AI & SOCIETY
https://doi.org/10.1007/s00146-019-00891-1

OPEN FORUM

AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines

Scott Robbins[1]

Received: 16 October 2018 / Accepted: 4 April 2019

Abstract
With Artificial Intelligence (AI) entering our lives in novel ways—both known and unknown to us—there is both the enhancement of existing ethical issues associated with AI as well as the rise of new ethical issues. There is much focus on opening up the 'black box' of modern machine-learning algorithms to understand the reasoning behind their decisions—especially morally salient decisions. However, some applications of AI which are no doubt beneficial to society rely upon these black boxes. Rather than requiring algorithms to be transparent we should focus on constraining AI and those machines powered by AI within microenvironments—both physical and virtual—which allow these machines to realize their function whilst preventing harm to humans. In the field of robotics this is called 'envelopment'. However, to put an 'envelope' around AI-powered machines we need to know some basic things about them which we are often in the dark about. The properties we need to know are the: training data, inputs, functions, outputs, and boundaries. This knowledge is a necessary first step towards the envelopment of AI-powered machines. It is only with this knowledge that we can responsibly regulate, use, and live in a world populated by these machines.

Keywords AI ethics · Machine ethics · Meaningful human control · Robot ethics

1 Introduction

Artificial intelligence (AI) and robotics are increasingly entering our lives—from smart assistants in the home, social robots in the hospital, to algorithms delivering our news. There is no shortage of proposals for algorithms and robots in the future to take on novel roles—from AI-powered sex robots (Sharkey et al. 2017) to AI therapy bots (Gaggioli 2017). If implemented responsibly, these algorithms will no doubt contribute to society in a positive way. However, each of these applications raises concerns over the possibility to create unique ethical issues and/or exacerbate existing ones. Therapy bots and sex robots are both, for example, being placed in moral roles. Some have argued that machines like these require moral reasoning capabilities to navigate the ethical dilemmas they are sure to face (Wallach and Allen 2010; Scheutz 2016). This raises issues regarding the moral status of the machine as well as an issue assigning moral responsibility when bad outcomes occur (Johnson 2006; Bryson 2010; van Wynsberghe and Robbins 2018). A problem society currently faces is one in which we do not have ethical norms, regulation, or policy guidelines to assist developers in the careful balance between harnessing the power of AI while at the same time avoiding negative ethical and societal impacts. The first step to solving this problem, however, requires closure of an epistemic gap, i.e., society does not know for sure what these algorithms do or how they were created. Before we can create sound regulation and policy to guide AI development, there must be made available specific knowledge of the products and services powered by AI algorithms. It is the aim of this article to start us down a path that will lead us out of the epistemic darkness with regard to AI-powered machines.

Much of the focus surrounding AI ethics has been on the opacity of how AI algorithms reach decisions. It is not currently possible to know the reasons for a particular decision reached by an AI algorithm.[1] In some cases (e.g., playing

✉ Scott Robbins
scott@scottrobbins.org

[1] Ethics and Philosophy of Technology, Faculty of Technology, Policy, and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

[1] Here I am discussing those AI algorithms falling under the umbrella of 'machine learning'. There is work to try and overcome this opacity (see e.g., Wachter et al. 2018; Gilpin et al. 2018); how-

chess) this may be a perfectly acceptable situation; however, if the algorithm is deciding whether someone will get a loan or not it is unacceptable. One deserves an explanation of how the decision to decline a particular loan application was reached because the consequences can result in harm to the person whose loan was denied. This harm may not be physical; however, their lives will be significantly impacted due to the decision. There are many cases, though, when it would be counterproductive to require such an explanation—we use AI in some cases because it works differently than humans. Given this, requiring it to give human reasoning for its decision may undermine its effectiveness. For example, while the classification of a mole as malignant is an important in that it "significantly affects him or her" (Vollmer 2018), the AI algorithm which makes such a classification works well precisely because it does not use human articulable reasons for its classification. Why is it that in some cases algorithms with opaque reasoning are acceptable, while in others not?

This tension surrounding algorithmic opacity described above is the inspiration for this paper. I argue that opaque algorithms are acceptable when they are enveloped.[2] The central idea of envelopment is that machines are successful when they are inside an 'envelope'. This envelop constrains the system in a manner of speaking, allowing it to achieve a desired output given limited capacities. However, to create an envelope for any given AI-powered machine we must have some basic knowledge of that machine—knowledge that we often lack.

The knowledge that we need to create such envelopes are knowledge of the: inputs, outputs, function, boundaries, and training data of the AI. In the case of the mole classification AI, we know about the: inputs (pictures of moles), the training data (lots of pictures of moles), the function (to classify moles), and the outputs (malignant or not malignant). We do not know how it decides to classify the moles, but with all these other knowledge, an explanation is not needed. Even when we know very little about many of these aspects, it can be acceptable to use given that we constrain the AI appropriately. To take a really simple example, if the outputs of an AI-powered machine are rotating blades and swinging hammers and we are ignorant about what its inputs are, how it decides what to do, and what training data it was given it would be simple to figure out that this machine is only acceptable in very limited circumstances. For example, the show Robot Wars could use such a machine within the confines of a protected arena to fight against other robots.

The importance of this knowledge becomes especially salient when the outputs of an AI-powered machine have the potential to be harmful. Here, harm is to be understood not only as physical harm, but also harms such as invasions of privacy, financial harms, and restrictions on autonomy. It is also important to note here that these harms are understood as a result of the AI—not the companies irresponsibly collecting data on users of their products. In theory, an AI digital assistant like the Amazon Echo could operate without Amazon violating users' privacy. The focus here is restricted to those harms that are possible due to the functioning of the AI (both intended and unintended). When harms like this are present, we must know as much as we can about the properties highlighted above to make informed choices regarding usage, implementation, and regulation of these machines.

This paper begins by going into more detail on the subject of opacity as it relates to applications of AI. Following from this is a discussion of the concept of envelopment as it offers what I argue to be a better solution to AI's opacity problem. This is because many features outside of the inner workings of the algorithm remain opaque to us as well. I argue that enveloped AI will help us regulate, use, and be bystanders to AI-powered machines without the need for so-called 'explainable' AI. I include users and bystanders because regulation is one part of an overall picture which will guide the responsible introduction of AI-powered machines into society. The people implementing and using these machines must do so responsibly and the people who are being processed by or are bystanders[3] to these machines must be able to navigate this AI augmented world ethically. Section 4 delves into the properties that I argue are needed to properly envelope any given machine. Before concluding, I briefly respond to some possible objections and limitations of the proper envelopment of AI-powered machines.

2 Opacity and algorithms

There is much discussion about a lack of transparency when it comes to algorithms. Frank Pasquale argues that we live in a 'black box' society (Citron and Pasquale 2014; Pasquale 2015). Decisions are made by algorithms which affect many facets of our lives. Many of the stories in the media regarding contemporary AI are about algorithms which fall under the umbrella of machine learning. Machine-learning

Footnote 1 (continued)

ever, nothing so far can give us the specific reasons used to make a particular decision.

[2] The term 'envelopment' comes from the robot ethics literature. See e.g., Luciano Floridi (2011a) for a discussion of envelopment in which it is argued that envelopment describes the conditions under which robots would be successful.

[3] Bystanders to AI-powered machines are those people who are forced to use engage with them in some way. For example, people biking to work may come across an autonomous car and not know how to act around it.

algorithms use statistics and probability to 'learn' from large datasets. The complexity of the statistics involved and what those statistics refer to 'learn' has led to a situation in which we do not know how these algorithms make the decisions they make.

This can be quite disconcerting—and probably unethical in many circumstances. A decision about who gets a loan or not or what length of sentences are given to convicted criminals seems to require reasons. The same can be said about decisions regarding who is placed on the No-Fly list (Robbins and Henschke 2017). Finding out you are on the No-Fly list or were denied a loan without explanation is arbitrary and unacceptable. The European Union's recent General Data Protection Regulation (GDPR) legislation has been interpreted to include a "right to explanation":

The right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her.[4]

Although some disagree that such a right can be derived from GDPR (Wachter et al. 2016), the debate is illustrative of the desire for such a right. While I am not opposed to such a right, I find that it has focused the discussion on how to open up the black box of machine-learning algorithms rather than simply bar such algorithms from making such decisions. The question is "how can these algorithms explain themselves" rather than "what decisions are acceptable for an opaque algorithm to make?". The difficulty in answering this second question can be found in our ignorance with regard to the basics of many AI-powered machines.

This is because we are in the dark regarding AI. By 'we' I mean consumers, policymakers, lawyers, and academics. By 'in the dark' I mean that we have a general lack of knowledge and understanding about the technology. Take the recent example of the Amazon Echo. In May 2018, a woman reported that an Amazon Echo recorded a private conversation between her and her husband and sent it to one of her husband's employees—all without their knowledge (Chokshi 2018). While it is still unclear exactly how this occurred, Amazon's explanation is disconcerting:

As the woman, identified only as Danielle, chatted away with her husband, the device's virtual assistant, Alexa, mistakenly heard a series of requests and commands to send the recording as a voice message to one of the husband's employees (Chokshi 2018).

This explanation of the event offers other consumers who have purchased Amazon's echo devices with little information regarding how to prevent this from happening to them as well. Consumers (and Amazon) do not understand the combination of sounds that served as inputs into this AI-powered device. Consumers do not know what its boundaries are with regard to what it can do. Being an internet-connected device with access to your files, contacts, emails, documents, etc., it seems that there are no virtual boundaries for this device. Consumers also do not know what the functions of this device are. It is presented as an assistant with unlimited capabilities and its slogan is "just ask". Its outputs include: turning on lights, providing information, reading bedtime stories, ordering products, sending emails, chatting, ordering an Uber, etc. There are lists online detailing what possible outputs there are (Martin and Priest 2017) which have to be updated as the software updates.

AI-powered machines can also have deadly results—as can be seen by the recent autonomous car crashes of Uber and Tesla. It is unknown what combination of inputs resulted in, for example, a Tesla slamming into the road barrier resulting in the passenger's death (Levin 2018). While we cannot control our environment (e.g., a drunk driver might slam into your car), AI-powered machines are the first example of us not being able to control the tools we use to navigate our environment. No utilitarian calculus can change the disturbing idea that your autonomous car, for reasons unknown, may slam into a pedestrian or barrier. Without basic knowledge surrounding these machines how are users supposed to use them in an ethical manner? How are bystanders supposed to appropriately navigate a world filled with these machines? Finally, how are governments supposed to craft effective policies and regulation for these machines?

One major problem with focusing on explanation as a fix for the opaque inner workings of AI-powered machines is that many of these machines are beneficial because they are not articulable in human language. A cancer detection algorithm which cannot explain why one mole is labeled as cancerous should not be considered a problem if it is more effective than dermatologists.[5] Due to the fact that many of the benefits of AI-powered machines come from inherently opaque inner workings we must zoom out as we did in this section to see the other opacities surrounding AI-powered machines. When seen from this perspective a better solution to this problem is needed. The solution, I argue, can be found in a concept borrowed from the robotics field: envelopment.

[4] GDPR Recital 71. The full text can be found at https://gdpr-info.eu/recitals/no-71/.

[5] Other ethical concerns may, however, be raised for this application, e.g., concerns regarding the appropriate training data when algorithms are proven to work far better on fair skin than on darker skin tones (Lashbrook 2018).

3 Envelopment

Luciano Floridi has claimed that robots will be successful when "we envelop microenvironments around simple robots to fit and exploit at best their limited capacities and still deliver the desired output" (Floridi 2011a, p. 113). The term 'envelop' is borrowed from the field of robotics. The 'envelope' of a robot is the "three-dimensional space that defines the boundaries that a robot can reach" (Floridi 2011b, p. 228). Luciano Floridi has discussed envelopment as a process which allows for robots and AI to be more effective. He provides the striking example of washing dishes. Dishwashers are effective because they have been properly enveloped within an environment conducive to its operations (a closed box we call a dishwasher). The alternative is a humanoid robot which would be decidedly ineffective with regard to washing dishes.

Using Floridi's dishwashing robot as an example, we can see two broad sets of issues with regard to non-enveloped robotics and AI. First, the humanoid robot would constantly face novel scenarios (i.e., its inputs are not precisely defined and constrained) in which it would have to make judgments which could result in harm. I would consider myself deeply harmed were such a robot to scrub my new Le Creuset nonstick skillet with an abrasive brush. Add in mistaking a tablet computer for a plate and we can see a few of the many complex decisions such a humanoid robot would encounter. Furthermore, this robot would have to share its environment with humans. This increases the potential for ethical dilemmas and harm to humans.

Second, the task for the robot is ill-defined. "Wash dishes" is not precise enough. This could mean finding dirty dishes throughout a household, washing and drying those dishes, and, putting them away. Giving a robot this umbrella task, one could easily envision further tasks which would need to be added on: notifying a human that the soap is running out, sweeping broken glass, etc. Human users of such a robot may justifiably expect the robot to do things it simply is unable to do. These expectations could be mitigated if the robot's boundaries and functions were explicitly defined. These two sets of issues (harmful judgments and undefined task) should not occur in robotic and AI systems.

Floridi also proposes that driverless vehicles will only enjoy success if envelopment happens for them:

If drones or driverless vehicles can move around with decreasing troubles, this is not because productive AI has finally arrived, but because the "around" they need to negotiate has become increasingly suitable to reproductive AI and its limited capacities.(Floridi, 2011a, p. 228).

The limits of driverless cars in a non-enveloped environment have been shown in dramatic fashion. In April 2018, one passenger and one pedestrian were killed by cars operated by artificial intelligence in separate incidents. To date, the focus on driverless cars has been to increase their 'intelligence' by self-learning algorithms and to increase the effectiveness and capabilities of their sensors. The missing ingredient, according to Floridi, is envelopment.

But envelopment with cars becomes increasingly difficult. First, we would need to increasingly make the roads and their surroundings machine readable. Rather than relying on image recognition AI to 'see' that a stop sign is coming up, sensors could be built into the road which the car is easily able to read. This prevents a stop sign from being missed by the car's cameras due to a mud splattered sign or heavy fog. The effectively enveloped environment for a driverless car would be one which closes out all unexpected variables. Pedestrians and cyclists would not be allowed on the road, all cars would be driverless (human drivers are unpredictable), and all the road signs, dotted lines, solid lines, etc., would emit signals for the driverless cars to read. Truly enveloped driverless cars would not be able to leave the enveloped zone. This is because its inputs outside of an enveloped zone are potentially anything going on near automobile infrastructure. Ignorance about the possible inputs has led to fatal crashes. There is little advice given on where and when these cars should be used in autonomous mode. Are they only intended for recently built infrastructure on sunny, clear days? We really do not know. Tesla does not make a claim about what context autonomous cars should be used in—they simply say that the human operator should have their hands on the wheel in case they need to take over.

Envelopment would solve a lot of problems; however, as Floridi notes, this raises the possibility that the world becomes a place that reduces our autonomy in that we will have created a world in which we are forced to adapt to the envelopment needed by machines. Floridi is concerned with ensuring that this process of envelopment occurs with our foresight and guidance to prevent a world which works well with robots and AI but is not desirable to human beings. People who cannot afford new driverless cars would be forced to cope with a crumbling infrastructure for their non-driverless cars as more and more resources are used for the infrastructure serving as the envelope for driverless cars. The privacy concerned may be put at risk because the world has been changed such that AI-driven machines rely on sensors implanted into human beings—sensors which the privacy concerned refuse rendering them invisible. Although I argue in this paper that we should envelop AI-driven machines, it is important to note that the envelopes themselves may be unethical. This is why we must know what the envelop would have to be before we thrust these machines into society. This gives us a chance to say that the required envelope would not be worth it.

While Floridi uses envelopment to describe the conditions under which AI-powered machines will be successful, I argue that envelopment describes the conditions under which AI-powered machines should be considered acceptable. The example of driverless cars shows the potential harm which can occur when operating non-enveloped AI-powered machines. While we may not know how the algorithm results in a particular action, decision, or output, we should know enough about the possible inputs and outputs to know under what conditions a particular AI system should be used. Some basic knowledge about the machine helps us to make its envelope—preventing harm whilst helping the machine reach its full potential. If the envelope is too difficult to create (e.g., driverless cars), then the machine in question would be unethical to implement. To say otherwise creates a dilemma. On the first horn lies the situation in which we do not know enough about the machine to create such an envelope and, therefore, cannot prevent harmful situations (e.g., digital assistants). On the second horn lies the situation in which we know that this machine will lead to harm in the context that it is placed in but it is too costly or implausible to build the required envelope (e.g., autonomous cars). Both horns should be unacceptable to regulators, users, and bystanders.

To achieve the envelopment of any one AI-powered machine requires a level of knowledge about the machine that we often lack. To be clear, knowledge alone does not prevent bad things from happening. Knowing that a machine is capable of an output that causes serious bodily harm should prevent us from putting it into contexts where that output would cause serious bodily harm. This is how knowledge is connected to solving the diverse ethical issues that will arise when using AI-powered machines. Before we have this knowledge, we will not know if regulation, policy, ethical norms, or an outright ban will be the path to the responsible development, implementation, and use of AI-powered machines. Just like we do not put chainsaws into daycare centers, we should not put trash compacting robots in places where babies are sleeping. This knowledge will allow us to envelope AI-powered machines. Only then can these machines be considered to be under meaningful human control (Santoni de Sio and van den Hoven 2018)—control that is needed to responsibly regulate, use, and be a bystander to such machines.

The knowledge that we are lacking not only refers to how the machine works, but the what, why, and where. The "what" refers to the training data, possible inputs, and possible outputs. The "why" refers to what the machine is intended to be used for, i.e., its function. And the "where" refers to what boundaries constrain this machine. There are simply too many unknowns with regard to some AI-powered machines to regulate and use them. Many products now powered by AI are like the Monolith in Stanley Kubrick's 2001: a Space Odyssey in that their purpose, capabilities, and inputs are a complete mystery. The point is that we are in no epistemic position to create ethical norms, enact policy and regulation, or engage with these AI-powered machines until we shine a light on these important properties.

4 Towards the envelopment of AI

If we are to make responsible decisions about regulating and using AI-powered machines we need to know a lot more about them than we often do. This is especially true for modern AI algorithms (e.g., deep learning) which are opaque with regard to their reasoning. The training data, inputs, outputs, function, and boundaries of these machines must be known to us.

4.1 Training data

The data used to train machine-learning algorithms are extremely important with regard to how that algorithm or machine will work. Two algorithms that share the exact same code could work wildly differently because they were trained using different datasets. A facial recognition system trained only using pictures of faces of old white men will not work very well for young black women. If someone is to buy a facial recognition algorithm then there should be some information about the faces used to train it. The number of faces and the breakdown of age, ethnicity, sex, etc., would be a basic start. The specifics regarding what information is needed about the training data will obviously vary depending on context and type of data.

The knowledge regarding training data will be important when implementing algorithms. Simply knowing that the training data lack a certain demographic would hopefully cause one to test the system before using it on such a demographic or to restrict its use to demographics covered by the training data. For example, algorithms made to detect skin cancer were trained on images of moles mostly from fair-skinned patients—meaning the algorithm does poorly with regard to darker skinned patients (Lashbrook 2018). Whatever the reasons for this biased training data, it is important to know this before such an algorithm is used on a dark-skinned patient.

Knowledge of training data can also help to determine unacceptable algorithms which will simply reinforce societal stereotypes (Koepke 2016; Ensign et al. 2017). Predictive policing algorithms which rely upon training data that is biased against African Americans simply should not be used. The knowledge of this bias would not lead to its envelopment; rather, it should, if possible, lead to fixing the training data.

4.2 Boundaries and inputs

The terms 'boundaries' are construed broadly. Not only does it mean physical boundaries in the case of a robot, but also virtual boundaries which refer to the possible inputs (or types of input) in the form of data that it could encounter. 'Boundaries', then, refers to an algorithm's or robot's expected scenarios. For example, AlphaGo expects as an input a GO board with a configuration of white and black pieces. AlphaGo is not expected to be able to suggest a chess move based on an input of a chess board with a configuration of pawns, knights, bishops, rooks, queens, and kings on it. An algorithm playing chess is fine, but is a different algorithm than AlphaGo.

Knowing precisely what the boundaries a machine is constrained by helps us know what the possible inputs are. For example, a Roomba vacuum will have the boundaries of one floor of a home or apartment. A user is given a limited space with which to make sure that the robot will function properly. We can imagine a seeing eye robot which is given the task of guiding the blind when they go outside of the home. Now we have a machine whose boundaries are potentially limitless. It would be impossible to know all the possible situations the machine could face. In other words, the inputs to the machine are limitless. With the Roomba, however, one can survey the floor and detect possible problem inputs—the human has the information needed to envelop the machine.

Boundaries are different from inputs. A machine's inputs are determined by its sensors or code. The seeing eye robot above may have cameras, microphones, and haptic sensors all serving as inputs into the machine. An 'input' as I want to talk about it here is the combined data from all sensors. We, as humans, make decisions based on a number of factors. For example, we might put on a rain jacket because: it is raining, it is not too cold outside (otherwise we would opt for a heavy jacket), and we are going to be outside. A machine might be able to tell a user to wear a rain jacket based on the same data because it has a temperature sensor to sense how cold it is outside, a data feed from a weather website (to 'sense' that it is raining), and a microphone to hear the user say they need to go outside. It is the combination of these data which determines what output will be given.

Therefore, we not only need to know what types of inputs there are (sound, image, temperature, specific voice commands, data feeds, etc.), but how these get combined to form one input. There are machines which take very limited inputs which make very important classifications. The machine capable of detecting cancerous moles can only accept an image of a mole as an input. We have a very clear understanding of the inputs of this machine. On the other hand, a driverless car has many sensors which combine to provide infinite combinations of inputs.

I do not mean to suggest that a machine which can accept infinite combinations of inputs should not be used. We simply must know that this is the situation. We may know that an AI app on our phone accepts data from weather stations, our voice commands, images of our face, etc., as well as feedback after its decision (so that it can improve). Furthermore, it may not have any real boundaries—that is, it has the ability to grab data from other sources if it helps to improve its decisions. However, the function of the machine may simply be to decide whether or not to advise the user to wear a jacket. That is, it only has two outputs: jacket, or no jacket. We can debate about the overkill regarding using AI for advice on our outdoor clothing; however, the point is that a decision about the acceptability of a machine requires not only knowing its boundaries and inputs, but its function and outputs as well.

4.3 Functions and outputs

Knowledge of the functions and possible outputs of a machine is essential if we are to achieve the goal of enveloping AI-powered machines. In the AlphaGo example, the output is a legal move in the game of GO. We might be shocked by it making a particular move, but it is nonetheless a legal move in the game of GO. It would be strange if the function of AlphaGo were defined as "not letting an opposing player win" and instead of making a move its output was to mess up the board (because it knew there was no chance of winning and this was the only way to ensure that the other player did not win).

It can be easy to think that functions and outputs are equivalent. In the case of the jacket-deciding machine in the previous section, the function of the machine is to advise the user on whether or not to wear a jacket. This is the same as its output which is either "jacket" or "no jacket". This, however, is often not the case. The function of a driverless car is to drive from point A to point B; however, this will involve many outputs. Each turn, acceleration, swerve, and brake is an output. Defined functions are of the utmost importance because they allow us to test the machines for efficacy. How well a machine functions is clearly salient with regard to its moral acceptability. If the malignant mole-detecting algorithm was seldom successful at categorizing moles, then it would be unethical to use it. Equally unethical is the use of the algorithm when we are ignorant with regard to how successful it is (i.e., use outside of a testing environment).

Outputs are not the same as a machine's function; however, they can be discussed in the same way that we talk about a machine's capabilities. What can the machine do? A driverless car may be able to go 200 mph—which means that this is a possible output. A drone may have a machine gun built in, giving it the capability to shoot bullets—which means a possible output is the shooting of bullets.

This example makes it clear why it is so important to have knowledge regarding the functions, outputs, boundaries, and inputs. A machine whose possible output is to shoot bullets may be acceptable if its only input is a user telling it to shoot and its boundaries are a bulletproof room. We need all these knowledge to make informed decisions regarding the acceptability of machines.

4.4 Stepping out of the dark

Knowing what the inputs, boundaries, training data, outputs, and functions of an AI-powered machine will allow us to have some clue as to the envelopes these machines should be operating in. Even when machines are operated in environments which are so broad that we cannot prevent novel scenarios, the knowledge that this is the case helps inform our decisions regarding such a machine's acceptability. If there are possible novel environments (and, therefore, we are ignorant to the possible inputs), then the outputs must be such that it does not matter. No matter what novel board configuration of the game GO is given to AlphaGo, the output is always a legal move of GO. It is simply not possible for a harmful output. It would not matter if AlphaGo took as its inputs live CCTV video feeds from all over the world—the outputs would always be the same benign GO moves (although such inputs would probably not help with the stated goal of winning the game of GO). This is in direct contrast to the situation we face with driverless cars. Their possible inputs are states of affairs on just about any road in the world—with the weather, pedestrians, other cars, etc., all combining to create consistently novel inputs. In this case though, the outputs are potentially fatal.

Machines which have clear specifications regarding the properties listed in Sects. 4.1, 4.2 and 4.3 limit these problems. Cortis is an algorithm which detects voice patterns associated with cardiac arrest (Vincent 2018). The algorithm exists explicitly for the purposes of aiding emergency call operators (we know its function). The algorithm takes as its input live sound from the calling line. Its output is true if the voice pattern is associated with cardiac arrest and false if it is not (explicitly defined outputs). This algorithm being so explicit means that we have the knowledge to determine that this is an acceptable machine. If the machine is used within the boundaries given, then we can easily figure out what the possible scenarios are—without understanding how the machine comes to its decision. The machine either outputs true or false. If true, and a person on the end of the phone line is indeed having a heart attack, then the machine may be instrumental in preventing death. If the output is true, and no one on the end of the phone line is having a heart attack, then emergency services may be sent out without it being necessary. While this is not an ideal situation, knowing that it could occur gives us the knowledge to decide whether this risk is worth it. If the machine outputs false, and no one on the end of the line is having a heart attack, then the emergency call is unaffected by the machine. The last scenario is the machine outputting 'false' when someone on the line is having a heart attack. This is the worst scenario; however, the consequences of the machine acting this way are no different from the consequences of the emergency call without the machine. Again, the knowledge that this could happen is necessary for us to decide whether this is an acceptable risk.

We can imagine a machine which would operate in a similar context which could result in unacceptable risk—because we do not have the knowledge necessary to make an informed choice. This machine would be a robot which would be assigned the task of triaging incoming patients. The robot would be able to 'see' the patients, talk to them, and decide their place in line. The sheer number of possible inputs to this machine makes it difficult to determine how people could be harmed. In one obvious way, the machine could underestimate the seriousness of a person's situation resulting in their death. The possible harms are numerous and unpredictable. It could be that the machine results in less harm than when human beings are responsible for triaging; however, empirically validating this is next to impossible—especially before these machines are implemented.

If we are in the dark about the inputs, boundaries, functions, and outputs, then we have a machine we do not know enough about to properly envelop—leading to its possible failure which will often be an unacceptable risk to human beings. For, with modern AI, we are already in the dark about how it makes decisions. An undeveloped machine means that we are also in the dark about what could happen with these machines.

Ideally, AI-powered machines will be designed for envelopment—with clear ideas about the training data, inputs, functions, outputs, and boundaries. This knowledge would clearly be necessary to properly design for values or to facilitate an ethicist as part of the design team (van Wynsberghe and Robbins 2014). Not only would this result in ethically better designs but may prevent a waste of resources on a machine which cannot be enveloped and, therefore, may be designed to fail.

5 Objections

One objection could be that envelopment prevents the ultimate dream (or nightmare depending upon your perspective) of AI: developing general artificial intelligence. While some still dream of general artificial intelligence which will outperform humans at just about any task (Bostrom 2006; Müller and Bostrom 2016), the knowledge I am arguing for would explicitly exclude such a machine. General is the opposite of precise and general AI would be expected

to perform many different tasks, have a variety of outputs, and accept unlimited inputs. Luckily, this is not even on the horizon for robotics and AI right now, despite some futurists making bombastic and outlandish claims about this possibility. As Floridi puts it:

True AI is not logically impossible, but it is utterly implausible. We have no idea how we might begin to engineer it, not least because we have very little understanding of how our own brains and intelligence work. This means that we should not lose sleep over the possible appearance of some ultraintelligence. (Floridi 2016).

We should not be basing our ethical considerations and discussions around the possibility of general or strong AI. The focus should be on what is happening now and what could be happening in the foreseeable future. We must remember that robots just recently learned how to open a door—a capability that may be dependent upon specific door handles (Sulleyman 2018). We must not put a cart of ethical issues before the horse of the possibility of strong AI. It would be absurd to discuss the ethics surrounding eating unicorn meat when the foreseeable future does not include unicorns. The point is that discussion of general, super, or strong AI is a distraction from the real problems surrounding AI and robotics.

More pressing is the objection that requiring such knowledge would stifle innovation in AI. When Elon Musk claims that those opposing autonomous cars are "killing people" (McGoogan 2016) he is claiming that innovation in autonomous cars will save lives in the long run—so we should do it despite concerns. Envelopment would, to be sure, stop his Teslas from having the "autopilot" option. This function is not enveloped—and, therefore, we do not have the knowledge to make informed choices regarding its implementation and usage. However, envelopment would leave plenty of room for artificial intelligence to thrive.

The AI machines which are successful are the ones which are already enveloped. The Cortis machine for detecting cardiac arrest, AlphaGo, machines for analyzing X-rays (Litjens et al. 2017), spam filtering, fraud detection, etc., are all enveloped—and many of them are valuable with regard to helping us solve serious problems. Furthermore, we can measure how effective all these machines are. Most importantly, envelopment is a workaround for AI's transparency problem. If enveloped, AI machines can remain black boxes—therefore, ensuring that the benefits of AI are kept.

One objection which is difficult to resolve is that contemporary AI machines often have multiple algorithms at work to take inputs and create outputs. Just what is it that should be enveloped? That is, what is the machine? In a driverless car there are many sensors feeding into many algorithms which in turn feed their outputs to an algorithm which actually results in action. Taken as one machine, we might reach one evaluation; namely, that we lack the knowledge we need to envelope the machine. However, if we take this machine apart we may have many machines which are in fact enveloped. For example, if there was an algorithm which takes as its input an image of the inside of the car while it is in motion which outputs how many people are in the car the algorithm itself does not seem to have much problem. There are clear inputs, outputs, boundaries and a function.

However, human users of a driverless car experience the outputs of the car—the turns, accelerations, the braking, etc. Human users may not even be aware of the camera on the inside of the car—or the sensors detecting the outside world. The outputs of concern are the turns and accelerations of the car—not of the individual sensors. Therefore, the driverless car as a whole should be the object of evaluation.

Importantly, however, each of the machines which makes up the driverless car should be enveloped as well. What is different is the users of the machine. In this case, the user of the machine is the automaker. Each AI machine which makes up the driverless car should be enveloped—that is we should know their possible inputs, possible outputs, boundaries, and function. Not knowing these things about a machine of importance to the functioning of the driverless car would be unacceptable.

6 The limits of envelopment

It must be said that envelopment is not enough on its own. Although the function of the machine must be known to us, this paper says nothing of what functions should be assigned to robotics and AI Systems. It is easy to conceive of a robotic or AI system in which we have the knowledge I have argued we should require but is tasked with creating a superbug, or killing someone. What functions should be excluded from acceptable applications of AI is an important question. This question is actively debated in the field of robot ethics and the ethics of AI. Tasks that are deemed unethical for AI systems, therefore, should not be considered by developers, and attempting to envelop machines is a step that only applies to those machines whose functions are deemed ethical. Knowledge about these machines can help us with this, however. If the boundaries and function of the machine are forced to be made explicit, then it will be much easier to focus on whether or not this machine's function and context are acceptable.

The envelopment of a machine does not mean that a particular machine is effective. An enveloped machine may be spectacularly bad at achieving its function. This should definitely be a reason not to use a particular machine. What knowledge of the features described above can do for us with regard to efficacy is help us understand what success means for a particular machine. How do we judge the success of a machine when we do not know what its function is or the

boundaries of its operation? A machine which is precise with regard to its inputs, outputs, boundaries, function, and training data comes ready-made with a rubric for the evaluation of its efficacy.

A more general issue that this knowledge and ideal state of envelopment do not cover is the subtle changes technology can have on society. Just because we have the necessary knowledge for envelopment does not ensure that society will be changed for the better due to the technology. Guns serve as a good example here. We have good knowledge about how they work—their inputs and outputs. We can even say that there is meaningful human control in relation to guns. However, the option of using a gun opens up choices that were not available before. The ability to quickly and easily kill people has led, despite meaningful human control, to situations like the US, where too many people are harmed and killed. It would be better if such choices did not exist—and many countries have passed legislation taking away this choice.

The same will need to happen with regard to certain machines. Already there are calls to enact a ban on autonomous weapons (Sampler 2017). There may be many other machines which are unacceptable for their societal impact despite meaningful human control. Evaluations on societal impact—given envelopment and efficacy—will be extremely important. I do not pretend that the arguments in this paper help with such an evaluation; rather, they can prevent us from wasting time evaluating machines that have a more immediate problem: we do not have the knowledge to make informed evaluations in the first place.

7 Conclusion

The techno-optimism surrounding AI is running high. There seems to be no limit to its applications and no bounds to the hype in the media. It can be difficult, therefore, to separate out real hope from fantasy, the good ideas from the ridiculous, and the responsible from the irresponsible. Luciano Floridi has helpfully highlighted the concept of envelopment for helping us to understand what makes for successful robotics—and, as I have argued—for responsible robotics. To get to an enveloped state, however, we must know some basics with regard to these machines: the inputs, functions, training data, outputs, and boundaries.

Not only would such knowledge inform further ethical evaluation with regard to whether or not a specific function is an acceptable task for a machine, but it achieves a necessary condition for meaningful human control. Despite concerns about stifling innovation, envelopment allows for opaque algorithms to do what they do best. It simply keeps that opacity constrained to how the machine makes decisions. I argued here that opacity which spreads well beyond the 'how' of the machine and into the what, where, why, etc., is unacceptable. This allows us to realize the great things AI promises to us whilst keeping the fantastical, unnecessary, and dangerous machines out.

Envelopment is simply one part of the puzzle which, when solved, will result in creating AI-driven machines that will benefit and not harm society. Given envelopment, there are still important ethical evaluations which need to be made regarding the appropriateness of delegating a particular task to a machine, whether or not the operation of that machine is under meaningful human control, and what subtle societal effects such machines will have. While envelopment would not answer these important questions, it is a necessary and important first step towards the responsible design, development, and implementation of AI-powered machines.

Acknowledgements I would like to thank Aimee van Wynsberghe, Adam Henschke, and two anonymous reviewers for feedback on earlier versions of this article.

Funding The research benefited from the activities undertaken in the European Research Council advanced grant project "Collective Responsibility and Counterterrorism" awarded to Professor Seumas Miller.

References

Bostrom N (2006) How long before superintelligence? Linguist Philos Invest 5(1):11–30

Bryson J (2010) Robots should be slaves. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issues. John Benjamins Publishing, Amsterdam, pp 63–74

Chokshi N (2018) Is Alexa listening? Amazon echo sent out recording of couple's conversation. In: The New York Times. https://www.nytimes.com/2018/05/25/business/amazon-alexa-conversation-sharedecho.html. Accessed 28 May 2018

Citron DK, Pasquale FA (2014) The scored society: due process for automated predictions. Wash Law Rev 89:1

Ensign D, Friedler SA, Neville S et al (2017) Runaway feedback loops in predictive policing. arXiv 1706:09847 **(cs, stat)**

Floridi L (2011a) Enveloping the world: the constraining success of smart technologies. In: Mauger J (ed) CEPE 2011: ethics in interdisciplinary and intercultural relations. Milwaukee, Wisconsin, pp 111–116

Floridi L (2011b) Children of the fourth revolution. Philos Technol 24:227–232. https://doi.org/10.1007/s13347-011-0042-7

Floridi L (2016) True AI is both logically possible and utterly implausible—Luciano Floridi | Aeon Essays. In: Aeon. https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible. Accessed 20 Mar 2018

Gaggioli A (2017) Artificial intelligence: the future of cybertherapy? Cyberpsychol Behav Soc Netw 20:402–403. https://doi.org/10.1089/cyber.2017.29075.csi

Gilpin LH, Bau D, Yuan BZ, et al (2018) Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp 80–89

Johnson DG (2006) Computer systems: moral entities but not moral agents. Ethics Inf Technol 8:195–204. https://doi.org/10.1007/s10676-006-9111-5

Koepke L (2016) Predictive policing isn't about the future. In: Slate. http://www.slate.com/articles/technology/future_tense/2016/11/predictive_policing_is_too_dependent_on_historical_data.html. Accessed 22 Nov 2016

Lashbrook A (2018) AI-driven dermatology could leave dark-skinned patients behind. In: The Atlantic. https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/. Accessed 3 Oct 2018

Levin S (2018) Tesla fatal crash: "autopilot" mode sped up car before driver killed, report finds. In: The Guardian. https://www.theguardian.com/technology/2018/jun/07/tesla-fatal-crash-silicon-valley-autopilot-mode-report. Accessed 16 Oct 2018

Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005

Martin T, Priest D (2017) The complete list of Alexa commands so far. In: CNET. https://www.cnet.com/how-to/amazon-echo-the-complete-list-of-alexa-commands/. Accessed 28 May 2018

McGoogan C (2016) "You're killing people": Elon Musk attacks critics of self-driving cars. In: The Telegraph. https://www.telegraph.co.uk/technology/2016/10/20/youre-killing-people-elon-musk-attacks-critics-of-self-driving-c/. Accessed 23 May 2018

Müller VC, Bostrom N (2016) Future progress in artificial intelligence: a survey of expert opinion. In: Müller VC (ed) Fundamental issues of artificial intelligence. Springer, Switzerland, pp 555–572

Pasquale F (2015) The Black box society: the secret algorithms that control money and information. Harvard University Press, Cambridge

Robbins S, Henschke A (2017) The value of transparency: bulk data and authoritarianism. Surveill Soc 15:582–589. https://doi.org/10.24908/ss.v15i3/4.6606

Sampler I (2017) Ban on killer robots urgently needed, say scientists. In: The Guardian. http://www.theguardian.com/science/2017/nov/13/ban-on-killer-robots-urgently-needed-say-scientists. Accessed 23 May 2018

Santoni de Sio F, van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Robot AI 5:15. https://doi.org/10.3389/frobt.2018.00015

Scheutz M (2016) The need for moral competency in autonomous agent architectures. In: Müller VC (ed) Fundamental issues of artificial intelligence. Springer, Switzerland, pp 515–525

Sharkey N, van Wynsberghe A, Robbins S, Hancock E (2017) Our sexual future with robots. In: Foundation for Responsible Robotics. https://responsiblerobotics.org/wp-content/uploads/2017/07/FRR-Consultation-Report-Our-Sexual-Future-with-robots_Final.pdf. Accessed 20 Feb 2019

Sulleyman A (2018) Boston dynamics robot dog opens door for another robot with no arms. In: The Independent. http://www.independent.co.uk/life-style/gadgets-and-tech/news/boston-dynamics-robot-dog-open-door-video-hold-open-no-arms-help-video-youtube-a8208096.html. Accessed 28 May 2018

van Wynsberghe A, Robbins S (2014) Ethicist as designer: a pragmatic approach to ethics in the lab. Sci Eng Ethics 20:947–961. https://doi.org/10.1007/s11948-013-9498-4

van Wynsberghe A, Robbins S (2018) Critiquing the reasons for making artificial moral agents. Sci Eng Ethics. https://doi.org/10.1007/s11948-018-0030-8

Vincent J (2018) AI that detects cardiac arrests during emergency calls will be tested across Europe this summer. In: The Verge. https://www.theverge.com/2018/4/25/17278994/ai-cardiac-arrest-corti-emergency-call-response. Accessed 23 May 2018

Vollmer N (2018) Recital 71 EU general data protection regulation (EU-GDPR). http://www.privacy-regulation.eu/en/recital-71-GDPR.htm. Accessed 16 Oct 2018

Wachter S, Mittelstadt B, Floridi L (2016) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Priv Law 7:76–99. https://doi.org/10.1093/idpl/ipx005

Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol 31:1–52

Wallach W, Allen C (2010) Moral machines: teaching robots right from wrong, 1st edn. Oxford University Press, New York