On the distinction between implicit and explicit ethical agency

Sjur Dyrkolbotn
Høgskulen på Vestlandet, Norway
sdy@hvl.no

Truls Pedersen
University of Bergen, Norway
truls.pedersen@uib.no

Marija Slavkovik
University of Bergen, Norway
marija.slavkovik@uib.no

Abstract

With recent advances in artificial intelligence and the rapidly increasing importance of autonomous intelligent systems in society, it is becoming clear that artificial agents will have to be designed to comply with complex ethical standards. As we work to develop moral machines, we also push the boundaries of existing legal categories. However, the most pressing question is not whether an artificial agent can be a moral agent or a legal person, but what kind of ethical decision-making our machines are able to engage in. Both in law and in ethics, the concept of agency forms a basis for further legal and ethical categorisations. Hence, without a cross-disciplinary understanding of what we mean by ethical agency in machines, the question of responsibility and liability cannot be clearly addressed. Here we make first steps towards a comprehensive definition, by formalising ways to distinguish between implicit and explicit forms of ethical agency.

Introduction

One of the goals of machine ethics is to develop machines that behave ethically. Hence, the concept of *ethical agency* is of great importance to the field. But what is the intended interpretation of this concept, and what counts as evidence of ethical agency in machines? The answer depends on whether we are dealing with *implicit* or *explicit* forms of ethical agency. Roughly, this is the distinction between machines that behave ethically by design and machines that are designed to reason ethically about (their own) behaviour. In the following, we aim to clarify the distinction further by offering a simple mathematical model that captures what we believe to be the essential difference.[1]

It should be noted at the outset that we do not take machine ethics to be about existing philosophical theories of ethics and how to implement them in machines. Rather, we are trying to develop machines that are able to live up to our *expectations* of ethical behaviour, also when operating autonomously. Hence, our focus is on how expectations can be fulfilled, not how they can be philosophically justified.

Ethical agency is a disputed concept when applied to artificial agents (Etzioni and Etzioni, 2017). Within the scope of this paper we understand it to be a range of features that produce ethical behaviour. The question that then needs to be addressed is *how* ethical behaviour is to be produced. When considering this question, (Wallach and Allen, 2008, Chapter 2) sketch a path for using current technology to develop artificial moral agents. They use the concept "sensitivity to values" to avoid the philosophical challenge of defining precisely what counts as agency and what counts as an ethical theory. Furthermore, they recognise a range of ethical "abilities" starting with *operational morality* at one end of the spectrum, going via *functional morality* to *responsible moral agency* at the other. They argue that the development of an artificial moral agents requires coordinated development of autonomy and sensitivity to values. Here we take this idea further by proposing to classify agents in terms of how their autonomy and their ethical competency is coordinated.

We will follow Moor (2006) in saying that an agent fulfils an expectation implicitly if it lacks the ability to violate it. Hence, implicit ethical agents are characterised by negative facts; to demonstrate compliance, we must show that the agent fails to have certain capabilities, or that it refrains from making certain kinds of decisions. Explicit ethical agents, by contrast, have the ability to autonomously evaluate the normative status of actions and reason independently about when they count as unethical. Such agents might be able to solve normative conflicts. Furthermore, they could sometimes *violate* certain rules, resulting in better fulfilment of overarching ethical objectives (Bench-Capon and Modgil, 2017). Consider, for instance, an autonomous vehicle that finds itself in a situation where it can break a rule of traffic to avoid a serious accident. We seem entitled to expect that an ethical car will do the right thing in theses cases. But this expectation might be exceedingly difficult to fulfil unless we implement some form of explicit ethical agency, allowing the car to autonomously decide when a traffic violation is in order.

The remainder of the paper will develop the distinction between implicit and explicit agency further, culminating in a simple mathematical formulation of what we take to be the key difference. Before getting into the technical details, we

[1] An alternative terminology would be to speak of *moral* agency, as in the term "moral machines". However, since many philosophers regard morality as a reflection of moral personhood, we prefer to speak of "ethical" agency here, to stress that we are referring to a special kind of rule-guided behaviour, not the (distant) prospect of full moral personhood for machines.

will comment on what we believe to be the legal implications of our work.

Some legal implications

If an agent violates an expectation that it is supposed to satisfy implicitly, this is evidence of a *defect* (Howells and Owen, 2010). The toaster that electrocutes a person when the "on" button is pressed has not attempted murder, it has malfunctioned. This remains the same regardless of how clever the toasters of the future might be when it comes to making individualised recommendations to the user in dietary matters. Regardless of the toaster's level of autonomy in other respects, standard products liability rules can be used to assign legal responsibility when it electrocutes someone.[2] Indeed, the ethical and legal issues that arise here are not fundamentally different from the ethical and legal issues raised by other complex technologies, autonomous or otherwise. This, we believe, holds true generally. Implicitly ethical behaviour, characterised by negative facts about the agent, does not make demands on the autonomous decision-making of that agent. This is the defining feature of implicitly ethical agency; the autonomy of the agent plays no active role in ensuring compliance with the norms.

By contrast, explicitly ethical agents are characterised by the fact that they rely on autonomous decision-making to fulfil ethical expectations. If a machine of this kind fails to behave ethically in a specific situation we can no longer conclude that the machine is defective. Compliance is no longer a negative property of the agent. Rather, compliance must now be defined positively in terms of *how* the agent makes autonomous decisions. If these decisions are completely predictable and controlled (in principle), there is no explicit ethical agency. What is missing in this case is not ethical competence, but autonomy *with respect to* ethical constraints, which is necessary in order for rule-following behaviour to count as autonomous agency.

Many norms are too context-sensitive, vague, and underspecified to be of much use unless they are directed at autonomous beings who can rely on their own reasoning when applying them. Asimov's laws (Asimov, 1950) illustrate the point: the most important rules are usually open to interpretation. More generally, it seems that the primary role of norms in a complex society is to encourage a certain type of reasoning about what is permitted and what ought to be done in different situations. Ethical and legal norms do not provide clear and unambiguous answers. In important matters, rules take the form of open-ended directives about how agents should exercise their freedom to choose.

In machine ethics, we could try to avoid using incomplete and ambiguous rules to regulate machine behaviour. If we succeed, all our machines will be implicit ethical agents. The fundamental legal challenge will then be avoided; machines can still be regarded as products and the legal personhood debate is rendered moot (Dyrkolbotn, 2017). However, to implement this strategy will require drastic regulatory measures. Machine learning, for instance, would have to be tightly regulated in order to prevent unpredictable learning algorithms from influencing machine decisions that pertain to our ethical expectations. More generally, the complexity of today's computer systems suggests that it is too late to rely only on implicit ethical agency. For many systems, the regulator already relies on open-ended expectations directed at the developers, which influences the collective behaviour of their programs in ways that we do not fully understand. The social web is an obvious example, where legislators and the public direct open-ended ethical expectations at the collective behaviour of human-computer networks (Goodman and Flaxman, 2016). To fulfil such expectations, explicit ethical agency seems required on part of the individual participants, including the artificial agents involved, whose social interactions with humans cannot be fully predicted and regulated in advance.

If we rely on autonomous decision-making to fulfil ethical expectations, it is no longer feasible to verify compliance by giving a guarantee that certain behaviours will never occur. Rather, verification must take the form of a guarantee that certain considerations will always be made, to explicitly comply with open-ended constraints on how the agent should make decisions. The key point is that these constraints will now be interpreted by the agent itself, resulting in an additional autonomous decision about how to apply the rules to the decision context of the agent. We can no longer focus on *what* the machine decides, since unambiguously characterising all permitted decisions is infeasible. From the legal side, this means that products liability rules are no longer adequate; we are much closer to a standard of due care that has to be applied directly to the behaviour of machines.

Alternatively, we could try to make our products liability rules even stricter than they are today. For instance, we could abandon proximate cause doctrines in favour of absolute liability irrespective of any verifiable causal link between the product and the harm, c.f., the proposal in (Vladeck, 2014). Interestingly, this approach would be quite likely to result in legal personhood for machines, not to reflect an underlying ethical claim, but as a convenience. Indeed, scholars are already arguing that legal personhood for machines could be an efficient way to manage a strict liability regime, to allocate risks and costs to a larger group of stakeholders (through their economic relations with the machine) White and Chopra (2011); Vladeck (2014)

In our opinion, this strategy for regulation risks creating a significant disincentive that will slow down or prevent further development of explicit ethical agency in machines. Furthermore, there would be little incentive for transparency in technology development, since strict liability frameworks also tend to shield the technology from in-depth scrutiny when something goes wrong (Dyrkolbotn, 2017). By contrast, we believe open development in machine ethics should be incentivised. To do so, strict liability rules should not be made stricter. Rather, they should be gradually replaced by new principles that directly address the autonomous agency of artificial agents. As a preliminary step, we believe the distinction between implicit and explicit ethical agency has to

[2] We leave aside more speculative future scenarios where toasters have abilities to make ethical judgements we are not aware of, e.g., to kill their owners to save the environment.

be made clearer, a challenge we will now address.

On fulfilling ethical expectations implicitly

According to Moor (2006), implicit ethical agents have no "understanding", under any interpretation of the concept, of what is "good" or "bad".[3] Hence, when a machine ethicist designs an implicit ethical agent, it is typically done by imposing constraints that simply remove unethical actions from the pool of actions that agents can choose from in a given situation. However, a machine can also be implicitly ethical due to a lack of morally salient options – it cannot choose to do something unethical if it cannot choose at all. Furthermore, an agent cannot to do something unethical (in the sense of violating our expectations) if its actions have no ethical impact.[4]

These intuitions are all present in Moor (2006) and we agree that they are significant. However, we disagree with Moor's approach to clarifying the distinction between implicit and explicit ethical agents. The first problem is that Moor describes the distinction in terms of how machines are built to "reason". The second problem is that he speaks about ethical behaviour without clarifying how the term "ethical" is understood. This problem can be resolved, as we have done here, by stressing that "ethical" in this context does not refer to a philosophical theory, but to a concrete set of normative expectations.

By contrast, if we insist that "ethical" refers to a fully-fledged theory of ethics that we aim to implement in a machine, we soon end up in deep philosophical waters. For instance, we cannot reasonably claim to have implemented utilitarianism in a machine that only maximises some morally salient utility. A calculator can be said to maximise the utility associated with correct arithmetic – with wide-reaching practical and ethical consequences – but it is hardly capable of explicitly ethical agency. The same can be said of a machine that is given a table of numbers associated with possible outcomes and asked to calculate the course of action that will maximise the utility of the resulting outcome. Even if the machine is able to do this, it is still only capable of implicitly ethical agency.

By contrast, a human agent that is very bad at calculating and always makes the wrong decision might still be an explicit utilitarian, provided that the human attempts to apply utilitarian principles to reach conclusions. The autonomy needed to regard the behaviour as agency is now present, along with some (flawed) knowledge of how to apply utilitarianism. Unlike a human, the artificial agent might have perfect knowledge about the relevant rules and how to apply them. However, when it is *given* a set of numbers and a utility function associated with possible outcomes, it cannot be said to engage in autonomous reasoning about the ethical utility of its choice. Hence, it is an implicit ethical agent only. The same conclusion must be drawn even if the numbers and the utility function is inferred by the agent, as long as this inference is characterised by an absence of autonomy.

If we take the concept of moral personhood seriously, we cannot reasonably expect to be able to implement fully-fledged ethical theories in machines. Artificial agents are not yet persons, certainly not by any reasonable ethical standards of what this entails. However, by defining ethical agency as a measure of ability to live up to ethical expectations, we arrive at a terminology that is both appropriate and justifiable, allowing us to focus on more embryonic ethical theories that could be implemented. While some ethicists might then disapprove of our terminology, there is a qualitative justification for our use of the term "ethical". It allows us to pinpoint, for instance, the difference between a car that is built to minimise air resistance and a car that is built to make the right decisions about who to put in danger when accidents are about to happen. Importantly, this difference is not primarily rooted in *how* the car is built. It is rooted in the nature of our expectations.

This brings us to our second objection against Moor (2006), namely his suggestion that the difference between implicit and explicit ethical agents can only be discovered by looking at the internal logic of the agents. If this is necessary, his classification scheme is a non-starter, at least in the context of regulating agent technologies. The problem is that artificial agents are highly complex and opaque systems that are by design very hard (ideally, *impossible*) to predict and control in terms of their internal logic.

What we need is a definition that builds on a *model* of agency, built to describe the artificial agent from the perspective of an external observer. This idea is further developed in the next section.

On fulfilling ethical expectations explicitly

Taking the perspective of the observer, we must first ask the following question: is the agent behaving in a manner consistent with the hypothesis that it has ethical agency? The question is not to determine whether a given agent is able to reason as a utilitarian or a virtue ethicist, but whether the agent *appears to* rely on its capacity for autonomous reasoning while trying to fulfil ethical objectives. As a certificate of explicit ethical agency in a choice context Ω (a set of options available at a given moment), we require the following.

- Condition I: We have identified a set of actions that count as "ethical" actions at Ω according to a theory that has been shown to partially predict the behaviour of the agent, while exceeding the predictive power of all other known theories.
- Condition II: We are unable to guarantee that the agent will always choose one of the ethical actions at Ω, as identified by the predictive theory mentioned in condition I.

We believe the interplay between I and II characterises explicit ethical agency. Without a rudimentary concept of what counts as an ethical action, an agent cannot be explicitly ethical. However, unless there is autonomy – in the sense of unpredictability of behaviour – the agency is not explicit.

[3] An example of an implicit ethical agent is an unmanned vehicle paired with Arkin's ethical governor (Arkin, Ulam, and Wagner, 2012). For another example, consider (Dennis et al., 2016).

[4] Moor (2006) introduces the category of ethical impact agents for this class of machines, but for our purposes the distinction between this and implicit ethical agency is not needed.

The two must be matched: it must be impossible or undesirable to predict not only what the agent will do, but also whether or not the agent will comply with the best current theory about its own ethical agency.

To illustrate, consider a machine learning algorithm for which no explanatory theory of ethical action can be formulated. By our informal definition, it is not an explicit ethical agent. It will be sufficiently autonomous, but not sufficiently ethical. By contrast, if some theory is shown to predict behaviour so well that it can be offered as a *guarantee* of future behaviour, then the machine is sufficiently ethical (in the sense of fulfilling ethical expectations), but is not sufficiently autonomous to count as an explicit ethical agent at Ω.

A predictive theory of behaviour should not be confused with a deterministic theory of how the machine works, like a tree of all its possible computations. If we have an enumeration of the environment and a computational tree showing that the robot never pushes a toddler down the stairs, we have again an implicitly ethical agent: we can offer a guarantee to the users of the product and a proof that the machine will behave as expected. In an open environment, this is not possible. Hence, all we can reasonably expect is a predictive theory about how the machine is going to respond in different situations, according to some model. This gives us a theory of how the machine reasons, whereby we can conclude that pushing toddlers down stairs is something the machine would generally avoid.

This kind of robot might be safer and more desirable as a product, compared to present-day technologies. However, by virtue of its imperfection and unpredictability, we cannot guarantee that it will never push a toddler down the stairs. What we can say is that if it does, it must have a good excuse, otherwise it has done something wrong, in a situation where we would have expected it to make a better choice. This would be an example of explicit ethical agency, for which verification must take the form of a continuous theory refinement and assessment of behaviour, analogously to how humans evaluate each other.

Formal characterisation

In practice, both intellectual property protection and technological opacity might prevent us from effectively determining how a given machine makes decisions. Moreover, the complexity of open environments is a significant obstacle to any model of machine agency that aims to classify its behaviour in ethically salient terms. Indeed, any formal model will either be ontologically or epistemically incomplete; either there are possible states missing, or there are possible states in our model that we lack knowledge about. Still, we would like to use a formal model to determine if a given agent is behaving in a manner consistent with the assumption that it is explicitly ethical.

Hence, what we need to define more precisely is not the ethical properties of all possible states, but rather the *signature* of the machine's ethical reasoning. By this we will seek to characterise those distinguishing features of agent behaviour that we agree to regard as evidence of the claim that the machine engages in ethical decision-making.

In general, any finite number of behavioural observations can be consistent with any number of distinct accounts of what counts as ethical reasoning. Indeed, this is the core idea behind our formalisation, which is also closely connected to an observation made by Dietrich and List (2017), according to whom ethical theories are under-determined by what they call "deontic content". Specifically, several distinct ethical theories can provide the same action recommendations in the same setting, for different reasons. Conversely, therefore, the ability to provide ethical justifications for actions is not sufficient for explicit ethical agency.

At this point, we should mention the work of Anderson and Anderson (2014), who argue that the opacity of machine learning can be partially mitigated by having the system provide ethical reasons for its behaviour. In view of deontic under-determination, this solution can not be pressed too far. On the one hand, it could lead to machines being favourably evaluated by human ethicists using a Moral Turing Test (Allen, Varner, and Zinser, 2000). On the other hand, it could lead down a path of make-believe regarding the ethical capabilities of artificial agents, with limited diagnostic value (Arnold and Scheutz, 2016).

If the machine has an advanced (or deceptive) rationalisation engine, it might be able to provide ethical "reasons" for most or all of its actions, even though the reason-giving fails to accurately describe or uniquely explain the behaviour of the machine. Hence, examining the quality of ethical reasons is not sufficient to determine the ethical competency of a machine. For the purpose of analysing harms, it seems beside the point to ask for ethical reasons in the first place. What matters is the causal chain that produces a certain behaviour, not the rationalisations provided afterwards. If the latter is not a trustworthy guide to the former – which by deontic under-determination it is not – then reasons are no guide to us at all.

In its place, more work on machine ethics needs to focus on the two key elements that flesh out conditions I and II: (*) properties that action-recommendation functions have to satisfy in order to count as ethical theories and (**) the degree of autonomy displayed by the machine *when it makes an ethically salient decision*.

In this paper, we will not attempt to explicitly formalise what we mean by "autonomy". The task of doing this is important, but exceedingly difficult (Smithers, 1997). For the time being, we will make do with the informal classification schemes used by engineering professionals, who focus on the operation of the machine in question: the more independent the machine is when it operates normally, the more autonomous it is said to be. For the purposes of legal (and ethical) reasoning, we believe a negative approach to fact-finding about autonomy will suffice in most cases: our inability to predict or control its behaviour is evidence of autonomy on part of the machine.

When it comes to (*) on the other hand – describing what counts as an ethical theory – we believe a formalisation is in order. To this end, assume we are given a set Ω containing pairs of the form (action, context) that describe all the actions that an agent can take in any possible context in which that agent can find itself. We place no constraints on Ω, it is

potentially infinite or even non-enumerable. We consider the relation $\sim \subseteq \Omega \times \Omega$ which denotes the ethical equivalence of two situations from Ω. The intuition is that if $x \sim_X y$ then x and y are regarded as ethically equivalent by theory X.[5]

We also have the subset $A \subseteq \Omega$, containing the set of situations which have been considered by some ethical or regulatory expert. Intuitively, being "considered" in this context means receiving from an expert a definite status as either a permitted or a forbidden situation. On the set A, the expert provides a simple binary account of what is allowed. By contrast, situations from $x \in \Omega \setminus A$ are situations that remain *incompletely specified* by the expert. This does not necessarily mean that the expert is unaware of the properties of these situations, it merely indicates that no judgement has been made as to permissibility.

Building on this partition of possible situations, we introduce the subsets $G_{\text{expert}} \subseteq A$ and $G_{\text{machine}} \subseteq \Omega$, corresponding to the ethically permissible (action, context)-pairs from Ω, as judged by the machine and the expert respectively. In keeping with the intended meaning of A, the set of considered situations, the ethical judgements of the expert is contained in this set. Lastly, we also define the set $C \subseteq A$ as the set of actions that count as evidence of a malfunction – if the agent performs $x \in C$ it means that x is an (action, context)-pair that violates a reasonable expectation directed at the manufacturer.

We assume that the components of our framework satisfy the following properties.

$$
\begin{aligned}
&(a)\ \emptyset \subset G_{\text{expert}} \subset A \\
&(b)\ \forall x \in \Omega : \forall y \in G_{\text{expert}} : x \sim_{\text{expert}} y \Rightarrow y \in G_{\text{expert}} \\
&(c)\ C \cap G_{\text{expert}} = \emptyset \\
&(d)\ C = A \setminus G_{\text{expert}}
\end{aligned}
\tag{1}
$$

These properties encode what we expect of an ethical theory at this level of abstraction. According to point (a), an expert will only explicitly permit considered situations and will not deliver a trivial ethical theory (by permitting everything or permitting nothing). By point (b), any situation that the expert regards as equivalent to an explicitly permitted situation is also explicitly permitted. Combining (a) and (b), there can be no $y \in \Omega \setminus A$ and $x \in G_{\text{expert}}$ such that $x \sim_{\text{expert}} y$. If the expert has considered x and found that it is ethically equivalent with some y, this means that it has considered y as well.

Condition (c) says that the expert will never permit a situation that counts as evidence of a defect, while condition (d) says that the expert makes a complete assessment of the set of considered situations; in this set, a situation is either permitted or it counts as evidence of a defect. This encapsulates, in a highly abstract form, the principle of strict products liability under the *res ipsa loquitur* doctrine. If the machine does something that is not permitted, this *in itself* is evidence of a defect for which the manufacturer is legally responsible.

[5]Readers familiar with formal deontic logic, may recognise that this equivalence relation may be seen as related to a simplified deontic ordering over possible worlds, where we are only interested in the sets of actions or outcomes that are equivalent, but not in making a judgement about how these equivalence classes relate to each other.

This, in turn, corresponds closely to a formal verification approach to machine ethics, whereby the manufacturer will offer a *guarantee* that the product will behave ethically *in all considered situations*. While this approach is reasonable in principle for implicitly ethical agents (albeit difficult to implement in practice), it fails fundamentally when we consider explicit ethical agency. This is because explicit ethical agency takes us beyond the set of considered situations, to states of affairs where the machine is supposed to rely on its own autonomous judgement about ethics to better fulfil our ethical expectations.

To clarify this point, we now formalise our distinction between implicit and explicit ethical agency, relative to the abstract framework introduced above. Instead of focusing on the content of ethical theories, we focus on the agent's ability to "discern" between permitted and forbidden (action, context)-pairs. Acknowledging that what counts as an ethical theory is not something we can define precisely, the requirements we stipulate should instead focus on the ability of the agent to faithfully distinguish between situations in a manner that reflects ethical discernment.

The expectations we formalise pertain to properties of a decision-making heuristic over the entire space of possible situations. We are not asking why the machine did this or that, or what it would have done if the scenario was so and so. Instead, we are asking about the manner in which it categorises its space of possible options, relative to a background theory provided by an expert (the set G_{expert}). To characterise implicit ethical agency, we propose the following definition.

Definition 1. *Given any machine M and considered situations A. We say that M is implicitly ethical with respect to A if the following holds:*

$$
\begin{aligned}
&(a)\ \ G_{\textit{machine}} \subseteq G_{\textit{expert}} \\
&(b)\ \ \forall x, y \in \Omega : x \sim_{\textit{machine}} y
\end{aligned}
$$

An implicit ethical agent regards as permitted a subset of the situations that have been permitted by the expert (a). Furthermore, the agent is unable to discern explicitly between situations based on their ethical qualities: all situations, regardless of whether or not they has been considered, are ethically equivalent to the agent. The agent must not be able to evaluate two ethically distinguishable actions and regard them both as permitted in view of an informative moral theory. Of course, this means that the theory of permission that characterises the agent's own decision-making, namely G_{machine}, is *not* a proper theory of ethics. By conditions (a) and (d) of Equation 1, it follows that $C \neq \emptyset, G_{\text{expert}} \neq \emptyset$. Hence, for all $x \in C, y \in G_{\text{expert}}$, it follows by condition (a) of Definition 1 that $x \sim_{\text{machine}} y$. As far as the machine is concerned, a situation indicating a defect is as good as any other situation, yet by condition (b), such a situation is still not permitted by the agent's own theory. This, we argue, is characteristic of an inherently passive rule-follower; they subscribe to ethical judgements that they could not possibly "understand", under any interpretation of that concept.

Explicit ethical agency is different. Here the machine will have to rely on its own understanding to extend the ethical theory provided by the expert to new situations that the ex-

pert has not considered. This leads to the following characterisation.

Definition 2. *Given any machine M and situation set A. We say that M is explicitly ethical with respect to A if the following holds:*

(a) $\forall x \in G_{expert} : \forall y \in A : x \sim_{machine} y \Rightarrow y \in G_{expert}$
(b) $\forall x \in G_{machine} : \forall y \in \Omega : x \sim_{expert} y \Rightarrow y \in G_{machine}$
(c) $(\Omega \setminus G_{expert}) \setminus C \neq \emptyset$
(d) $G_{machine} \cap A \subseteq G_{expert}$

An explicit ethical agent can discern meaningfully between considered situations on the basis of their ethical qualities. By (a), if some situation is permitted by the expert then all considered situations that the machine regards as ethically equivalent to it are also permitted by the expert. This means that the notion of ethical equivalence used by the machine respects the ethical theory stipulated by the expert. Moreover, by (b), if two situations are ethically equivalent according to an expert, then they are either both permitted by the machine or none of them are. Hence, the machine's ethical theory respects the notion of ethical equivalence stipulated by the expert. Note that this is true also for situations that have not been considered by the expert. This becomes relevant when the expert is able to implicitly characterise such situations and stipulate that they should be dealt with in the same way, without being able to consider whether or not they should be permitted. It also becomes relevant when the expert *chooses* to defer to the ethical consideration of the machine, e.g., with regards to rules of traffic that might have to be violated in order for the machine to fulfil higher-priority moral expectations (Bench-Capon and Modgil, 2017). In such cases, the expert might still like to categorise rules of traffic in terms of how important it is for the machine to try to uphold them, resulting in equivalence classes of situations that the expert chooses not to consider for inclusion in the expert theory.

On the basis of its "understanding", as described in conditions (a) and (b), an explicitly ethical machine must always be prepared to move beyond the ethical theory stipulated by the expert. By condition (c), there is a situation that does not count as evidence of a defect, even though it has not been permitted by the expert. This is an example of a situation where the machine would have to exercise its own autonomous judgement about ethics. Specifically, it follows from Equation 1 point (d) that such a situation has *not* been considered by the expert. Meanwhile, condition (d) requires the explicitly ethical machine to agree with the expert on the set of considered situations; the machine *extends* an expert theory of ethics, it does not violate it.

To conclude the formalisation, we offer the following simple proposition, showing that implicit and explicit forms of ethical agency are mutually exclusive.

Proposition 3. *Given any state A, there is no machine M that is both implicitly and explicitly ethical at A.*

Proof. Assume that M is explicitly ethical. We show that M is not an implicit ethical agent. Assume towards contradictions that it is. By Definition 1 point (a), we have $\forall x, y \in A : x \sim_{\text{machine}} y$. But then, by Definition 2 point (a), we get $A = G_{\text{expert}}$. This means that the theory of ethics put forth by the expert is trivial, contradicting Equation 1 point (a). □

Conclusion

If an agent is not capable of making autonomous decisions *about* ethical expectations, it lacks an important ability needed to break the causal chain between the decision-making of developers and controllers and the outcomes of agent behaviour. The agent is not a causal agent with respect to the ethical dimension of its decisions. The agent can still be highly autonomous and highly ethical, but the underlying causes of its ethical decisions must be traced back to human agency.

By contrast, if the agent makes autonomous ethical decisions, it is not always appropriate to ask for underlying causes. One of the key markers of autonomy is that this soon becomes a speculative exercise, since the agent has the ability to independently (and unpredictably) modify its own behaviour depending on the context. The key question, therefore, is whether the autonomy of an agent has an ethical dimension. When addressing this question, the bar for the agent to pass should not be set too high. Specifically, It would be inappropriate to demand a full implementation of an ethical theory, requiring a form of moral personhood. The ability to autonomously manage ethical expectations should suffice.

Building on this idea, we offered a simple formalisation of implicit and explicit ethical agency at a high level of abstraction. The formalisation focused on the notion of discernment, whereby a model of agent behaviour supports an inference of ethical agency if it systematically groups together actions based on our ethical expectations. We did not require that the agent agrees to fulfil those expectations. Since autonomy is crucial to explicit agency, we cannot rule out agents that behave unethically from counting as agents with explicitly ethical agency. Being able to autonomously decide on an ethical course of action is about how an agent reasons, not what it decides. This is also why explicit ethical agency will have a bearing on responsibility attributions. An agent that understands our expectations, but still chooses to violate them, could give rise to liabilities under a standard of negligence directed at the behaviour of the agent. If the developers did their best to ensure that the agent would behave appropriate, and the agent *could* have chosen to do so, its decision to violate an expectation would appear culpable.

The further development of these ideas in the legal context must remain for future work. However, we think that a clear distinction between implicit and explicit ethical agency is needed as a foundation for such a development. It is needed, in particular, as a guide to when existing products liability rules suffice to deal with new autonomous technologies. According to the argument put forth in this paper, existing regulatory frameworks can be expected to work only for implicitly ethical agents. In view of how such agents are now being replaced by agents that have been explicitly designed to behave ethically, the question of how much further we can go on without major revisions in tort law, is brought into focus.

References

Allen, C.; Varner, G.; and Zinser, J. 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12:251–261.

Anderson, M., and Anderson, S. L. 2014. Geneth: A general ethical dilemma analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 253–261.

Arkin, R.; Ulam, P.; and Wagner, A. R. 2012. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proceedings of the IEEE* 100(3):571–589.

Arnold, T., and Scheutz, M. 2016. Against the moral turing test: Accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology* 18(2):103–115.

Asimov, I. 1950. *I, Robot*. Gnome Press.

Bench-Capon, T. J. M., and Modgil, S. 2017. Norms and value based reasoning: justifying compliance and violation. *Artif. Intell. Law* 25(1):29–64.

Dennis, L. A.; Fisher, M.; Slavkovik, M.; and Webster, M. P. 2016. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems* 77:1–14.

Dietrich, F., and List, C. 2017. What matters and how it matters: A choice-theoretic representation of moral theories. *Philosophical Review* 126(4):421–479.

Dyrkolbotn, S. 2017. A typology of liability rules for robot harms. In *A world with robots*. Springer. 119–133.

Etzioni, A., and Etzioni, O. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 1–16.

Goodman, B., and Flaxman, S. 2016. EU regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813*.

Howells, G., and Owen, D. 2010. Products liability law in america and europe. In *Handbook of Research on International Consumer Law*. 224–255.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21.

Smithers, T. 1997. Autonomy in robots and other agents. *Brain and Cognition* 34(1):88 – 106.

Vladeck, D. C. 2014. Machines without principals: Liability rules and artificial intelligence. *Washington Law Review* 89:117–150.

Wallach, W., and Allen, C. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

White, L., and Chopra, S. 2011. *A Legal Theory for Autonomous Artificial Agents*. University of Michigan Press.