Synthese (2021) 199:11423–11443
https://doi.org/10.1007/s11229-021-03296-0

ORIGINAL RESEARCH

Ethical machine decisions and the input-selection problem

Björn Lundgren[1,2,3]

Received: 29 March 2021 / Accepted: 30 June 2021 / Published online: 30 August 2021

Abstract
This article is about the role of factual uncertainty for moral decision-making as it concerns the ethics of machine decision-making (i.e., decisions by AI systems, such as autonomous vehicles, autonomous robots, or decision support systems). The view that is defended here is that factual uncertainties require a normative evaluation and that ethics of machine decision faces a triple-edged problem, which concerns what a machine ought to do, given its technical constraints, what decisional uncertainty is acceptable, and what trade-offs are acceptable to decrease the decisional uncertainty.

Keywords Machine decisions · Uncertainty · Factual uncertainty · AI ethics · Data choices · Input choices · Input-selection problem · Trade-offs · Data protection · Privacy · Transparency · Opacity · Cost–benefit · Time-sensitive machine decisions

1 Introduction

Uncertainty yields problems for moral decision-making in at least two ways. First, we have the issue of 'moral uncertainty', which is uncertainty about which normative principles should guide what we ought to do.[1] Second, we have the issue of 'factual uncertainty', which is uncertainty about the (possible) state of affairs or the

[1] See, e.g., Bykvist (2017) for an introductory overview. See also MacAskill et al. (2020) for a recent substantial treatment of the issues of moral uncertainty (cf. fn. 3).

This article belongs to the topical collection "Indeterminacy and Underdetermination", edited by Mark Bowker and Maria Baghramian.

✉ Björn Lundgren
bjorn.lundgren@umu.se; bjorn.lundgren@iffs.se; bjorn.lundgren@philosophy.su.se

[1] Department of Historical, Philosophical and Religious Studies, Umeå University, Umeå, Sweden

[2] Institute for Futures Studies, Stockholm, Sweden

[3] Department of Philosophy, Stockholm University, Stockholm, Sweden

(possible) consequence(s) of actions.[2] Both forms of uncertainty affect the question of what we ought to do in a given situation. Moral uncertainty because we do not know which principles ought to guide us, or which facts matter for decision-making. Factual uncertainty because we do not know all morally relevant facts about the given situation. In this article, I will focus on factual uncertainty, as it concerns the ethics of machine decisions (by which I mean decisions made by any input/output machine, but the focus will be on AI systems).[3]

In ethics, a common method for dealing with factual uncertainty is to first analyze idealized cases. Once we know what we ought to do in idealized cases, we can analyze what to do based on a theory of rational decision-making for situations involving factual uncertainty. That is, the *ethical analysis* can be restricted to the idealized cases. I will call this method 'the standard approach'.[4]

Contrarily, some argue that the standard approach is problematic because factual uncertainty changes the normative evaluation of a given situation.[5] Hence, if we want to determine what we ought to do in a situation involving factual uncertainty, we cannot idealize; rather we must analyze the situation simpliciter, including the associated factual uncertainty. I call this 'the uncertainty approach'.[6]

[2] See, e.g., Greaves (2016) and Hansson (2003). It is worth to point out that some hold that there is a third category of uncertainty: modal uncertainty (see Bradley & Drechsler, 2014). That is, if we think that there is a distinction between uncertainty about what *is* the case (what Bradley and Dreschler call 'empirical uncertainty') and what *could* be the case (i.e., modal uncertainty). I take it that Bradley and Dreschler's view may be regarded as somewhat controversial for various reasons (i.e., while there is a clear distinction between what is and could be the case, it may be sensible to lump our epistemic uncertainty about actual and possible states-of-affairs under one concept). For example, it seems reasonable to think that if we increase an agent's uncertainty about what is the case, the agent might eventually become uncertain about what could be the case. Moreover, if an agent is uncertain about what could be the case, then they necessarily do not know what is the case (i.e., if we are talking about the same proposition). Anyway, I remain neutral as to whether we should accept Bradley and Dreschler's distinction, but—as you can see in the text—I have included both modal and empirical uncertainty under the heading of factual uncertainty, since this will not affect the arguments as such. However, even if I lump what Bradley and Dreschler claim are distinct forms of uncertainty under one concept, I think my examples will illustrate that the distinction may nevertheless be important in the sense that modal uncertainty increases the decisional problem.

[3] Recently, MacAskill and Ord (2020, p. 350) argued that in cases "where the magnitudes of choice-worthiness can be compared across theories, normative and empirical uncertainty should be treated in the same way" (here empirical uncertainty can be understood as factual uncertainty, cf. fn. 2). Since I do not think any argument in the present article turns on this issue, I will not consider their arguments or conclusions. (*N.B.,* the arguments of MacAskill and Ord can also be found in Chapter 2 of MacAskill et al. (2020).

[4] It is not unreasonable to think that rationality ought to be analyzed *normatively*, but that does not necessary imply that the normativity is ethical.

[5] See, e.g., Hansson (2003).

[6] There are other possible ways to distinguish the methodological disagreement. For example, one may agree with the uncertainty approach that risks and uncertainties must be normatively evaluated, but one may hold that it need not be evaluated *in context*. That is, one may hold that we can follow the standard approach in analyzing the idealized cases and then separately resolve the normative aspects about factual uncertainty. For simplicity, I will not consider such variants.

We can see this distinction plays a role in the applied debate on the ethics of machine decisions, irrespective of whether one of these approaches is *explicitly* endorsed. For example, there is a large literature on the ethics of crashing with autonomous vehicle(s), which is concerned with the ethics of machine decision-making for autonomous vehicles in situations of an unavoidable crash.[7] In this context, proponents that explicitly or implicitly adhere to the standard approach mostly discuss so-called 'applied trolley problems'. These exist in many variants but most commonly involve one autonomous vehicle facing an unavoidable accident, involving a bivalent choice with idealized factual descriptions.On the other side, promoting the uncertainty approach, are those arguing that the idealization of these applied trolley problems is problematic because (1) they ignore risk uncertainties, and (2) risk uncertainties must be normatively evaluated. If we want to give an answer to what principles ought to guide autonomous vehicles' decision-making in these situations, then it is necessary to determine which approach that should be applied.[8]

In this article, I will use examples from the discussion on the ethics of crashing to make a broader point about the methods used in applied ethics to address the ethics of machine decisions (i.e., AI decision-making). Thus, the point I am making here is not really concerned with the ethics of crashing, rather it concerns the broader methodology for addressing the ethics of machine (AI) decision-making.

On the view I will defend, I concur with the uncertainty approach that we must *normatively* analyze the uncertainty component of machine decisions. Yet, I will argue that this analysis is insufficient because we are dealing with a moving target. That is, the ethics of machine decisions involve a triple-edge problem. First, the question of machine decision is not only a question about what a machine ought to do in situations (of uncertainty), given its technical limits, but it also concerns how the machine needs to be constituted to achieve the right decisions. That is, what inputs are needed to achieve the ethically correct—or a sufficiently correct—decision? Second, it concerns the question about how much decision uncertainty we can accept and what inputs we need to achieve that. Third, given that increasing inputs in most cases implies various trade-offs or risks thereof, the question is what trade-offs are justified for reducing that decisional uncertainty? Thus, the ethics of machine decisions is a moving target in so far as all three aspects of the problem involve the question of how the machine ought to be constituted, because how the

[7] See, e.g., Awad et al. (2018), Borenstein et al. (2019), Casey (2017), Cunneen et al. (2018), Goodall (2014, 2016), Hern (2016), Himmelreich (2018), JafariNaimi (2018), Keeling (2020a), Lin (2013, 2015), Lundgren (2020a), Mirnig and Meschtscherjakov (2019), Nyholm and Smids (2016), Santoni de Sio (2017), and Wolkenstein (2018). See Nyholm (2018) for an overview. *Nota bene,* some writers on the topics that I discuss might not be aware of the distinction I make here. Many of these proponents may nevertheless imply one view or the other, but the aim here is not to identify specific proponents of different camps; rather the point is that this distinction is important for the issues being debated. For that reason, I have not sorted these references into the two categories (i.e., the standard approach and the uncertainty approach).

[8] One may find reliance on these types of examples problematic for various reasons. For example, given that the discussion on the ethics of crashing has been criticized for being irrelevant (e.g., because these unavoidable crashes are rare—see, e.g., Lundgren, 2020a). While I agree with this criticism, this is not a problem as long as the examples I use can be generalized to a broader context.

machine is constituted affects its decision-making abilities and—at the same time—it can yield potential harms (or so I will argue). This trilemma is what I call the 'input-selection problem', which concerns the question of which inputs that are needed (for ethical decision-making with sufficient certainty) and which inputs are acceptable (granted the possible harms of using those inputs). The conclusion of this article is that ethical machine decisions have to be analyzed as a response to the input-selection problem.

The remainder of the article will be structured as follows. In Sect. 2, I will discuss the standard and uncertainty approaches and I will present some brief reasons why I think we ought to prefer the uncertainty approach to the standard approach. In doing so I will also argue for the importance of inputs in order to yield the right decisions. In Sect. 3, I will introduce what I will call 'the grandma problem', which aims to illustrate that reducing decisional uncertainty involves trade-offs. My point is not that these trade-offs are necessary (examples of AIs that are relatively harmless include, for example, AIs playing chess), but that they are common for many types of AI applications and hence an ethical evaluation machine decision must address such potential trade-offs. In Sect. 4, I sketch an idea of how ethics of machine decision should proceed in light of these insights. The article ends with a section summing up the main conclusions.

Lastly, in the article I will use terms such as 'data' and 'information' in a fairly colloquial sense (i.e., I will not clearly distinguish them). While more precise definitions, which clearly distinguish these concepts, are available in the literature, those definitions and distinctions will not matter for the issues I am addressing; hence, I will set that aside. Moreover, to simplify the language, I will often refer to *normative ethical questions* as *normative questions* (and likewise for similar formulation), even if normative questions are not limited to ethics.

2 The standard or the uncertainty approach?

In this section, I will defend the uncertainty approach against the standard approach. Moreover, the aim is also to establish that a machine needs to have access to the right inputs in order to achieve an acceptable level of certainty that the machines' decisions are ethically correct. As noted in the introduction I will use the literature on the ethics of crashing as an illustrative example.

Early writers on the ethics of crashing, such as Patrick Lin, argued that autonomous vehicles will unavoidably crash and, thus, we must determine how they should crash. In determining how the vehicle should crash, it is argued that the vehicle will face choices such as that between crashing into a kid or a grandmother.[9] As I

[9] This example is a fairly common in the literature (see, e.g., Lin 2015, p. 70; Awad et al., 2018; Hern, 2016). As previously noted, it is not always evident that the authors take a clear position in favor of either the standard or the uncertainty approach; I have therefore avoided trying to sort the literature into the two camps. Instead, I hope that I have been clear about the fact that the literature allows for some ambiguous readings (see fn. 7).

mentioned in the introduction, these unavoidable accident scenarios are often called applied trolley problems due to their similarity with, and inspiration from, the trolley problem.[10]

Critics have pointed out that applied trolley problems are missing the important fact that AIs are dealing with probabilities. For example, Sven Nyholm and Jilles Smids note that "[an autonomous vehicle] necessarily has to work with […] estimate[s]".[11] One way to read these critiques is to take their claims to be that we must include (ranges of) probabilities into our ethical evaluation. Yet, proponents of the standard approach would not necessarily deny that. Arguably, this is precisely how Geoff Keeling has responded to their arguments. Keeling defends the standard approach, arguing that we first ought to settle the question of *what utility is* and then it is a matter of decision-making under risk (or uncertainties), in which case Keeling endorses *expected utility maximization*.[12]

Although Keeling's point may satisfy some critics, it seems that it does not address the alternative reading of Nyholm and Smids—that risks and uncertainty must be normatively evaluated (as I have pointed out previously).[13] That is, another way of reading these critiques is that they adhere to the uncertainty approach. Given that Nyholm and Smids refer to Sven Ove Hansson—who strictly defends the uncertainty approach—this is arguably the best interpretation of their arguments.

So far, I have not said much to favor either the standard or the uncertainty approach. However, in the context of the ethics of machine decision-making, I think there are several reasons why I think we need to opt for the uncertainty approach. I will present three main reasons below; while doing so I will also establish the importance of inputs for achieving an acceptable level of certainty in decision-making.

First, there is moral uncertainty about how to evaluate risks and uncertainties. That is, the standard approach seems to ignore that a large set of questions about decisions under risk and uncertainty must be normatively evaluated (in context). For example, is there a *pro tanto* right against risk exposure or should the evaluation of risks be done according to purely consequentialist principles? Does fairness in distribution of risks and rewards matter?

Remember that the idea behind the standard approach is that we can resolve all normative questions in idealized cases (i.e., uncertainties and risks can be handled by a theory of rational decision-making), but there seems to be substantial normative questions concerning decision-making under factual uncertainty that require

[10] Thompson (1985), inspired by Foot (1967), is often listed as a standard example of the trolley problem. It is worth pointing out that the similarities to trolley problems are somewhat superficial given that many applications for the trolley problem do not apply to autonomous vehicles. For example, trolley problems are sometimes used in studies of moral psychology and/or to illustrate a distinction between action and inaction (none of which applies to autonomous vehicles).

[11] Nyholm and Smids (2016, p. 1286). Cf. fn. 7 for further reference (although the list includes many proponents of the standard approach).

[12] Keeling (2020a, p. 300). In all fairness to Keeling, it should be mentioned that he seems to have taken a different perspective more recently, in his PhD-thesis (see Keeling 2020b). Thanks to anonymous reviewer for bringing this to my attention.

[13] Lundgren (2020a).

contextual analysis. For example, what makes a machine decision right or wrong arguably depends on what factual uncertainty is normatively acceptable, which is a normative question. To see this, suppose that we are determining an appropriate speed of an autonomous vehicle in a given situation. How fast we ought to drive arguably depends on the uncertainty of the machine's ability to identify and avoid objects in proximity of its travel path and to determine whether these objects are trees, pedestrians, dogs, or other vehicles, and so forth. However, the answers to those questions are not static. That is, in order to determine the appropriate speed in a given situation, we must normatively evaluate how much factual uncertainty is acceptable in the given context (e.g., how certain do we need to be that A will not crash into x in situation S?). This speaks in favor of the uncertainty approach, because it means that factual uncertainty changes the normative evaluation of the situation and that it cannot be separately analyzed.[14]

Second, while we might be tempted to think that a measure such as the one suggested by Keeling is acceptable when we are talking about decisions under known (ranges of) probabilities, we must also keep in mind that uncertainty also includes the possibility of information gaps (and potentially false information). No machine can input all possible data; there are and always will be technical constraints. Because of these constraints, it is possible that ethically relevant data inputs are missing from the machine. Moreover, it is possible that ethically relevant data cannot be predicted from other available information. This is problematic, because even if the machine makes the *perfect decision based on the available inputs*, it can still make the wrong decision, because the available inputs are flawed (i.e., a perfect reasoner may fail if she is reasoning based on false information).[15] This example illustrates that decisions based on flawed information (information that may even be false) creates special normative challenges.

Let us pause for a moment to look more closely at the details of this example. To do so, I will create a simplified thought example. Suppose that we know what the correct fact-relative ethical theory is. That is, suppose that we have a complete description of what is right/wrong "in the ordinary sense if we knew all of the morally relevant facts".[16] Moreover, suppose that we could program, or train, an AI to

[14] An example that illustrates this problem is the tragic accident that occurred in Temple (Arizona, United States), in which an operator controlled "automated test vehicle struck and fatally injured a 49-year old pedestrian" (National Transport Safety Board 2019, see Abstract). Although the probable cause was ruled to be "the failure of the vehicle operator to monitor the driving environment and the operation of the automated driving system because she was visually distracted throughout the trip by her personal cell phone", with a number of contributing factors (ibid, p. 59), the example, nevertheless, illustrates two failures of the automatic system. First, the system failed to identify the pedestrian correctly—shifting from vehicle, to other, to bicycle. Second, the system was unable to correctly predict the path of the pedestrian until it was too late—the pedestrian was considered *not* on path of the vehicle until 1.5 s before the accident (ibid, pp. 15–16).

[15] Here it may be illustrative to keep in mind that a valid argument is not necessarily sound.

[16] The concept *fact-relative* comes from Derek Parfit (2011). According to Parfit, acts are "*wrong* in the *fact-relative* sense just when this act would be wrong in the ordinary sense if we knew all of the morally relevant facts" (p. 150).

apply it perfectly.[17] Lastly, suppose that we have created such an AI. Under these suppositions, the AI would be able to *for each given description of a situation, correctly determine the morally right action to perform*. This raises the question of whether that resolves all ethical queries involving ethical machine decisions. Here is a generalizable illustration to show that it would not. Suppose that for any given situation the possible ethically relevant factors are A, B, and C. Suppose further that the machine can only input A and B. Suppose that two situations are ethically equivalent, but that they vary in relation to A and B. The machine will incorrectly determine that these situations are ethically non-equivalent, because the machine can only describe the situation using A and B. Furthermore, suppose that two situations are ethically non-equivalent because they differ in relation to factor C (although they are equivalent relative to A and B). The machine will incorrectly determine that these situations are ethically equivalent.[18]

To give an example from the literature, suppose that an autonomous vehicle is facing the bivalent choice of crashing into a kid or a grandmother. Suppose that the correct priority, given the contextual factors—such as the speed of the vehicle—is to avoid crashing into the grandmother. Moreover, suppose that in a counterfactual situation—such that the machine was facing the bivalent choice of crashing into a kid or an adult, *ceteris paribus*—the machine should avoid the kid over the adult. Lastly, suppose that the machine cannot distinguish between different forms of adults (i.e., it cannot separate adult grandmothers from other adults). If so, the machine mistakenly assumes that it is facing the choice in the counterfactual situation and although it makes the right decision, given the available information, it ends up making the wrong decision (all things considered).

This hopefully makes it clear that even a "perfect" decision-making machine (i.e., as defined earlier), would yield erroneous decisions if it is missing ethically relevant facts (*mutatis mutandis* for false information). One problem for the standard approach is that the possibility of false information and information-gaps cannot be dealt with in the same way as risks and other uncertainties. For example, Keeling suggested a method that arguably depends on known possibilities, but given that we are dealing with information gaps (unknown unknowns) and potentially false information, this problem arguably cannot be solved as Keeling suggests. Of course, one may modify Keeling's proposal to suggest that there are *other* rational decisional principles that apply in the case of unknown unknowns and false information. However, it is difficult to see how the choice of those principles will not involve ethical questions. For example, under which conditions should we accept the possibility that an autonomous vehicle fails to identify a pedestrian crossing the street? That seems to be an inherently normative question.

Third, the reason why I favor the uncertainty approach is also tied to what I previously called the 'science-fiction presumption'.[19] This is a name for various examples from the

[17] Whether our best normative theories can be applied in machine decisions is a substantial empirical question that is normatively relevant, but I will not analyze this question here.

[18] *Mutatis mutandis* for false information.

[19] Lundgren (2020a).

literature on the ethics of crashing in which these systems are presumed to have capabilities that are currently not available. This idealization is different from the idealization standardly used in the applied trolley problem, since it concerns an idealization of the features of the machine making the decision. For example, Derek Leben supposes "that it is possible for an autonomous vehicle to estimate the likelihood of survival for each person in each outcome".[20] The problem is that proposals of what a machine, with some specific capabilities, ought to do, tells us very little about what machines with other features ought to do. If the standard approach was correct, then one may argue that these examples could still be useful because they show us what the basic decision-making principles ought to be. However, decision-making principles based on science-fiction presumptions may be incongruent with the best available technical solutions. Furthermore, as I will argue in the upcoming section, technical choices must be normatively evaluated.

To summate: In this section I have argued that we must adhere to the uncertainty approach, because the ethics of machine decision-making is not reducible to what a machine ought to do in a given situation, but it also depends on the machines' uncertainty about the facts of the situation. Moreover, what the machine ought to do can only be determined in relation to what degree of factual uncertainty is acceptable in the given situation (e.g., how fast an autonomous vehicle should be allowed to drive depends, in part, on its ability to correctly identify and avoid objects in its proximity). Lastly, we need to consider what inputs are needed to achieve an acceptable level of certainty that a machine is making the right decision.

3 The grandma problem

Based on the previous section it should be clear that to achieve an acceptable level of certainty that a machine makes the right decisions we need to ensure that it has access to the relevant inputs. In this section, I will argue that adding inputs in most cases creates trade-offs. As previously noted these trade-offs do not need to be necessary; what is important is that the potential of trade-offs are so common that they must be part of the normative evaluation. The main trade-offs are between the inputs needed to achieve *an acceptable degree of certainty that the machine decisions are ethically correct* and *the risks of harms* from using these inputs. Moreover, I will also argue that more inputs may yield problems in situations with time constraints. To establish all of this, I will start by making use of an example from the discussion of ethics of crashing.

Suppose an ethical machine decision depends on whether someone is a grandmother.[21] Granted this presumption, a machine must be able to model the property of being a grandmother in order to achieve an ethical decision (i.e., without an evaluation of the relevant ethical factors, the machine will not be able to make the correct decision). The problem is that it is difficult to determine whether someone has

[20] Leben (2017, p. 110).

[21] Again, the example that the property of being a grandmother is important can be found in the literature (see, e.g., Awad et al., 2018; Hern, 2016; Lin, 2015, p. 70).

the property of being a grandmother.[22] A simple model could predict that *x is a grandmother* by first determining that *x is a human* (a feature that autonomous vehicles arguably need to have anyway), then using a simple image analyze that *x is a woman* and that *x is old*. Setting aside the uncertainty involved in these evaluations, it is clear what the problem with the proposed model is: The predictor is neither necessary, nor sufficient (i.e., there are young grandmothers and there are old women who are not grandmothers). Hence, if determining whether someone is a grandmother is necessary in order to make an ethical decision, we need a model with a more complex informational input. One idea is to equip the autonomous vehicle with facial recognition capability and access to an appropriate database.[23] Although such a model could be highly successful (granted the completeness of the database and the ability to perform timely, accurate, and precise facial recognition), it should be obvious that equipping autonomous vehicles with such technologies would be highly detrimental. It would not only be a privacy invasion for the individual, we would also enable an extreme mass surveillance system (making all vehicles moving parts of a joint visual surveillance system).[24]

This illustrates a trade-off between the ability to predict whether *x has the property y* and *risks of harms from information needed to predict that x has the property y*. On the one hand, we have a relatively (but not entirely) innocent image analysis that yields unsuccessful results. On the other hand, we have a potentially successful system with a really dangerous and detrimental integrated information analytic system. The question is if the trade-offs involved in this example apply more generally and if so to what extent?[25] In two upcoming subsections I will aim to establish that we have at least two

[22] One may wonder how this relates to the frame problem (i.e., "the challenge of representing the effects of action in logic without having to represent explicitly a large number of intuitively obvious non-effects", Shanahan, 2016). Indeed, the representational challenges of the frame problem are an explicit part of the grandma problem (cf. also fn. 25). Nevertheless, in order to avoid turning this already long article into a book I will not be able to address the particular literature on the frame problem in any detail. This restriction is partly motivated by the fact that some of the technical challenges addressed by the frame problem has been resolved (see, e.g., Shanahan, 2016 for references) and other issues relating to philosophy of mind are less of an interest. Moreover, there is also some conceptual debate about how the frame problem best should be understood (see, e.g., Miracchi, 2020), which means that any discussion thereof would necessarily be longer than what is motivated based on the aim of this article.

[23] One may worry that this example implicitly suffers from the science-fiction presumption. However, it is not a problem in this case since I am merely using this possibility to illustrate the problems with using such inputs (if possible), not to conclude how the machine ought to act.

[24] It is worth noting that there are alternative ways to model the property of being a grandmother. For example, a facial analysis can be used to predict the property of being a grandmother—given that there has been some success with determining other properties from people's faces (see, e.g., Kosinski, 2021). However, if such techniques would be possible, it would not solve the problems of trade-offs, which means that a normative analysis is needed and that is exactly the point I wish to make here.

[25] One may worry that I have sidestepped the option of solving this by machine learning. However, even if machine learning algorithms can solve problems in surprising ways, they do need access to some sort of information that actually allows the ability to predict the relevant property (in the example, the property of being a grandmother). Therefore, the basic problems apply to machine learning systems as well, even if the models are difficult to describe (something that I will return to in the next subsection). However, it is worth pointing out that machine learning involve further problems. For example, many AI systems are training using goals. The problem is that even goals that *prima facie* seem sensible can turn out to be problematic. An illustrative example of this is an AI that was designed to play Tetris suc-

common types of trade-offs between inputs that we may need for making the right decisions and the risks of using those inputs. As I have said before, the goal here is not to establish that these trade-offs apply to *all* decision-making machines. Arguably, it holds for many, if not most, machines that have a sufficiently broad application and capacity. I will not settle this distinction precisely, but it should be clear that there is a difference between an AI playing chess and an autonomous vehicle.

The grandma problem can also be used to illustrate a difficulty with time-sensitive decisions. Hence, in a third subsection I will turn to the trade-offs needed for right decision-making and the time needed to process these inputs.

Before turning to the subsections, it is worth pointing out that there is an overall trade-off for all these issues: cost and benefit. Arguably, adding an input often implies a cost (e.g., adding a sensor or processing the data), hence there is a trade-off between that cost and the benefit of adding that input. Although such cost–benefit analyses often are performed according to various methodological rules, any such analyses are substantially normative in nature, and cost–benefit analyses are not without problems.[26] This first trade-off is quite simple, but I mention it since it enters all input-selection choices.

Lastly, to summate: What the grandma problem showed was that inputs are needed to reduce decisional uncertainty. Specifically, inputs are necessary for the machine to know (with sufficient certainty) what is going on. Because it is difficult to a priori or *ex ante* determine what might be relevant facts in any possible situation and because there is factual uncertainty about which facts may matter for instrumental reasons or serve as a model for some instrumental or intrinsic value, we face a problem of adding inputs broadly (because we have reasons to believe that they may be of relevance) and the potential trade-off of adding these inputs. In the upcoming subsections I will deal with trade-offs of adding inputs, as well as the problem of time-sensitive decisions (which may be considered a trade-off in its own right). In the next section, I will make a brief sketch of how ethics of machine decision-making ought to proceed in light of the input-selection problem.

3.1 Transparency

I take for granted what I have argued previously, that is, that inputs are needed. In this subsection, I will deal with one negative aspect of adding inputs: how it affects the transparency of the system and why that matters. Simply put I will argue that adding inputs—*ceteris paribus*—increases the complexity and sophistication for many AI systems (such as artificial neural networks), which in turn would decrease

Footnote 25 (continued)
cessfully. The AI was given the goal to not loose (Tetris has no winning conditions, but goes on as long as the player manages to not loose). The AI ended up solving this by pausing the game (see Murphy, 2013). This illustrates how an AI can fulfill a goal in a way that is contrary to what we want it to achieve, because the goal logically does not include contextual premises that humans normally understand to be part of the goal condition. Defining such goal conditions is not generally a trivial task (cf. fn. 22).

[26] For a critical overview, see Hansson (2007).

the transparency of the system.[27] Hence, this would create a trade-off between inputs and transparency. In fact, some hold that the trade-off is a trilemma, between transparency, accuracy, and robustness.[28] Moreover, I will give a few examples demonstrating why a lack of transparency may be a problem.

Generally speaking a model can be opaque (i.e., non-transparent) or uninterpretable for two reasons: the internals of the model are unknown or we cannot assign any meaning or understandable explanation to the internals.[29] The problem of understanding the systems' internals arguably has to do with its complexity. That is, while complexity is defined in different ways—relative to different techniques—it is standardly viewed as an opposed term of interpretability.[30] Some "define the model complexity as the model's size",[31] which indicates that increasing the inputs increases the complexity and decreases interpretability (or transparency).

Understanding it in this rough and simplified way we get that transparency is decreased when model size increases. Given that adding inputs increases the size of the model, adding inputs, *ceteris paribus*, generally decreases model transparency. To see this more clearly, it might be illustrative to consider that different authors have defined complexity in terms of the number of regions, non-zero weights, the depth of the decision tree, or "the length of the rule condition".[32] So, for example, increased inputs would *prima facie* add to the length of the rule condition by adding criteria that must be considered in the rule condition (likewise for the other definitions). Thus, as a rule of thumb, adding inputs decreases transparency (i.e., there is a trade-off between inputs and transparency).[33]

Transparency is broadly promoted in the literature on ethical AI.[34] Ethically, we can distinguish between two different transparency demands, which we may desire for various reasons. On the one hand, we may demand that the system satisfy a demand of *explainability* (i.e., that the machine decision, or the justification thereof, is understandable). On the other hand, we may demand that the system satisfy a demand of *traceability* (i.e., that we have the ability to trace the decision from input to output).[35]

Before explaining why these demands matter, it should be recognized that explainability, in all fairness, does not link directly to model complexity, since what we need to understand is not necessarily the model, but the result of the model. Yet, the argument between increased inputs, complexity, and transparency, arguably holds as a rule-of-thumb, which

[27] See, e.g., Guidotti et al. (2018) and Noga (2018).

[28] Thieltges et al. (2016).

[29] Guidotti et al. (2018, p. 5).

[30] Ibid., p. 9.

[31] Ibid., p. 6.

[32] Ibid., p. 9.

[33] I need not establish that the relation is necessary. It is sufficient that it is common and problematic and hence deserves consideration.

[34] Transparency is broadly considered an important property (see, e.g., AI HLEG, 2019; Brey et al., 2019; Danaher, 2016; Floridi et al., 2018; Wachter et al., 2017; Walmsley, 2020; cf. also Zerilli et al., 2018 for a more critical perspective and more references).

[35] See, e.g., Brey et al. (2019).

is sufficient. (Keep in mind that the point is to establish that these are concerns that deserve our attention when evaluating the ethics of machines' decision-making.)

Explainability is important in legal, political, and medical contexts, for example. In a legal context, we usually want to avoid procedural opacity, because you have a right to understand and (in many cases, if it applies to you) appeal a legal decision (and in order to so, one must understand the decision). Moreover, a legal decision is often strongly connected to the legal reasoning that it is based upon.[36]

Understanding political decisions is also important, at least in a democracy. It is also important for political participation, since if you do not understand the political process or political decision-making, then participation will be difficult. Hence, political usage of algorithmic decision-making may make political participation more difficult.[37]

In the medical context, informed consents are considered a gold standard for medical decision-making (e.g., because they protect individuals against harms and abuse; protect their autonomy, self-ownership, and personal integrity; and increase trust and decrease domination),[38] and decisional opacity is a problem for informed consent since it makes it difficult to inform the individual of the reasons for her treatment. Even if we hold that decisional accuracy is more important than explainability, that does not mean that there is no trade-off.[39]

Explainability can also increase trust,[40] which may be important to alleviate fear of new technologies, such as the fear of riding a fully autonomous vehicle.[41]

Traceability is important for responsibility and accountability (e.g., when something has gone wrong or to increase trust in a system by allowing it to be monitored and evaluated), and to increase safety and reliability (e.g., in the case of an autonomous vehicle crashing we might not only need to determine who is responsible or accountable, but also how we can improve the system). Our ability to fully understand the system can also be important to reveal bias in algorithms.

These are just a few examples to illustrate the importance of transparency and that it must be part of the normative evaluation of ethical machine decisions.

In summation, this subsection has shown that we need to consider a possible trade-off between transparency and adding inputs. As previously noted, the point here is

[36] See, e.g., Walmsley (2020) for a discussion of how contestability can be decoupled from transparency.

[37] See Danaher (2016) for an argument of the risks involved in using algorithms in political decision-making.

[38] See, e.g., Eyal (2019) for an overview.

[39] London (2019) goes further and argues that "a blanket requirement that machine learning systems in medicine be explainable or interpretable is unfounded and potentially harmful" (p. 15). However, London's point is that "Recommendation to prioritize explainability or interpretability over predictive and diagnostic accuracy are unwarranted in domains where our knowledge of underlying causal systems is lacking. Such recommendations can result in harms to patients whose diseases go undiagnosed or who are exposed to unnecessary additional testing" (p. 20). However, my point here is not that there should be a blanket requirement, but that there is a trade-off that requires a normative analysis. Moreover, in cases where the doctors strongly disagree with the findings of a machine's analyses, some kind of explainability may be strongly warranted.

[40] See, e.g., Herlocker et al. (2000).

[41] A telephone survey with 1008 completed interviews of adult Americans (18-year-olds or older) shows that 71% of all Americans "are afraid to ride in fully self-driving vehicles," which has increased over time from 63% (Edmonds 2019).

not that transparency necessarily matters in all situations, nor that adding inputs necessarily reduces transparency in a relevant way in any situation. The point is, as I just stated, that it needs to be part of the normative analyses of machine decisions.

3.2 Privacy and data protection

The grandma problem quite clearly illustrated a trade-off between privacy and data protection and the adding of inputs to decrease decisional uncertainty. While privacy has been recognized as a substantial problem for autonomous vehicles,[42] it ought to be clear that issues relating to privacy and data protection apply much more broadly to most kinds of machine decisions.

However, one may think—based on the grandma problem—that privacy is only at risk when we are dealing with sensitive data. Thus, one may worry that the trade-off is restricted to situations in which one is dealing with sensitive information. For this reason, in this subsection I will aim to show that we have very broad reasons to minimize usage of data and information. Simply put, I will show that the idea of restricting machines to non-sensitive information inputs does not guarantee a protection of sensitive information. Moreover, I will exemplify why we should be concerned about an individual's privacy, besides a right to privacy or an individual desire for secret keeping.

There are various examples of how seemingly innocent data-sets can be used to predict fairly sensitive information. For example, it has been shown how 'Likes' on Facebook (i.e., giving a virtual thumbs-up to a social media posting) can be used to predict personal information such as political leaning (Republican/Democrat), sexuality, parental separation before 21 years old, etcetera.[43] Once we predicate more substantial and/or sensitive information there is a risk that such information could be used to manipulate individuals and blackmail them.[44] Manipulation based on information harvesting is arguably a business model used by many so-called "free" online services.

As this illustrates, there are further reasons for data protection beyond privacy concerns. For example, Jeroen van den Hoven argues that there are at least three reasons for data protection, beyond privacy: information-based harm, informational inequality, and informational injustice. Information-based harm is harm to an individual that makes use of personal information; informational inequality is concerned with (a lack of) transparency and fairness in the informational marketplace (i.e., access to information is power); and informational injustice is concerned with how

[42] See, e.g., Borenstein et al. (2019), Glancy (2012), Hevelke and Nida-Rümelin (2015), Himmelreich (2018), Holstein et al. (2018), Lin (2013, 2015), McBride (2016), Mladenovic and McPherson (2016), Nyholm (2018), Ryan (2019), Santoni de Sio (2017), Stone et al. (2020), and Wolkenstein (2018). Most of these references merely note that that privacy is important; the only substantial discussion is in Glancy (2012). See also Vrščaj et al. (2020) for two empirical studies of attitudes on autonomous vehicles (including privacy) and some more references; and Zimmer (2005) for a discussion on privacy and vehicle-to-vehicle communication.

[43] Kosinski et al. (2013). See also Ohm (2010) and Lundgren (2020b) for several other examples.

[44] Ohm (2010) and Lundgren (2020b).

information is used to discriminate against an individual.[45] As you can see, some of these reasons overlap with transparency considerations.

There are also surveillance problems. For example, autonomous vehicles must have the ability to track both the passenger(s) and people in its vicinity. The surveillance capability and tracking of people in its vicinity is arguably an extremely substantial problem, since it also affects the privacy of non-users (meaning that they are negatively affected without receiving the benefits and without being able to properly opt out, which poses a problem for solving this by using informed consents). If combined, these surveillance capabilities can also be used for undemocratic purposes, to control the population.

This can put individuals in a problematic situation where they have to choose between using AI services and protecting their privacy. For example, Dorothy J. Glancy discusses an example involving an autonomous vehicle in which you must choose between increasing mobility (which increases user autonomy) or giving up your informational privacy (because the service requires access to, e.g., travel data).[46]

In summation, this subsection has shown that we need to consider a possible trade-off between, on the one hand, privacy and other informational wrongdoings and, on the other hand, adding inputs. As previously noted, the point here is not that all situations of adding inputs will affect an individual's privacy or cause informational wrongdoings. Nevertheless, it is clear that adding inputs cannot only affect an individual's privacy because the data is sensitive; even insensitive data can be privacy-problematic. Moreover, information can be used and abused in various ways and that gives us reason to consider limits on information access (as well as creation, for that matter). Furthermore, AI systems can, when combined, also lead to a risk of mass surveillance. Thus, the choice of inputs must be evaluated against various privacy concerns and reasons for data protection, whether directly or indirectly, and more broadly against risks of mass surveillance. The overall point is, as I just stated, that these trade-offs need to be part of the normative analyses of machine decisions.

3.3 Time-sensitive decisions

Adding inputs not only adds a monetary cost, decreases transparency, or affects an individual's privacy, it also yields an increased decision time (because the input must be processed). That is, adding inputs postpones the machine's decision-making, *ceteris paribus*. The problem with postponing decisions in general is that it can—all things considered—lead, to more harm, due to delayed response time.

In this subsection, I will argue that this implies an ethical problem, in the design process, between adding a function that allows for a more highly grained ethical analysis versus making a decision *in time*. The problem is that we *can* end up with a

[45] Van den Hoven (1999).

[46] Glancy (2012, p. 1186).

machine that although it makes "better decisions" (if allowed to run through the full process), ends up performing worse, because making better decisions takes more time. One may suppose that this is a matter of optimization, but that depends on knowing beforehand the trade-offs between time and best analysis, decisions, and action, in *any* given situation. That it, for systems that will be used in varied contexts with varied complexity there will also be a variation in decision time. Although this is partly an engineering problem, it is not only an engineering problem. It includes the normative choice on how much decisional uncertainty we can accept relative to how quick decisions can be made in a given situation.

It is easy to see how this conflict may create a situation in which we end up with a suboptimal decision. I will establish this for both absolutist rule-based ethics and consequence-based ethics. In situations of uncertainty, absolutist decision rules (e.g., a constraint) are usually applied as follows: "it is permissible to Φ only if" the probability that Φing will breach the constraint "is lower than some threshold".[47] Given that more inputs add processing time, this means that there would be some set(s) of *input data* and some time constraint(s) for a given machine-choice mechanism such that the machine would miscalculate the threshold at the time- limit because of the added processing time from added inputs, while a smaller sets of inputs would yield the correct decision (even if the estimate of the probability that Φing breaches the constraint would be more imprecise if the algorithm were allowed to run without time constraints). That is, more is sometimes less (or worse).

Similarly, for a consequence-based ethics, the evaluation of the utility of the two actions can take more time because of added inputs, which may also skew the balance of the choice in a way that during a specific time limit gives the incorrect result in accordance with the theory applied, while an alternative with fewer inputs gives the correct outcome.

In this example, it is clear that the problem is that some inputs were not needed, for otherwise, fewer inputs could not possibly generate the correct decision. However, the fact that some inputs were not needed to reach a correct decision in this situation does not imply that they would not be necessary in another situation (e.g., if more precision would be needed and the time constraints would differ). Thus, this example illustrates that the selection of inputs is a substantial normative choice that we must engage with.

[47] More formally: "it is permissible to Φ only if $p(\neg C)$ is less than some threshold," where C is "your reason to abide by a constraint, and $p(\neg C)$ is the probability that Φing will breach the constraint" (Lazar & Lee-Stronach, 2017, p. 6). However, we might be critical of the underlining presumptions of such standard procedures of adaptation for various reasons (as Lazar and Lee-Stronach are). What matters in this context is that we might think that such a threshold should be contextually sensitive rather than the same in all situations (ibid., p. 7). Moreover, the situation arguably becomes a bit more complex, for absolutistic rules can allow for less strict decision-making when uncertainties are involved (ibid.). However, the argument I present above can *mutatis mutandis* be modified to address such considerations.

4 So what should we do?

Given that the current practices of applied ethics have largely ignored the role factual uncertainty plays in what I have called the input-selection problem, what should ethicists do differently henceforth? First, ethicists need to take inputs and input limitations into consideration when analyzing what constitutes ethical machine decisions. Although this may seem obvious, this is currently largely neglected (at least in the applied ethical literature on autonomous vehicles). Second, taking inputs and input limitations into consideration requires an analysis of the trade-off between the benefits for ethical machine decisions by including various inputs $X_1,\ldots,X_n$ and the potential negative effects (e.g., for privacy) of including such inputs. Thus, the discussion of ethical machine decisions needs to change radically to take these potential trade-offs into consideration.

To illustrate the above point, consider the examples of unavoidable crash scenarios in which we must choose whether a vehicle should crash into A or B. Suppose, for example, that we think that in situations with a choice between accidentally killing individual A or B, the ethical choice depends on A's and B's individual properties. (As before, I am merely using this as an example, with all caveats of simplification.)

If we think that the individual's properties matter, then we can attempt to put a value on this. For example, if A has property x and B has property y, then A should be prioritized over B. That is how it is commonly discussed in the literature, with the caveat of adding conditions for varying degrees of uncertainty. However, I have argued that these analyses are incomplete because we must also consider *how* the machine can conclude that A has the property x (with some sufficient degree of probability, whatever that is), because we need to evaluate the risks of potential harms from using that kind of machine and selecting the inputs needed for that machine.

There are different ways of doing this. One way would be to attempt to spell out all conditions that apply (all decisions about how to handle utilities and/or non-consequentialist values, uncertainties, and trade-offs). Alternatively, we can consider available alternatives and see if anyone of them is permissible and/or obligatory.

For example, considering the grandma problem we may conclude that using facial recognition technologies implies too many risks of serious harm. Here we ought to consider not only risks involved if the machine is used as intended, but also risks involved with abuses and accidental misuse. Therefore, when evaluating potential trade-offs, we might conclude that the downsides of adding these inputs dominate, all-things-considered, the benefits of adding them. If so, we can consider other options, with a, *ceteris paribus*, lower degree of certainty in the evaluation. If any option is *prima facie* permissible in its own right (i.e., the benefits of the machine decisions seem to outweigh its risks), then we set that option aside so that it can be considered against other *prima facie* permissible options.

In any case, we have to repeat the process for a representative sample of alternatives. With all alternatives, we must evaluate the trade-off against the different

degrees of (un)certainty about the ethical machine decision (what that is may in itself be uncertain). For example, how important is it—from an ethical perspective—to be able to say that the probability that A has property x is between 0.99 and 1 or that it is between 0.8 and 1? Furthermore, how would any new inputs and functionality affect the machines' decision-making capabilities in time-sensitive situations?

Simplified, what I suggest is that there are five steps that I believe need to be part of the ethical evaluations of machine decision-making henceforth. Note that we might have to go back-and-forth through the steps. First, we start with a normative investigation of the basic goals of machine decisions, which is already considered in the literature. Second, we need to find out what technical options there are for achieving that goal, which is something often ignored in the literature, sometimes—as previously noted—in favor of science-fiction presumptions. Third, for each option we need to identify the potential ethical trade-off between achieving the goal and the normative cost of doing so (e.g., how does the proposed functionality of the system affect its transparency or individuals' privacy). This is the core of what I argued for in the previous section. Fourth, with the trade-off in mind we need to evaluate the normative value of accuracy and robustness of machine decisions relative to the potential trade-offs. That is, what degree of certainty in achieving our goals justifies the associated risks? Are there alternatives that better protect data and achieve a higher degree of transparency? Fifth, for time-sensitive decision-making we need to evaluate all the previous considerations relative to the risks involved in time-limited decision-making. This fifth step is perhaps best considered not a separate step, but as part of steps 3 and 4.

Currently, the ethical analysis of machine decisions focuses only on the first step. It ignores technical limitations; it ignores the potential trade-offs of having certain machine decision-making capabilities; proponents for the standard approach also largely ignore the normative evaluation needed for the value and disvalue of certainty and uncertainty; and it does not address the particulars of time-sensitive decision-making.

5 Summation and conclusions

In conclusion, to analyze the ethics of machine decisions we need to consider how much decisional uncertainty we can accept. Achieving an ethically *perfect* decision requires not only that the machine has a well-calibrated decisional algorithm, but also that it has access to all the ethically relevant facts (i.e., we need to consider which inputs are needed for decision-making). Thus, to achieve an acceptable level of decisional uncertainty, a central question is what inputs the machine needs access to. The problem is that it is *prima facie* difficult to know precisely which inputs are ethically relevant. More importantly, as I have argued in this article adding inputs in most cases implies a trade-off (e.g., adding inputs puts various important values such as transparency and privacy at risk). These trade-offs are not necessary, but they are so common that one must evaluate them when one evaluates what constitutes an ethical machine decision. Moreover, for decisions under time constraints fewer inputs

might yield a better output because of added processing time. All these conclusions imply a revision of the way that the ethics of machine decisions are currently being discussed: We need to take the machine and possible ways that the machine could be constituted into consideration, we have to consider potential trade-offs, and we have to pay particular attention to decisions under time constraints. Hence, we cannot think about machine ethics in isolated idealized terms, we need to analyze it in context and analyze the associated uncertainty and the possibly related trade-offs with reducing the uncertainty to an acceptable level—and that is the input-selection problem.

Lastly, it should be mentioned that while I have focused on two trade-offs in this text, there are arguably other trade-offs that deserve attention in the normative analysis. For example, how should we deal with a data-set that is highly accurate but biased? That is, this article should not be read as an indication that transparency, privacy, cost–benefit, and time constraints are the only problems that need to be addressed.

Acknowledgements In preparing this manuscript I have benefited from comments received at two conferences: *International Association for Computing and Philosophy* (Warsaw, 2018) and *Computer Ethics–Philosophical Enquiry* (Norfolk, 2019). I have also benefited from comments at seminars at the *Institute for Futures Studies* and the *Royal Institute of Technology*. I am grateful for the comments I received on these occasions. I am also grateful for comments from Kalle Grill and two anonymous reviewers for *Synthese*.

Funding Open access funding provided by Umea University. This work has been supported by a grant from the *Swedish Transport Administration* and it was also partially supported by the *Wallenberg AI, Autonomous Systems and Software Program—Humanities and Society* (WASP-HS) funded by the *Marianne and Marcus Wallenberg Foundation* and the *Marcus and Amalia Wallenberg Foundation*.

References

AI HLEG (High-Level Expert Group on Artificial Intelligence). (2019). Ethics guidelines for trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Sharif, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Borenstein, J., Herkert, J. R., & Miller, K. W. (2019). AVs and engineering ethics: The need for a system level analysis. *Science & Engineering Ethics, 25*(2), 383–398. https://doi.org/10.1007/s11948-017-0006-0

Bradley, R., & Drechsler, M. (2014). Types of uncertainty. *Erkenntnis, 79*(6), 1225–1248. https://doi.org/10.1007/s10670-013-9518-4

Brey, P., Lundgren, B., Macnish, K., & Ryan, M. (2019). Guidelines for the development and the use of SIS. Deliverable D3.2 of the SHERPA project. https://doi.org/10.21253/DMU.11316833.v3.

Bykvist, K. (2017). Moral uncertainty. *Philosophy Compass, 12*(3), e12408. https://doi.org/10.1111/phc3.12408

Casey, B. (2017). Amoral machines, or: How robotics can learn to stop worrying and love the law. *Northwestern University Law Review, 112*(5), 1–20. https://doi.org/10.2139/ssrn.2923040

Cunneen, M., Mullins, M., Murphy, F., & Gaines, S. (2018). Artificial driving intelligence and moral agency: Examining the decision ontology of unavoidable road traffic accidents through the prism of the Trolley Dilemma. *Applied Artificial Intelligence, 33*(3), 267–293. https://doi.org/10.1080/08839514.2018.1560124

Danaher, J. (2016). The threat of Algocracy: Reality, resistance and accommodation. *Philosophy and Technology, 29*, 245–268. https://doi.org/10.1007/s13347-015-0211-1

Edmonds, E. (2019). Three in Four Americans Remain Afraid of Fully Self-Driving Vehicles. Retrieved from: https://newsroom.aaa.com/2019/03/americans-fear-self-driving-cars-survey/

Eyal, N. (2019). Informed Consent. *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/informed-consent/.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review, 5*, 5–15.

Glancy, D. J. (2012). Privacy in autonomous vehicles. *Santa Clara Law Review* 52(4):1171–1239. https://digitalcommons.law.scu.edu/lawreview/vol52/iss4/3.

Goodall, N. J. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2424*, 58–65. https://doi.org/10.3141/2424-07

Goodall, N. J. (2016). Away from Trolley problems and toward risk management. *Applied Artificial Intelligence, 30*(8), 810–821. https://doi.org/10.1080/08839514.2016.1229922

Greaves, H. (2016). Cluelessness. *Proceedings of the Aristotelian Society CXV, I*(3), 311–339. https://doi.org/10.1093/arisoc/aow018

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 51(5):Article 93. https://doi.org/10.1145/3236009.

Hansson, S. O. (2003). Ethical criteria of risk acceptance. *Erkenntnis, 59*(3), 291–309. https://doi.org/10.1023/A:1026005915919

Hansson, S. O. (2007). Philosophical problems in cost-benefit analysis. *Economics & Philosophy, 23*(2), 163–183. https://doi.org/10.1017/S0266267107001356

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (CSCW '00). Association for Computing Machinery, New York, NY, USA, pp. 241–250. https://doi.org/10.1145/358916.358995.

Hern, A. (2016). AVs don't care about your moral dilemmas. *The Guardian*. Retrieved from: https://www.theguardian.com/technology/2016/aug/22/self-driving-cars-moral-dilemmas.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics, 21*(3), 619–630. https://doi.org/10.1007/s11948-014-9565-5

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice, 21*, 669–684. https://doi.org/10.1007/s10677-018-9896-4

Holstein T, Dodig-Crnkovic G, & Pelliccione P (2018) Ethical and social aspects of self-driving cars. arXiv preprint. https://arxiv.org/pdf/1802.04103.pdf.

JafariNaimi, N. (2018). Our bodies in the Trolley's path, or why self-driving cars must *not* be programmed to kill. *Science, Technology, & Human Values, 43*(2), 302–323. https://doi.org/10.1177/0162243917718942

Keeling, G. (2020a). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics, 26*, 293–307. https://doi.org/10.1007/s11948-019-00096-1

Keeling, G. (2020b). The ethics of automated vehicles. PhD thesis. https://doi.org/10.13140/RG.2.2.28316.10889.

Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Report*. https://doi.org/10.1038/s41598-020-79310-1

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS, 110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Lazar, S., & Lee-Stronach, C. (2017). Axiological absolutism and risk. *Noûs, 53*(1), 97–113. https://doi.org/10.1111/nous.12210

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology, 19*, 107–115. https://doi.org/10.1007/s10676-017-9419-3

Lin, P. (2013). The Ethics of Autonomous Cars. *The Atlantic* October 8. Retrieved from: https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

Lin, P. (2015). Why Ethics Matters for Autonomous Cars. In Maurer M., Gerdes J., Lenz B., Winner H. (eds) *Autonomes Fahren*. Springer Vieweg, Berlin. Also available in a English-titled 2016-edition (*Autonomous Driving*).

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report, 49*(1), 15–21. https://doi.org/10.1002/hast.973

Lundgren, B. (2020). Safety requirements vs. crashing ethically: What matters most for policies on autonomous vehicles. *AI & Society*. https://doi.org/10.1007/s00146-020-00964-6.

Lundgren, B. (2020b). Beyond the concept of anonymity: what is really at stake? In: Macnish, K., and Galliot, J. (eds.) *Big data and democracy*. Edinburgh University Press, Edinburgh, pp. 201–216. https://edinburghuniversitypress.com/pub/media/resources/9781474463522_Chapter_13.pdf.

MacAskill, W., Bykvist, K., & Ord, T. (2020). Moral uncertainty. *Oxford University Press*. https://doi.org/10.1093/oso/9780198722274.001.0001 (N.B., the book is open access).

MacAskill, W., & Ord, T. (2020). Why maximize expected choice-worthiness? *Noûs, 54*(2), 327–353. https://doi.org/10.1111/nous.12264

McBride, N. (2016). The ethics of driverless cars. *SIGCAS Computer Society, 45*(3), 179–184. https://doi.org/10.1145/2874239.2874265

Miracchi, L. (2020). Updating the frame problem for AI research. *Journal of Artificial Intelligence and Consciousness, 7*(2), 217–230. https://doi.org/10.1142/S2705078520500113

Mirnig A.G. & Meschtscherjakov A. (2019). Trolled by the trolley problem: on what matters for ethical decision making in automated vehicles. *CHI '19: Proceedings of the 2019 CHI conference on human factors in computing systems* Paper No 509: 1–10. https://doi.org/10.1145/3290605.3300739.

Mladenovic, M. N., & McPherson, T. (2016). Engineering social justice into traffic control for self-driving vehicles? *Science and Engineering Ethics, 22*(4), 1131–1149. https://doi.org/10.1007/s11948-015-9690-9

Murphy, T. (2013). The first level of super mario bros. is easy with lexicographic orderings and time travel . . . after that it gets a little tricky. Retrieved from: http://www.cs.cmu.edu/~tom7/mario/mario.pdf.

National Transport Safety Board. (2019). Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018. Highway Accident Report NTSB/HAR-19/03. Washington, DC. Retrieved from: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf.

Noga, M. (2018). Bringing transparency Into AI. *Digitalist Maganize.* 27 November. Retrieved from: https://www.digitalistmag.com/future-of-work/2018/11/27/bringing-transparency-into-ai-06194523.

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

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for AVs: An applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289. https://doi.org/10.1007/s10677-016-9745-2

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57:1701–1777. https://www.uclalawreview.org/pdf/57-6-3.pdf.

Parfit, D. (2011). *On what matters* (Vol. 1). Oxford University Press.

Ryan, M. (2019). The future of transportation: Ethical, legal, social and economic impacts of self-driving vehicles in the year 2025. *Science and Engineering Ethics, in Press.* https://doi.org/10.1007/s11948-019-00130-2

Santoni de Sio, F. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice, 20*, 411–429. https://doi.org/10.1007/s10677-017-9780-7

Shanahan, M. (2016). The frame problem. *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition). https://plato.stanford.edu/archives/spr2016/entries/frame-problem/.

Stone, T., Santoni de Sio, F., & Vermaas, P. E. (2020). Driving in the dark: Designing autonomous vehicles for reducing light pollution. *Science and Engineering Ethics, 26*, 387–403. https://doi.org/10.1007/s11948-019-00101-7

Thieltges, A., Schmidt, F., & Hegelich, S. (2016). The Devil's triangle: Ethical considerations on developing bot detection methods. *2016 AAAI Spring Symposium Series.* https://www.aaai.org/ocs/index.php/SSS/SSS16/paper/view/12696.

Thompson, J. J. (1985). The trolley problem. *Yale Law Journal, 94*(5), 1395–1515. https://doi.org/10.2307/796133

van den Hoven, J. (1999). Privacy and the varieties of informational wrongdoing. *Australian Journal of Professional Applied Ethics, 1*(1), 30–43.

Vrščaj, D., Nyholm, S., & Verbong, G. P. J. (2020). Is tomorrow's car appealing today? Ethical issues and user attitudes beyond automation. *AI & Society, 35*(4), 1033–1046. https://doi.org/10.1007/s00146-020-00941-z

Wachter, S., Mittelstadt, B. D., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics, 2*(6), eaan6080. https://doi.org/10.1126/scirobotics.aan6080

Walmsley, J. (2020). Artificial intelligence and the value of transparency. *AI & Society*. https://doi.org/10.1007/s00146-020-01066-z

Wolkenstein, A. (2018). What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of AVs. *Ethics and Information Technology, 20*(3), 163–173. https://doi.org/10.1007/s10676-018-9456-6

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*, 661–668. https://doi.org/10.1007/s13347-018-0330-6

Zimmer, M. (2005). Surveillance, privacy and the ethics of vehicle safety communication technologies. *Ethics and Information Technology, 7*(4), 201–210. https://doi.org/10.1007/s10676-006-0016-0