmic decision-making in this domain is likely to be more accurate precisely *because* it dispenses with the demand for explainability in the form of some possibly false causal theory. Alex John London (2019, p.18) cites the example of how doctors used to perform mastectomies over more mild interventions because of a theory that removing as much as tissue as possible would reduce the likelihood of breast cancer reoccurring. This theory was shown to be false even though it sounded theoretically plausible and is intuitively explainable to patients. The basic point is that privileging explainability over accuracy is particularly mistaken in medicine given the history and mode of reasoning in that particular field. This point will be relevant to Section 3 when we consider what kinds of decisions algorithms should or should not be allowed to make.

### *AI and Human Decision-making: A Double Standard?*

The comparison between doctors and medical diagnostic algorithms is meant to illustrate a deeper point. Demands that AI ought to be explainable should be consistently applied to human decision-makers and to do otherwise would reveal an arbitrary and possibly harmful double standard, especially if explainability is treated as a *constraint* on their usage (Zerilli *et al.*, 2019). To be clear, the claim that we currently apply an irrational double standard for explainability to AI is not meant to suggest that explainability is unimportant. The medical example can lead one to think that accuracy ought clearly to trump explainability, but there may be contexts (like the parole case) where explainability is more important. Alternatively, one might think that as the practical stakes of a decision get higher, so too should the demands for explainability. But making our demands for transparency sensitive to the stakes (as we plausibly should) does not necessarily speak in favor of either a human or algorithmic decision-maker.

Interestingly, recent advances in the field known as explainable or XAI seem to show that we can get a very good approximation of how some deep learning algorithm came to its decision by training an algorithm to reproduce its performance (Zerilli, 2021, p. 37). These ‘models-of-models’ can reveal the primary weights used by the original algorithm and give decision-makers the ingredients of an explanation. Perhaps this might show that we can get more reliable explanations for decisions from AI. This is possible if, as seems empirically

[4] To be clear, Vredenburgh does not intend for her argument to apply to all such contexts.

demonstrated in cognitive science, that human beings are riddled with unconscious biases and often appeal to *post hoc* rationalization that does not reflect the motivation behind their original decision. In short, we have compelling reasons to think both kinds of decision-makers are equally legitimate in terms of their capacity to explain their decisions to others.

Some philosophers reject this view and think there is something significantly different about human-given explanations. For example, Seth Lazar finds the view of human rationality that underlies the 'double standard' challenge to be overly pessimistic. In his view, he stresses that a necessary condition for morally legitimate decision-making is that it is possible for us to see and grasp how such decisions were made (Lazar, 2024). This ties the idea of explainability to what political philosophers call a *publicity* condition (that justice must also be *seen* to be done). His objection is that those who press the double standard worry overemphasize the importance of the 'luminosity' of our mental states for legitimate decisions. In his view, "explanations show how decisions were made: what procedures were followed; what evidence was used; what rationale was presented… Human decision-makers in institutional settings can explain their decisions by addressing these questions without analysing their private motivation."

Lazar is right to tie the idea of explainability to publicity for legitimate decision-making, but his objection misses the point. The problem is that those features of human decision-making are likely to be corrupted at their *source* by bias, post-hoc rationalization and other sub-doxastic factors. Whatever other salutary features these explanations have (such as publicity) are simply made less reliable and less accurate as truthful answers to the question of explainability*, i.e., why did you make the decision*? If explainability is suitably linked to publicity, it may be that the recent successes of XAI will give algorithms a better chance of faithfully explaining why they made particular decisions. All this underscores how complicated debates about explainability are and how reasonable disagreement exists around its moral importance.

## Section 3: Algorithms Need Not Apply: Hiring, Noise and Human Decision-making

### Case Study #3: Hiring and Interviews

Recently, a large body of evidence has shown that interviews are a bad mechanism for predicting who will be the best candidate for a job (see studies cited in Bhargava and Assadi, 2024). One set of concerns focuses on bias in hiring, but another surprisingly common one is noise. According to Kahneman, Sibony and Sunstein (2021) noise occurs when there is "unwanted variability" in human judgment. For example, we might look at the same kind of candidate and make a different decision if we happen to be hungry, tired or if it is Friday. Alternatively, one of the interviewers might think the candidate should definitely be hired whereas the other vehemently disagrees. Algorithms don't get tired or hungry or inconsistent; in other words, they are noiseless. What value, if any, is lost when we use algorithms to replace human interviewers?

*Treating Like Cases Alike (or, What's Wrong with Noise)*

To get a little clearer on the difference between noise and bias, consider firing shots or arrows at a target. One player might hit all of their shots in one corner whereas the other player hits all over the target. Neither player hits the bullseye but for different reasons. The first player represents bias because their shots are systematically skewed in one direction. The second player represents noise because their shots deviate from the target in unpredictably varied ways.[5] Noise is the conceptual (or statistical) cousin of bias and we tend to have an intuitive grasp of what is wrong with biased decision-making, namely, it is *unfair*. The most high-profile case of this was the COMPAS algorithm used to assist the criminal justice system. It was found to predict that black people were more likely than white people to be incorrectly classified at a high-risk of re-offending.[6] How does fairness relate to noise though? If our worry is that certain factors like hunger, mood can make decisions on similar candidates vary wildly, then we probably have the following view of fairness in mind: *always treat like cases alike* (Sunstein, 2022).

This view of fairness is a very old one stretching back all the way to Aristotle. It is extremely intuitive. There are at least two aspects of the view that we need to make sense of. First, what does it mean for like cases to be alike? Second, why is it good to treat like cases alike? Of course, it may seem like we already have an answer to this question: it is good to treat like cases alike because that is fair. But as I point out in footnote 6, there is reasonable disagreement about what fairness consists of. So, we should want to see if there are other considerations besides stating that it is good to follow this rule because it is fair. There are at least two broad ways in which it could be good to 'treat like cases alike'. It could be instrumentally valuable, i.e., it produces good results. Alternatively, it could be *intrinsically* valuable, i.e., it is good in itself.

Consider the first question. It may seem trivial to point out but like cases will never *strictly* be alike in every respect. No two job candidates will literally be the same, however similar. Therefore, we must mean something different with this maxim. Perhaps the view is that there are plausible features for which cases *should* be treated alike. In other words, we should want to exclude traits that are *morally irrelevant* from our comparison of two cases. This is easy to do if two candidates are identical but one was interviewed on Tuesday and another on Wednesday. The day on which a candidate is interviewed is morally irrelevant and so these like cases can be treated alike. In general, the context of choice will play an important role here because that may modify the ease with which these comparisons can be made. This is an important point I will come back to in the next sub-section.

The second question is more substantive. Some argue that treating like cases alike is creates better outcomes. Certainly, this is what Kahneman et al., (2021, p.17) have in mind when they cite the infamous case of "refugee roulette", where one judge admitted 88% of asylum seekers whereas another admitted 5% (where these cases were alike in relevant respects). Another benefit of treating like cases alike is the way in which it increases accountability (Strauss, 2002, p. 16). Knowing that decision-makers adopt this consistency-based rule is incredibly useful because it allows us to challenge deviations in a simple way. The refugee roulette case is also one where there is too much discretion afforded to these decision-makers. Rules, like the maxim of treating like cases alike, can silence this noise which points to the maxim's important instrumental value.

[5] Bias and noise occur simultaneously but I omit these cases for clarity of exposition.

[6] There is serious disagreement about this case precisely because of the different conceptions of fairness at play. Obviously, there is something unfair if we think that fairness consists of equalizing the rate of false positives and false negatives across two groups. However, COMPAS defended its algorithm by showing it satisfied an alternative conception of fairness called *calibration* where risk scores mean the same regardless of one's ethnicity (where we try to omit entering 'protected attributes' into the data, features of persons that should not be discriminated against). Obviously the difficulty arises because black people are a smaller proportion of the population than white people, so, by definition, we cannot satisfy both of these criteria simultaneously (e.g., trying to promote equal error rates across groups means the predictions will be inaccurate across one of the groups)(Kleinberg, Mullainathan and Raghavan, 2016).

What is the intrinsic value of treating like cases alike? This view expresses or embodies respect for the value of equality. The philosopher Bernard Williams (1973, p. 231) puts the point clearly when he says this amounts to a claim that "for every difference in the way [one is] treated, some general reason or principle of differentiation must be given". We should treat like cases alike unless some justification in the form of a compelling moral reason can be given for deviating from (this conception of) equality. There is something deeply intuitive about this view. What is effective about appealing to this value is that it defends the maxim regardless of the results it produces. It is good and fair to treat like cases alike because that respects equality.

There is an important upshot of these arguments. As Kahneman et al., (2021, p. 18) provocatively claim, "Wherever you look at *human* judgments, you are likely to find noise [own emphasis]". The prescription, where possible, is to replace noisy human judgment with definitionally noiseless algorithms which will always treat like cases alike. Algorithms will not grant different asylum decisions depending on the day of the week. They are consistent and incapable of operating with discretion. If we want to avoid unequal and unfair treatment, the message is that algorithms or rules are the way to go.

### *Is Noise Always Bad?*

This question might seem strange. After all, Kahneman et al., (2021) explicitly define noise as *unwanted* variability. One could be tempted to just conclude that all noise is bad. But this definition is misleading and problematic. Recall Williams' observation above. The tacit assumption in Kahneman et al.'s argument is that all of these are cases where *no justification* exists for treating like cases *differently*. This assumption makes a lot of sense in the refugee roulette and possibly some cases of job interviews. But variability is not always undesirable.

First, consider a fairly trite counterexample. If an algorithm is used by a foreign power (with whom we're at war) to indiscriminately kill all civilians, it treats like cases alike. It does not discriminate, does not use discretion and is therefore noiseless. Obviously, there would be something absurd about claiming that eliminating noise here would be a good thing to do. In this case, we want the foreign power to be more discerning and – depending on one's views of just war theory – exercise restraint and care when selecting targets. They should not treat every person the same, they should treat combatants differently to noncombatants. So literally speaking, variability cannot *always* be unwanted.

The more sophisticated way of making this point appeals to a point I made earlier about how the context of choice or judgment makes a normative difference. Ruth Chang helpfully distinguishes between what she calls 'messy' and 'neat' cases of judgment. In a neat case, we can evaluate judgment against a single bullseye and easily determine whether a mistake has been made. The refugee roulette is the best example of a neat case. Alternatively, missing a tumor because one is tired is an example where a clear mistake in judgment has occurred. In these cases, an algorithm should be put in place.[7] By contrast, messy cases are those where there are *multiple* bullseyes, that is, multiple legitimate targets which *require* human judgment to choose between them.

[7] The refugee roulette case might be messier though especially if we think there can be value to discretion in legal adjudication. One such value is mercy. Typically, mercy is a form of discretion that judges enact when sentencing a criminal less harshly than they deserve according to the rules of the justice system (Tasioulas, 2003). Perhaps merciful judgment is an expression of a uniquely human sensitivity to the special circumstances of defendants that law would treat too bluntly. As David Strauss (2002) points out, rigid rules can sometimes make *unlike* cases be treated alike in a way that distorts our judgment of them.

Hiring candidates for a job is great example of this. What is it that makes for a good employee? Perhaps we think a good employee is efficient and makes a company the most money possible. But robotic employees like this might be bad team-workers causing internal friction in a company. Maybe then a good employee is one that has great interpersonal skills. But it's not a good idea to just hire friendly people. Perhaps this is why we need human interviewers even when they might disagree about who to hire. One interviewer might have caught onto the fact that a person lacks interpersonal skills despite being very efficient. Perhaps they weigh up these factors differently than the other interviewer. The deliberations between interviewers bring out these considerations; they learn from one another and it is precisely *because of* variability in their judgment that we get a great employee. This is at least one instrumental benefit to sticking with human judgment even with its noisiness.

But there are other non-instrumental values realized in human judgment. Bhargava and Assadi (2024, p.216) point out that there might be 'representative value' in *us* being the ones choosing whom we work with even if they turn out not to be the best employee. It is valuable-in-itself for us to see features of our choice represented in the world around us and again this can only be manifested through allowing discretionary and admittedly noisy human decision-making. Algorithms might pick the 'right' employee (along one of the bullseyes), but the employee might not be *right for us* precisely because the act of choosing and human judgment can changes the nature and value of our decisions (Scanlon, 1998).

### *Going Forward: Right to a Human Decision?*

In the previous section we considered whether there was a right to explanation. Do the considerations offered in this section offer support for a right to a *human* decision? Such a right might kick in as enabling persons to pursue an *ex-post* appeal of a decision that was made by a machine. This is the subject of recent legal scholarship in the European Union and is increasingly being discussed in philosophical and other academic circles (Huq, 2020; Shany, 2023). Though I cannot address this question here in great detail, there are many relevant instrumental considerations such as feasibility and effectiveness as well as non-instrumental considerations such as standing in equal or solidaristic relations with human decision-makers who preside over us (Tasioulas, 2023). In a world where an increasing number of choices will likely be delegated to algorithms, the question of whether we have rights to human decision-makers will be a fraught one.

## Section 4: Overview and the Rise of LLM's

In this chapter, we've focused on the challenges posed to responsibility, legitimacy and human decision-making by automated technologies and artificial intelligence. In Section 1, we covered the debate about the alleged responsibility gaps created by autonomous technologies like self-driving cars. In Section 2, we looked at black-box algorithms and how their inability to provide explanations for their outputs seems in tension with the legitimacy of their usage. Finally, in Section 3 we took up the question of whether there are contexts where we should take humans out of the equation and replace them with noiseless algorithms. In each of those sections, I sought to show that many of these worries can either be tackled using the tools of philosophy or at least that we can make some progress on them.

In this final section, I want to briefly look at the most recent uses of AI, namely, large-language models. There is always a risk in writing about the ethics of modern technologies that one says becomes obsolete or overshadowed by some other important developments. Nevertheless, we can still point to some interesting questions already posed by these rapidly adopted technologies.

We briefly touched on how large-language models work by using neural networks to process and generate natural language based off a large corpus of data. There are many fascinating issues that one could take up here. For example, have these massive amounts of data been gathered in a way that is compliant with copyright law? How can copyright violations be proved? Another issue is how these technologies were simply unleashed on the public. We were not given time to handle the impact they have on important social institutions such as our educational system. The concerns here are not just about cheating but about how they affect the process of learning altogether as well as our capacities for creativity. If we come to rely on these technologies in all of our writing tasks and LLM's rely on our usage of them, then we might risk a homogenization and stagnation in the ideas we come up with as a species. One other usage of LLM's that worries people is their supposed capacity to spread misinformation. Now, concerns about misinformation are not unique to these technologies but it is worth noting that LLM's could and have been found to durably reduce belief in conspiracy theories (Costello, Pennycook and Rand, 2024).

One ethical concern I have about the use of these technologies lies in their power to persuade and paternalistically interfere with our inquiries. Already, these technologies are being touted as the possible co-pilots to the Internet. In time, it will become increasingly difficult for persons to even look up supposed falsehoods. This may sound unproblematic if the rationale for blocking exposure to misinformation is to prevent harm to others. Nevertheless, vesting massive corporations with the power to shape what we believe must meet a very high burden of justification. For one thing, the sheer level of influence that a few US corporations already have over everyday life is cause for concern.

Secondly, we should always be wary of failing to treat autonomous adults with respect and this extends to the inquiries they might wish to make online. If our worry is that they will come to believe and act on misinformation, that is an empirical claim that *has to be proved*, not assumed. There is a great deal of skepticism about our capacities to discern what is true and false but it is simply at odds with an enormous amount of research on our evolutionarily evolved capacities for epistemic vigilance (Mercier, 2020). In a different vein, it is strange to insist that it is categorically bad for people to simply have false beliefs. We all have false beliefs and it is hardly the responsibility of private companies to use AI to correct those beliefs. An interesting area of future research might be to examine the ways in which we use LLM's to outsource paternalistic activity. Outsourcing otherwise morally problematic actions to entities incapable of action does something strange to the moral status of an action and our ability to make moral claims on others (see Semler (MS) for more).[8] In short, the rise of LLM's in social life will force us to ask difficult moral questions which actions we may permissibly delegate to automated technologies and AI.

---

**References**

Bhargava, V.R. and Assadi, P. (2024) 'Hiring, Algorithms, and Choice: Why Interviews Still Matter', *Business Ethics Quarterly*, 34(2), pp. 201–230. Available at: https://doi.org/10.1017/beq.2022.41.

Costello, T.H., Pennycook, G. and Rand, D.G. (2024) 'Durably reducing conspiracy beliefs through dialogues with AI', *Science*, 385(6714), pp. 1–12. Available at: https://doi.org/10.1126/science.adq1814.

[8] For a great start on this topic, see Eggert (2023).

Eggert, L. (2023) ‘Autonomised harming’, *Philosophical Studies* [Preprint]. Available at: https://doi.org/10.1007/s11098-023-01990-y.

Huq, A.Z. (2020) ‘A RIGHT TO A HUMAN DECISION’, *Virginia Law Review*, 106, pp. 611–678.

Kahneman, D., Sibony, O. and Sunstein, C.R. (2021) *Noise: a flaw in human judgement*. London: William Collins.

Keating, G.C. (2014) ‘Strict Liability Wrongs’, in J. Oberdiek (ed.) *Philosophical Foundations of the Law of Torts*. Oxford University Press, pp. 292–311. Available at: https://doi.org/10.1093/acprof:oso/9780198701385.003.0015.

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2016) ‘Inherent Trade-Offs in the Fair Determination of Risk Scores’. arXiv. Available at: http://arxiv.org/abs/1609.05807 (Accessed: 12 November 2024).

Lazar, S. (2024) ‘Legitimacy, Authority and Democratic Duties of Explanation’, in S. Wall and D. Sobel (eds) *Oxford Studies in Political Philosophy*. Oxford: Oxford University Press, pp. 28–56.

Liao, S.M. (2020) ‘A short introduction to the ethics of artificial intelligence’, in *Ethics of Artificial Intelligence*. Oxford: Oxford University Press, pp. 1–42. Available at: https://doi.org/10.1093/oso/9780190905033.003.0001.

London, A.J. (2019) ‘Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability’, *Hastings Center Report*, 49(1), pp. 15–21. Available at: https://doi.org/10.1002/hast.973.

Matthias, A. (2004) ‘The responsibility gap: Ascribing responsibility for the actions of learning automata’, *Ethics and Information Technology*, 6(3), pp. 175–183. Available at:

Mercier, H. (2020) *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton, NJ.: Princeton University Press.

Moore, P. (2014) *Almost everyone jaywalks, but many still want it to be illegal*, *YouGov*. Available at: https://today.yougov.com/society/articles/10571-almost-everyone-jaywalks.

Scanlon, T.M. (1998) *What We Owe To Each Other*. Cambridge, MA.: Harvard University Press.

Semler, J. (MS) ‘Artificial Moral Behavior’.

Shany, Y. (2023) ‘The Case for a New Right to a Human Decision Under International Human Rights Law’.

Sparrow, R. (2007) ‘Killer Robots’, *Journal of Applied Philosophy*, 24(1), pp. 62–77.

Strauss, D.A. (2002) ‘Must Like Cases Be Treated Alike?’, *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.312180.

Sunstein, C.R. (2022) ‘Governing by Algorithm? No Noise and (Potentially) Less Bias’, *Duke Law Journal*, 71(6), pp. 1176–1205. Available at: https://doi.org/10.2139/ssrn.3925240.

Tasioulas, J. (2003) ‘Mercy’, *Proceedings of the Aristotelian Society*, 103, pp. 101–132.

Tasioulas, J. (2023) ‘The Rule of Algorithm and the Rule of Law’, in C. Bezemek, M. Potacs, and A. Somek (eds) *Vienna Lectures on Legal Philosophy, Volume 3*. Bloomsbury, pp. 17–40.

Véliz, C. (2021) ‘Moral zombies: why algorithms are not moral agents’, *AI & SOCIETY*, 36(2), pp. 487–497. Available at: https://doi.org/10.1007/s00146-021-01189-x.

Vredenburgh, K. (2022) ‘The Right to Explanation*’, *Journal of Political Philosophy*, 30(2), pp. 209–229. Available at: https://doi.org/10.1111/jopp.12262.

Williams, B. (1973) *Problems of the Self : Philosophical Papers 1956–1972*. Cambridge: Cambridge University Press.

Zerilli, J. *et al.* (2019) ‘Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?’, *Philosophy & Technology*, 32(4), pp. 661–683. Available at: https://doi.org/10.1007/s13347-018-0330-6.

Zerilli, J. (2021) *A Citizen’s Guide to Artificial Intelligence*. Cambridge, MA.: MIT Press.

Zerilli, J. (2022) ‘Explaining Machine Learning Decisions’, *Philosophy of Science*, 89(1), pp. 1–19. Available at: https://doi.org/10.1017/psa.2021.13.