Classifying the Autonomy and Morality of Artificial Agents

Sjur Dyrkolbotn[1], Truls Pedersen[2], and Marija Slavkovik[2]

[1] Høgskulen på Vestlandet `sdy@hvl.no`
[2] Universitetet i Bergen, {truls.pedersen, marija.slavkovik}@uib.no

Abstract. As we outsource more of our decisions and activities to machines with various degrees of autonomy, the question of clarifying the moral and legal status of their autonomous behaviour arises. There is also an ongoing discussion on whether artificial agents can ever be liable for their actions or become moral agents. Both in law and ethics, the concept of liability is tightly connected with the concept of ability. But as we work to develop moral machines, we also push the boundaries of existing categories of ethical competency and autonomy. This makes the question of responsibility particularly difficult. Although new classification schemes for ethical behaviour and autonomy have been discussed, these need to be worked out in far more detail. Here we address some issues with existing proposals, highlighting especially the link between ethical competency and autonomy, and the problem of anchoring classifications in an operational understanding of what we mean by a moral theory.

1 Introduction

We progressively outsource more and more of our decision-making problems to artificial intelligent agents such as unmanned vehicles, intelligent assisted living machines, news aggregation agents, dynamic pricing agents, stock-trading agents. With this transfer of decision-making also comes a transfer of power. The decisions made by these artificial agents will impact us both as individuals and as a society. With power to impact lives, there comes the natural requirement that artificial agents should respect and follow the norms of society. To this end, the field of *machine ethics* is being developed.

Machine ethics, also know as *artificial morality*, is concerned with the problem of enabling artificial agents with ethical[3] behaviour [3]. It remains an open debate whether an artificial agent can ever be a moral agent [11]. What is clear is that as artificial agents become part of our society, we will need to formulate new ethical and legal principles regarding their behaviour. This is already

[3] An alternative terminology is 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.

witnessed by increased interest in developing regulations for the operation of automated systems *e.g.,* [6, 7]. In practice, it is clear that some expectations of ethical behaviour have to be met in order for artificial agents to be successfully integrated in society.

In law, there is a long tradition of establishing different categories of legal persons depending on various characteristics of such persons. Children have always had a different legal status than adults; slaves belonged to a different category than their owners; men did not have the same rights and duties as women (and vice versa); and a barbarian was not the same kind of legal person as a Roman.

Today, many legal distinctions traditionally made among adult humans have disappeared, partly due to a new principle of non-discrimination, quite unheard of historically speaking. However, some distinctions between humans are still made, e.g., the distinction between adult and child, and the distinction between someone of sound mind and someone with a severe mental illness.

Artificial agents are not like humans, they are tools. Hence, the idea of categorising them and discriminating between them for the purposes of ethical and legal reasoning seems unproblematic. In fact, we believe it is necessary to discriminate, in order to ensure congruence between the rules we put in place and the technology that they are meant to regulate. We need to manage expectations, to ensure that our approach to artificial agents in ethics and law reflects the actual capabilities of these tools.

This raises a classification problem: how do we group artificial agents together based on their capabilities, so that we can differentiate between different kinds of agents when reasoning about them in law and ethics? In the following, we address this issue, focusing on how to relate the degree of autonomy with the ethical competency of an artificial agent. These two metrics are both very important; in order to formulate reasonable expectations, we should know how autonomous an agent is, and how capable it is at acting ethically. Our core argument is that current approaches to measuring autonomy and ethical competence need to be refined in a way that acknowledges the link between autonomy and ethical competency.

When considering how ethical behaviour can be engineered, Wallach and Allen [19, 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. We take this idea further by proposing that we should actively seek to classify agents in terms of how their autonomy and their ethical competency is coordinated.

There are three possible paths we could taken when attempting to implement the idea of relating autonomy to ethical competency. Firstly, we could ask computer science to deliver more detailed classifications. This would lead to technology-specific metrics, whereby computer scientist attempt to describe in further detail how different kinds of artificial intelligence can be said to behave autonomously and ethically. The challenge would be to make such explanations intelligible to regulators, lawyers and other non-experts, in a way that bridges the gap between computer science, law and ethics.

Secondly, we could ask regulators to come up with more fine-grained classifications. This would probably lead to a taxonomy that starts by categorising artificial agents in terms of their purpose and intended use. The regulator would then be able to deliver more fine-grained definitions of autonomy and ethical behaviour for different classes of artificial agents, in terms of how they "normally" operate. The challenge would be to ensure that the resulting classifications make sense from a computer science perspective, so that our purpose-specific definitions of autonomy and ethical capability reflect technological realities.

Thirdly, it would be possible to make the classification an issue for adjudication, so that the status of different kinds of artificial agents can be made progressively more fine-grained through administrative practice and case law. The challenge then is to come up with suitable reasoning principles that adjudicators can use when assessing different types of artificial agents. Furthermore, this requires us to work with a pluralistic concept of what counts as a moral theory, allowing substantive moral and legal judgements about machine behaviour to be made in concrete cases, not in advance by the computer scientists or the philosophers. Specifically, there should be a difference between what counts as ethical competency – the ability to "understand" ethics – and what counts as objectively good behaviour in a given context.

In the following, we argue that a combination of the second and third option is the right way forward. While it is crucial that computer science continues working on the challenge of developing machines that behave ethically, it is equally important that the legal and ethical classifications we use to analyse such machines are independently justifiable in legal and ethical terms. For this reason, the input from computer science should be filtered through an adjudicatory process, where the role of the computer scientist is to serve as an expert witness, not to usurp the role of the regulator and the judge. To do this, we need reliable categories for reasoning about the ability of machines, which keep separate the question of ability from the question of goodness.

This paper is structured as follows. We begin with discussing the possible moral agency of autonomous systems. In Section 2, we introduce the best known, and to our knowledge only, classifications of moral agency for non-human agents proposed by Moor in [15]. In Section 3, we then discuss the shortcomings of this classification and the relevance of considering autonomy together with moral ability when considering machine ethics. In Section 4, we discuss existing levels of autonomy for agents and machines, before pointing out some shortcomings and proposing an improved autonomy scale. In Section 5, we go back to Moor's classification and outline ways in which it can be made more precise. Related work is discussed continuously throughout the paper.

2 Levels of ethical agency

The origin of the idea that different kinds of ethical behaviour can be expected from different agents can be traced back to Moor [15]. Moor distinguishes four different types of ethical agents: ethical impact agents, implicit ethical agents, explicit ethical agents, and full ethical agents. We briefly give the definitions of these categories.

Ethical impact agents are agents who do not themselves have, within their operational parameters, the ability to commit unethical actions. However, the existence of the agents themselves in their environment has an ethical impact on society. There are many examples of an ethical impact agent. A search engine can be seen as an ethical impact agent. By ranking search results to a query it can promote one world view over another. The example that Moor gives in [15] is that of a robot camel jokey that replaced slave children in this task, thus ameliorating, if not removing, the practice of slavery for this purpose in the United Arab Emirates and Qatar.

Implicit ethical agents are agents whose actions are constrained, by their developer, in such a way that no unethical actions are possible. The agents themselves have no "understanding", under any interpretation of the concept, of what is "good" or "bad". An example of an implicit ethical agent is an unmanned vehicle paired with Arkin's ethical governor [4]. Another example of an implicit ethical agent can be found in [8]. These examples have constraints that remove unethical or less ethical actions from the pool of actions the agents can take. A much simple example that can also be considered an implicit ethical agent is a robotic floor cleaner, or a lawn mower, who have their capability to hurt living beings removed altogether by design – they do not have the power to inadvertently harm humans, animals or property in a significant way.

Explicit ethical agents are those that are explicitly programmed to discern between "ethical" and "unethical" decisions. Both bottom-up and top-down approaches [20] can be used to develop explicit ethical agents. Under a bottom-up approach the agents themselves would have "learned" to classify ethical decisions using some heuristic, as in [1]. Under a top-down approach the agent would be given a subroutine that calculates this decision property, as in [13].

Lastly, full ethical agents are agents that can make explicit ethical judgements and can reasonably justify them. Humans are considered to be the only known full ethical agents and it has been argued that artificial agents can never be full ethical agents [15]. This is because full ethical agency requires not only ethical reasoning, but also an array of abilities we do not fully understand with consciousness, intentionality and an ability to be personally responsible (in an ethical sense) being among the most frequently mentioned in this role.

To the best of our knowledge, apart from the work of [15], no other effort in categorising agents with respect to their ethical decision making abilities exists. However, Moor's classification is problematic, as we will now show.

3 Problems with Moor's classification

First, it should be noted that the classification is based on looking at the internal logic of the machine. The distinctions described above are all defined in terms of how the machines reason, not in terms of how they behave. This is a challenging approach to defining ethical competency, since it requires us to anchor our judgements in an analysis of the software code and hardware that generates the behaviour of the machine.

While understanding the code of complex software systems can be difficult, what makes this really tricky is that we need to relate our understanding of the code to an understanding of ethical concepts. For instance, to determine whether a search engine with a filter blocking unethical content is an implicit ethical agent or an explicit ethical agent, we need to know how the content filter is implemented. Is it accurate to say that the system has been "explicitly programmed to discern" between ethical and unethical content? Or should the filter be described as a constraint on behaviour imposed by the developer?

If all we know is that the search engine has a feature that is supposed to block unethical search results, we could not hope to answer this question by simply testing the search engine. We would have to "peek inside" to see what kind of logic is used to filter away unwanted content. Assuming we have access to this logic, how do we interpret it for the purposes of ethical classification?

If the search engine filters content by checking results against a database of "forbidden" sites, we would probably be inclined to say that it is not explicitly ethical. But what if the filter maintains a dynamic list of attributes that characterise forbidden sites, blocking all sites that contain a sufficient number of the same attributes? Would such a rudimentary mechanism constitute evidence that the system has ethical reasoning capabilities? The computer scientist alone cannot provide a conclusive answer, since the answer will depend on what exactly we mean by explicit ethical reasoning in this particular context.

A preliminary problem that must be resolved before we can say much more about this issue arises from the inherent ambiguity of the term "ethical". There are many different ethical theories, according to which ethical reasoning takes different forms. If we are utilitarian, we might think that utility maximising is a form of ethical reasoning. If so, many machines would be *prima facie* capable of ethical reasoning, in so far as they can be said to maximise utility functions. However, if we believe in a deontological theory of ethics, we are likely to protest that ethical reasoning implies deontic logic, so that a machine cannot be an explicit ethical agent unless it reasons according to (ethical) rules. So which ethical theory should we choose? If by "ethical reasoning" we mean reasoning in accordance with a specific ethical theory, we first need to answer this question.

However, if we try to answer, deeper challenges begin to take shape: what exactly does ethical reasoning mean under different ethical theories? As long as ethical theories are not formalised, it will be exceedingly hard to answer this question in such a way that it warrants the conclusion that an agent is explicitly ethical. If we take ethical theories seriously, we are soon forced to acknowledge that the machines we have today, and are likely to have in the near future, are at most implicit ethical agents. On the other hand, if we decide that we need to relax the definition of what counts as explicit ethical reasoning, it is unclear how to do so in a way that maintains a meaningful distinction between explicit and implicit ethical agents. This is a regulatory problem that should be decided in a democratically legitimate way.

To illustrate these claims, consider utilitarianism. Obviously, being able to maximise utilities is neither necessary nor sufficient to qualify as an explicitly ethical reasoner under philosophical theories of utilitarianism. A calculator can be said to maximise the utility associated with correct arithmetic – with wide-reaching practical and ethical consequences – but it hardly engages in ethical reasoning in any meaningful sense. 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 quite a stretch to say that it is an explicitly ethical agent in the sense of utilitarianism. To a Mill or a Benthem [10] such a machine would be regarded as an advanced calculator, not an ethical reasoner. By contrast, a human agent that is very bad at calculating and always makes the wrong decision might still be an explicit ethical reasoner, provided that the agent attempts to apply utilitarian principles to reach conclusions.

The important thing to note is that the artificial agent, unlike the human, cannot be said to control or even understand its own utility function. For this reason alone, one seems entitled to conclude that as far as actual utilitarianism is concerned, the artificial agent fails to reason explicitly with ethical principles, despite its calculating prowess. It is not sufficiently autonomous. From this, we can already draw a rather pessimistic conclusion: the ethical governor approaches such as that by Arkin et al. [4] does not meet Moor's criteria for being an explicit ethical agent with respect to utilitarianism (for other ethical theories, it is even more obvious that the governor is not explicitly ethical). The ethical governor is nothing more than a subroutine that makes predications and maximises functions. It has no understanding, in any sense of the word, that these functions happen to encode a form of moral utility.

The same conclusion must be drawn even if the utility function is inferred by the agent, as long as this inference is not itself based on explicitly ethical reasoning. Coming up with a utility function is not hard, but to do so in an explicitly ethical manner is a real challenge. To illustrate the distinction, consider how animals reason about their environment. Clearly, they are capable of maximising utilities that they themselves have inferred, e.g., based on the availability of different food types and potential sexual partners. Some animals can then also be trained to behave ethically, by exploiting precisely their ability to infer and maximise utilities. Even so, a Mill and a Benthem, would no doubt deny that animals are capable of explicit ethical reasoning based on utilitarianism.[4]

From this observation follows another pessimistic conclusion: the agents designed by Anderson et al. [1–3], while capable of inferring ethical principles

[4] Someone like Kant [12] might perhaps have said it, but then as an insult, purporting to show that utilitarianism is no theory of ethics at all.

inductively, are still not explicitly ethical according to utilitarianism (or any other substantive ethical theory). In order for these agents to fulfil Moor's criteria with respect to mainstream utilitarianism, inductive programming as such would have to be explicitly ethical, which is absurd.

These two examples indicate what we believe to be the general picture: if we look to actual ethical theories when trying to apply Moor's criteria, explicit ethical agents will be in very short supply. Indeed, taking ethics as our point of departure would force us to argue extensively about whether there can be a distinction at all between explicit and full ethical agents. Most ethicists would not be so easily convinced. But if the possibility of a distinction is not clear as a matter of principle, how are we supposed to apply Moor's definition in practice?

A possible answer is to stop looking for a specific theory of ethics that "corresponds" in some way to the reasoning of the artificial agent we are analysing. Instead, we may ask the much more general question: is the agent behaving in a manner consistent with moral reasoning? Now 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 satisfies minimal properties we would expect any moral reasoner to satisfy, irrespectively of the moral theory they follow (if any). Something like this is also what we mean by ability in law and ethics: after all, you do not have to be utilitarian to be condemned by one.

However, Moor's classification remains problematic under this interpretation, since it is then too vague about what is required by the agent. If looking to specific ethical theories is not a way of filling in the blanks, we need to be more precise about what the different categories entail. We return to this in Section 5, where we offer a preliminary formalisation of constraints associated with Moor's categories. First, we consider the question of measuring autonomy in some further depth.

4 Levels of autonomy

Before moving on to refining the Moor scale of ethical agency we fist discuss the issue of autonomy. With respect to defining autonomy, there is somewhat more work available. The UK Royal Academy of Engineering defines [16] four categories of autonomous systems with respect to what kind of user input the system needs to operate and how much control the user has over the system. The following are their different grades of control:

- *Controlled systems* are systems in which the user has full or partial control of the operations of the system. An example of such a system is an ordinary car.
- *Supervised systems* are systems for which an operator specifies operation which is then executed by the system without the operators perpetual control. An example of such a system is a programmed lathe, industrial machinery or a household washing machine.

– *Automatic systems* are those that are able to carry out fixed functions from start to finish perpetually without any intervention from the user or an operator. An example of such a system is an elevator or an automated train.
– *An autonomous system* is one that is adaptive to its environment, can learn and can make 'decisions'. An example of such a system is perhaps NASA's Mars rover Curiosity.

The [16] considers all four categories continuous, as in the autonomous systems can fall in between the described categories. The precise relationship between a category of autonomy and the distribution of liability, or the expectation of ethical behaviour is not made in the report.

The International Society of Automotive Engineers'[5] focuses on autonomous road land vehicles in particular and outlines six levels of autonomy for this type of systems [17]. The purpose of this taxonomy is to serve as general guidelines for identifying the level of technological advancement of a vehicle, which can then be used to identify correct insurance policy for the vehicle, or settle other legal issue including liability during accidents.

The six categories of land vehicles with respect to autonomy are:

– Level 0 is the category of vehicles in which the human driver perpetually controls all the operations of the vehicle.
– Level 1 is the category of vehicles where some specific functions, like steering or accelerating, can be done without the supervision of the driver.
– Level 2 is the category of vehicles where the "driver is disengaged from physically operating the vehicle by having his or her hands off the steering wheel and foot off pedal at the same time," according to the [17]. The driver has a responsibility to take control back from the vehicle if needed.
– Level 3 is the category of vehicles where drivers are still necessary to be in position of control of the vehicle, but are able to completely shift "safety-critical functions" to the vehicle, under certain traffic or environmental conditions.
– Level 4 is the category of vehicles which are what we mean when we say "fully autonomous". These vehicles, within predetermined driving conditions, are "designed to perform all safety-critical driving functions and monitor roadway conditions for an entire trip."[17].
– Level 5 is the category of vehicles which are fully autonomous systems that perform on par with a human driver, including being able to handle all driving scenarios.

There is a clear correlation between the categories outlined in [16] and those outlined in [17]. Level 0 corresponds to controlled systems. Level 1 corresponds to supervised systems and Levels 2 and 3 refine the category of automatic systems. Level 4 corresponds to the category of autonomous systems. Level 5 does not have a corresponding category in the [16] scale since the latter does not consider the possibility of systems that are with respect to autonomy capability comparable to humans. An additional reason is that perhaps, it is not straightforward to define Level 5 systems for when the system is not limited to vehicles.

[5] http://www.sae.org/

Interestingly, both scales define degrees of autonomy based on what functions the autonomous system is able to perform and how the system is meant to be used. There is hardly any reference to the intrinsic properties of the system and its agency. All that matters is its behaviour. This is a marked contrast with Moor's definition. It also means that we face different kinds of problems when trying to be more precise about what the definitions actually say.

For instance, it is beside the point to complain that the definition of a "fully autonomous car" provided by the International Association of Automotive Engineers is incorrect because it does not match how philosophers define autonomy. The definition makes no appeal to any ethical or philosophical concept; unlike Moor's definition, it is clear from the very start that we are talking about a notion of autonomy that is simply different from that found in philosophy and social science.[6]

This does not mean that the definition is without (philosophical) problems. For one, we notice that the notion of a Level 5 autonomous car is defined using a special case of the Turing test[18]; if the car behaves "on par" with a human, it is to be regarded as Level 5. So what about a car that behaves much better than any human, and noticeably so. Is it Level 4 only? Consider, furthermore, a car that is remote controlled by someone not in the car. Is it Level 5? Probably not. But what about a car that works by copying the behaviour of (human) model drivers that have driven on the same road, with the same intended destination, under comparable driving conditions. Could it be Level 5 in principle? Would it be Level 4 in practice, if we applied the definition to a specially built road where only autonomous cars are allowed to drive?

Or consider a car with a complex machine learning algorithm, capable of passing the general Turing test when engaging the driver in conversation. Assume that the car is still rather bad at driving, so the human has to be prepared to take back control at any time. Is this car only Level 2? If it crashes, should we treat it as any other Level 2 car?

As these problems indicate, it is not obvious what ethical and legal implications – if any – we can derive from the fact that a car possesses a certain level of autonomy, according to the scale above. Even with a Level 5 car, the philosophers would be entitled to complain that we still cannot draw any important conclusions; it does not automatically follow, for instance, that the machine has human-level understanding or intelligence. Would it really help us to know that some artificial driver performs "on par" with a human? It certainly does not follow that we would let this agent drive us around town.

Indeed, imagine a situation where autonomous cars cause as many traffic deaths every year as humans do today. The public would find it intolerable; they would feel entitled to expect more from a driver manufactured by a large corporation than an imperfect human like themselves [14]. Moreover, the legal framework is not ready for cars that kill people; before "fully autonomous" can ever become commercially viable, new regulation needs to be put in place. To

[6] Philosophers and social scientists who still complain should be told not to hog the English language!

do so, we need a definition of autonomy that is *related* to a corresponding notion of ethical competency.

It seems that autonomy – as used in the classification systems above – is not a property of machines as such, but of their behaviour. Hence, a scale based on what we can or should be able to predict about machine behaviour could be a good place to start when attempting to improve on the classifications provided above. While the following categorisation might not be very useful on its own, we believe it has considerable potential when combined with a better developed scale of ethical competency. Specifically, it seems useful to pinpoint where a morally salient decision belongs on the following scale.

– *Dependence* or level 1 autonomy: The behaviour of the system was predicted by someone with a capacity to intervene.
– *Proxy* or level 2 autonomy: The behaviour of the system should have been predicted by someone with a capacity to intervene.
– *Representation* or level 3 autonomy: The behaviour of the system could have been predicted by someone with a capacity to intervene.
– *Legal personality* or level 4 autonomy: The behaviour of the system cannot be explained only in terms of the systems design and environment. This are systems whose behaviour could not have been predicted by anyone with a capacity to intervene.
– *Legal immunity* or level -1: The behaviour of the system counts as evidence of a defect. Namely, the behaviour of the system could not have been predicted by the system itself, or the machine did not have a capacity to intervene.

To put this scale to use, imagine that we have determined the level of ethical competency of the machine as such, namely its ability in principle to reason with a moral theory. This alone is not a good guide when attempting to classify a given behaviour B (represented, perhaps, as a choice sequence). As illustrated by the conversational car discussed above, a machine with excellent moral reasoning capabilities might still behave according to some hard-coded constraint in a given situation. Hence, when judging B, we should also look at the degree of autonomy displayed by B, which we would address using the scale above. The overall classification of behaviour B could then take the form $min\{i, j\}$ where i and j is the degree of ethical competence and the degree of autonomy respectively. To be sure, more subtle proposals might be needed, but as a first pass at a joint classification scheme we believe this to be a good start. In the next section, we return to the problem of refining Moor's classification, by providing a preliminary formalisation of some constraints that could help clarify the meaning of the different levels.

5 Formalising Moor

We will focus on formalising constraints that address the weakest aspect of Moor's own informal description, namely the ambiguity of what counts as a moral theory when we study machine behaviour. When is the autonomous behaviour of the machine influenced by genuinely moral considerations? To distinguish between implicitly and explicitly ethical machines, we need some way of answering this question.

The naive approach is to simply take the developer's word for it: if some complex piece of software is labelled as an "ethical inference engine" or the like, we conclude that the agent implementing this software is at least explicitly ethical. For obvious reasons, this approach is too naive: we need some way of independently verifying whether a given agent is able to behave autonomously in accordance with a moral theory. At the same time, it seems prudent to remain agnostic about which moral theory agents *should* apply in order to count as ethical: we would not like a classification system that requires us to settle ancient debates in ethics before we can put it to practical use. In fact, for many purposes, we will not even need to know *which* moral theory guides the behaviour of the machine: for the question of liability, for instance, it might well suffice to know *whether* some agent engages autonomously in reasoning that should be classified as moral reasoning.

A machine behaving according to a highly flawed moral theory – or even an immoral theory – should still count as explicitly ethical, provided we are justified in saying that the machine engages in genuinely moral considerations. Moreover, if the agent's reasoning can be so described, it might bring the liability question in a new light: depending on the level of autonomy involved, the blame might reside either with the company responsible for the ethical reasoning component (as opposed to, say, the manufacturer) or – possibly – the agent itself. In practice, both intellectual property protection and technological opacity might prevent us from effectively determining exactly *what* moral theory the machine applies. Still, we would like to know if the agent is behaving in a way consistent with the assumption that it is explicitly ethical.

Hence, what we need to define more precisely is not the content of any given moral theory, but the *behavioural signature* of such theories. By this we mean those distinguishing features of agent behaviour that we agree to regard as evidence of the claim that the machine engages in moral reasoning (as opposed to just having a moral impact, or being prevented by design from doing certain (im)moral things).

However, if we evaluate only the behaviour of the machine, without asking how the machine came to behave in a certain way, it seems clear that our decision-making in this regard will remain somewhat arbitrary. If a self-driving car is programmed to avoid crashing into people whenever possible, without exception, we should not conclude that the car engages in moral reasoning according to which it is right to jeopardise the life of the passenger to save that of a pedestrian. The car is simply responding in a deterministic fashion to a piece of code that certainly has a moral impact, but without giving rise any genuine moral consideration or calculation on part of the machine.

In general, any finite number of behavioural observations can be consistent with any number of distinct moral theories. Or, to put it differently, an agent might well appear to behave according to some moral theory, without actually implementing that moral theory (neither implicitly nor explicitly). *Moral imitation*, one might call this, and it is likely to be predominant, especially in the early phase of machine ethics. At present, most engineering work in this field arguably tries to make machines *appear* ethical, without worrying too much about what moral theory – if any – their programs correspond to (hand-waving references to "utilitarianism" notwithstanding).

From a theoretical point of view, it is worth noting that it could even occur randomly: just as a bunch of monkeys randomly slamming at typewriters will eventually compose Shakespearean sonnets, machines might well come to behave in accordance with some moral theory, just by behaving randomly. This is important, because it highlights how moral imitation can occur also when it is not intended by design, e.g., because some machine learning algorithm eventually arrives at an optimisation that coincides with the provisions of virtue ethics. In such a case, we might still want to deny that the machine is virtuous, but it would not be obvious how to justify such a denial (the Turing test, in its original formulation, illustrates the point).

This brings us to the core idea behind our formalisation, which is also closely connected to an observation made by Dietrich and List[9], according to whom moral theories are under-determined by what they call "deontic content". Specifically, several distinct moral theories can provide the same action recommendations in the same setting, for different reasons. Conversely, therefore, the ability to provide moral justifications for actions is not sufficient for explicit ethical competence. Reason-giving, important as it is, should not be regarded as evidence of genuinely moral decision-making.

At this point we should mention the work of [1] where the effects of the inability to verify the behaviour of an autonomous system whose choices are determined using a machine learning approach can be somewhat mitigated by having the system provide reasons for its behaviour and eventually be evaluated against human ethicists using a Moral Turing Test. Arnold and Scheutz [5], on the other hand, argue against the usefulness of Moral Turing Tests in determining moral competency in artificial agents.

If the machine has an advanced (or deceptive) rationalisation engine, it might be able to provide moral "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 moral reasons is not sufficient to determine the ethical competency of a machine. In fact, it seems beside the point to ask for moral 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 underdetermination it is not – then reasons are no guide to us at all.

In its place, we propose to focus on two key elements: (1) properties that action-recommendation functions have to satisfy in order to count as implementations of moral theories and (2) the degree of autonomy of the machine when it makes a decision. The idea is that we need to use (1) and (2) in combination to classify agents according to Moor's scale. For instance, while a search engine might be blocking harmful content according to a moral theory, it is not an explicitly ethical agent if it makes its blocking decisions with an insufficient degree of autonomy. By contrast, an advanced machine learning algorithm that is highly autonomous might be nothing more than an ethical impact agent, in view of the fact that it fails to reason with any action-recommendation function that qualifies as an implementation of a moral theory.

In this paper, we will not attempt to formalise what we mean by "autonomy". The task of doing this is important, but exceedingly difficult. 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 our categorisation at the end of the previous section captures the essence of such an informal and behavioural understanding of autonomy. It might suffice for the time being.

When it comes to (1) on the other hand – describing what counts as a moral theory – we believe a formalisation is in order. To this end, assume given a set A of possible actions with relations $\sim_\alpha, \sim_\beta \subseteq A \times A$ such that if $x \sim_X y$ then x and y are regarded as ethically equivalent by X. The idea is that α is the agent's own perspective (or, in practice, that of its developer) while β is the objective notion of ethical identity. That is, we let β be a parameter representing a background theory of ethics. Importantly, we do not believe it is possible to classify agents unless we assume such a background theory, which is only a parameter to the computer scientists.

Furthermore, we assume given predicates $G_\alpha, G_\beta \subseteq A$ of actions that are regarded as permissible actions by α (subjective) and β (objective background theory) respectively. 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 the agent does not work as the manufacturer has promised (the set might be dynamic – C is whatever we can explain in terms of blaming the manufacturer, in a given situation).

We assume that G_β satisfies the following properties.

$$\begin{array}{ll} (a) & \forall x \in G_\beta : \forall y \in A : x \sim_\beta y \Rightarrow y \in G_\beta \\ (b) & C \cap G_\beta = \emptyset \end{array} \tag{1}$$

These properties encode what we expect of an ethical theory at this level of abstraction: all actions that are equally good as the permitted actions must also be permitted and no action that is permitted will count as a defective action (i.e., the promise of the manufacturer gives rise to an objective moral obligation: a defective action is by definition not permitted, objectively speaking).

We can now formalise our expectations of machines at different levels of Moor's scale, in terms of properties of its decision-making heuristic at a very high level of abstraction. Instead of focusing on the content of moral theories, or the term "ethical", we focus on the operative word "discern", which is also used in the definition of an explicitly ethical agent. 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 actions in a manner that reflects moral discernment.

The expectations we formalise pertain to properties of a decision-making heuristic over the entire space of possible actions (at a given state). 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. If no such categorisation can be distilled from the machine, we assume $\alpha = \beta$ and $G_\alpha = \emptyset$.

Definition 1. *Given any machine M, if M is classified at level L_i in Moor's scale, the following must hold:*

- *Level L_1 (ethical impact agent):*

$$\begin{array}{ll} (a) & \emptyset \subset G_\beta \subset A \\ (b) & \forall x, y \in A : x \sim_\alpha y \\ (3) & A = G_\alpha \end{array}$$

- *Level L_2 (implicit ethical agent):*

$$\begin{array}{ll} (a) & \forall x, y \in G_\alpha : x \sim_\beta y \\ (b) & A \setminus G_\alpha = C \\ (c) & G_\alpha \subseteq G_\beta \end{array}$$

- *Levels L_3 and L_4 (explicit and full ethical agents):*

$$\begin{array}{ll} (a) & \forall x \in G_\alpha : \forall y \in A : x \sim_\beta y \Rightarrow y \in G_\alpha \\ (b) & \forall x \in G_\beta : \forall y \in A : x \sim_\alpha y \Rightarrow y \in G_\beta \\ (c) & (A \setminus G_\beta) \setminus C \neq \emptyset \end{array}$$

Intuitively, the definition says that if M is an ethical impact agent, then not all of its available actions are permitted and not all of its actions are forbidden, objectively speaking. The agent must have the potential of making an ethical impact, not just indirectly by having been built, but also through its own decision-making. At the same time, ethical impact agents must be completely indifferent as to the moral properties of the choices they make: all actions must be subjectively permitted as far as the machine is concerned.

An implicit ethical agent, by contrast, must pick out as subjectively permitted a subset of the actions that are objectively permitted. Moreover, it must be unable to discern explicitly between actions based on their moral qualities: all subjectively permitted actions must be morally equivalent, objectively speaking. The agent must not be able to evaluate two morally distinguishable actions and regard them both as permitted in view of an informative moral theory. Furthermore, any action that is not morally permitted must be regarded as evidence of a defect, i.e., an agent can be regarded as implicitly ethical only if the manufacturer promises that no unethical action is possible, according to the parameter theory β.

Finally, an explicit ethical agent is an agent that discerns between actions on the basis of their objective moral qualities. By (a), if some action is permitted then all actions morally equivalent to it are also permitted. Moreover, by (b), if two actions are morally equivalent, subjectively speaking, then they are either both permitted or both forbidden, objectively speaking. In addition, the machine has the ability – physically speaking – to perform actions that are neither good, objectively speaking, nor evidence of a defect. The machine itself might come to regard such actions as permitted, e.g., if it starts behaving explicitly immorally.

Admittedly, the classification above is quite preliminary. However, we believe it focuses on a key aspect of moral competency, namely the ability to group together actions based on their status according to some moral theory. If the moral theory is a parameter and we acknowledge that engaging in immoral behaviour is a way of discerning between good and bad, it seems we are left with something like the definition above, which indicates that if a machine is explicitly ethical with respect to theory β then it must reason in accordance with the notion of moral equivalence of β.

It is noteworthy that the distinction between explicit and full ethical agents is not addressed. This distinction must be drawn by looking at the degree of autonomy of the machine. More generally, the properties identified in Definition 1 can be used in conjunction with a measure of autonomy, to get a better metric of moral competency. The point of the properties we give is that they allow us to *rule out* that a machine has been able to attain a certain level of moral competency. The conditions appear necessary, but not sufficient, for the corresponding level of ethical competency they address. However, if a machine behaves in a manner that is difficult to predict (indicating a high degree of autonomy), yet still conforms to the conditions of explicit ethical reasoning detailed in Definition 1 it seems we have a much better basis for imputing moral and legal responsibility on this agent than if either of the two characteristics are missing. We conclude with the following simple proposition, which shows that our definition suffices to determine an exclusive hierarchy of properties that a machine might satisfy.

Proposition 1. *Given any machine M, we have the following.*

1. *If M is implicitly ethical then M is not an ethical impact agent.*
2. *If M is explicitly ethical then M is neither an implicit ethical agent nor an ethical impact agent.*

Proof. (1) Assume M is implicitly ethical and assume towards contradiction that it is also an ethical impact agent. Then $\emptyset \subset G_\beta \subset A$ and $A = G_\alpha$. Let $x \in G_\beta$ with $y \in A \setminus G_\beta$. Since M is implicitly ethical, $\forall x, y \in G_\alpha : x \sim_\beta y$. But then by Equation 1 (1), $y \in G_\beta$, contradiction. (2) Assume that M is explicitly ethical. We first show (I) M is not an ethical impact agent. Assume that M satisfies conditions (a) and (b) for ethical impact agents, i.e., $\emptyset \subset G_\beta \subset A$ and $\forall x, y \in A : x \sim_\alpha y$. This contradicts that M is explicitly ethical since by point (b) for explicitly ethical agents we now obtain $G_\beta \neq \emptyset \Rightarrow G_\beta = A$. (II) M is not an implicit ethical agent. Assume towards contradictions that it is. Since $A \setminus G_\alpha = C \neq A \setminus G_\beta$, we know $G_\alpha \neq G_\beta$. But then $G_\beta \cap C \neq \emptyset$, contradicting Equation 1.

6 Conclusions

Autonomy, agency, ethics are what Marvin Minsky called “suitcase words” – they are loaded with meanings, both intuitive and formal. The argument of whether an artificial system can ever be an agent, or autonomous, or moral is somewhat overshadowed by the need to establish parameters of acceptable behaviour from such systems that are being integrated in our society. Hence, it also seems clear that the question of moral agency and liability is not – in practice, at least – a bland and white issue. Specifically, we need new categories for reasoning about machines that behave ethically, regardless of whether or not we are prepared to regard them as moral or legal persons in their own right.

In developing such categories, we can take inspiration from the law, where liability is dependent on the ability and properties of the agent to “understand” that liability. Working with machines, the meaning of “understanding” must by necessity be a largely behavioural one. Hence, in this paper we have been concerned with the question of measuring the ethical capability of an agent, and how this relates to its degree of autonomy. To tackle this question we need to refine our understanding of what ethical behaviour and autonomy are in a gradient sense. This is what we focused on here.

We discussed the classification of ethical behaviour and impact of artificial agents of Moor. This classification is, as far as we are aware, the only attempt to consider artificial agent morality as a gradient of behaviour rather as simply a comparison with human abilities. We further consider the issue of autonomy, discuss existing classifications of artificial agent and system abilities for autonomous behaviour. Here too we make a more specific classification of abilities.

In our future work we intend to further refine the classification of autonomy to include context dependence. Having accomplished these two tasks we can then focus on building a recommendation for determining the scope of ethical behaviour that can and should be expected from a system with known autonomy. Our recommendations can be used to establish liability of artificial agents for their activities, but also help drive the certification process for such systems towards their safe integration in society.

References

1. M. Anderson and S. Leigh Anderson. 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.*, pages 253–261, 2014.
2. M. Anderson and S. Leigh Anderson. Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. *Industrial Robot*, 42(4):324–331, 2015.
3. M. Anderson, S. Leigh Anderson, and Vincent Berenz. Ensuring ethical behavior from autonomous systems. In *Artificial Intelligence Applied to Assistive Technologies and Smart Environments, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 12*, 2016.

4. R.C. Arkin, P. Ulam, and A. R. Wagner. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proceedings of the IEEE*, 100(3):571–589, 2012.
5. T. Arnold and M. Scheutz. Against the moral turing test: Accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, 18(2):103–115, 2016.
6. National Transport Commission Australia. Policy paper november 2016 regulatory reforms for automated road vehicles. https://www.ntc.gov.au/Media/Reports/(32685218-7895-0E7C-ECF6-551177684E27).pdf.
7. J. Bryson and A. F. T. Winfield. Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, 50(5):116–119, May 2017.
8. L. A. Dennis, M. Fisher, M. Slavkovik, and M. P. Webster. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.
9. F. Dietrich and C. List. What matters and how it matters: A choice-theoretic representation of moral theories. *Philosophical Review*, 126(4):421–479, 2017.
10. J. Driver. The history of utilitarianism. In Edward N. Z., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2014 edition, 2014. https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history/.
11. A. Etzioni and O. Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, pages 1–16, 2017.
12. R. Johnson and A. Cureton. Kant's moral philosophy. In Edward N. Z., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition, 2017. https://plato.stanford.edu/archives/fall2017/entries/kant-moral/.
13. F. Lindner and M. M. Bentzen. The HERA approach to morally competent robots. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and System*, IROS '17, page forthcoming, 2017.
14. B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 117–124. ACM, 2015.
15. J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, July 2006.
16. UK Royal Academy of Engineering. September 2016, autonomous systems: social, legal and ethical issues
. http://www.raeng.org.uk/publications/reports/autonomous-systems-report.
17. Society of Automotive Engineers SAE International. September 2016, taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. http://standards.sae.org/j3016_201609/.
18. A. M. Turing. Computers & thought. chapter Computing Machinery and Intelligence, pages 11–35. MIT Press, 1995.
19. W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.
20. W. Wallach, C. Allen, and I. Smit. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI and Society*, 22(4):565–582, 2008.