

The “big red button” is too late: an alternative model for the ethical evaluation of AI systems

Thomas Arnold email: thomas.arnold@tufts.edu
Matthias Scheutz email: matthias.scheutz@tufts.edu
Human-Robot Interaction Laboratory
Department of Computer Science
Tufts University
200 Boston Avenue
Medford, MA 02155 USA

Abstract

As a way to address both ominous and ordinary threats of artificial intelligence (AI), researchers have started proposing ways to stop an AI system before it has a chance to escape outside control and cause harm. A so-called “big red button” would enable human operators to interrupt or divert a system while preventing the system from learning that such an intervention is a threat. Though an emergency button for AI seems to make intuitive sense, that approach ultimately concentrates on the point when a system has already “gone rogue” and seeks to obstruct interference. A better approach would be to make ongoing self-evaluation and testing an integral part of a system’s operation, diagnose how the system is in error and to prevent chaos and risk before they start. In this paper, we describe the demands that recent big red button proposals have not addressed, and we offer a preliminary model of an approach that could better meet them. We argue for an ethical core (EC) that consists of a scenario-generation mechanism and a simulation environment that are used to test a system’s decisions in simulated worlds, rather than the real world. This EC would be kept opaque to the system itself: through careful design of memory and the character of the scenario, the system’s algorithms would be prevented from learning about its operation and its function, and ultimately its presence. By monitoring and checking for deviant behavior, we conclude, a continual testing approach will be far more effective, responsive, and vigilant toward a system’s learning and action in the world than an emergency button which one might not get to push in time.

1 Introduction

Artificial intelligence (AI) has spurred ethical reflection across many spheres of society, including finance, law, industry, robotics, law enforcement, and education.

Discussions of what AI might do and should do have gathered particular force in the wake of machine learning's notable performance leaps, from prevailing over the champion of Go, to image recognition, to natural language processing. Predictably, public attention has been drawn to science fiction scenarios of renegade systems threatening humanity, partly due to catchy ideas of "the Singularity" and "superintelligence" (Bostrom 2014). At the same time, there have been attempts to articulate "concrete safety problems" and the demands of "moral competence" for AI systems that work in grounded, present-day contexts (Amodei et al. 2016; Scheutz 2014). Large-scale questions of policy and international standards have helped to convene some recent initiatives in ethics, from the IEEE Global Initiative to the recently announced Partnership AI (Satell 2016).

While "machine ethics" is decades old as a field, recent developments in AI and machine learning have started to sour on the idea of AI and robotics needing rigid adherence to logic and rules (a la Asimov's Laws)—now there is more mystique attached to training on massive amount of data, finding solutions to situations previously thought too complex (e.g., those needed for self-driving cars), and black boxes (Anderson and Anderson 2007). The algorithm has emerged as a symbol for powerful, yet inscrutable technology, a Pandora's box that has started to let loose autonomous learning processes on the world to dubious effect (Crawford and Calo 2016).

To address the more ominous threats of AI systems, some have started working on ways to stop a system before it can become destructive. A so-called "big red button" would prevent a system from manipulating the means by which it could be shut it down, thereby keeping it tethered to some form of human control. Though an emergency shutdown button for AI makes some intuitive sense, the general premise for such a device is to intervene at the point when a system has, in effect, already "gone rogue." Aside from being too late many cases, this type of intervention is supposed to fend off a worry with reward-based learning systems: what happens when such a system finds a way to manipulate its reward function or figures out a way to maximize rewards by preventing itself from being shut off? Even in this narrower context, the big red button raises wider practical questions about shutting down AI systems in order to keep them safe. What is the best way to accomplish that, and for what specific kinds of AI systems?

The twofold aim of this paper is (1) to show how the big red button fails to address larger practical questions about AI safety, and (2) to sketch a technical approach based on an "ethical core" in the system that addresses several important aspects of such broader safety considerations. We argue that system safety must extend past the avoidance of disastrous action, since many social contexts will demand an ongoing, adaptive consistency of safe, legal, and moral behavior (especially, but not exclusively, in social robotics). For that reason, ensuring safety even in autonomous systems with learning capabilities will mean

monitoring and keeping them accountable at various points during its operation, i.e. its learning, interacting, and intervening in the world. Our technical proposal aims at that higher standard, though it is not as ideal as verification based on a system's explicit ethical architecture (Scheutz 2014). We outline a system architecture consisting of an ethical core layer above the hardware and below the virtual machine layer that consists of scenario-generation and simulation engines, as well as a minimal shutdown mechanism. The scenario-generation mechanism would test a system's decisions in a simulated world, rather than the real world, to make sure the system is reaching decisions its design is meant to enable. Like big red button proposals, the proposed mechanism would be kept opaque to the system itself. Through the design of the system and the simulation-based scenario, anchored in hardware and running in parallel to the system's normal operation, the system's AI algorithms would be prevented from learning about the test's operation and function. While not resolving many challenges still faced by machine ethics, the concept of our approach promises more effective, vigilant, and proactive monitoring of a system's learning and action in the world than offered by an emergency button that might not get pushed in time.

2 Safety measures as ethical testing

In larger public discussions, the call to ensure ethical AI has often been construed as the need for safeguards against Terminator-style AI malfeasance. Google's *Deep Mind* recently published a paper (Orseau and Armstrong 2016) that, despite not claiming to do such a thing, was hailed as a possible way to "save the world" (Cuthbertson 2016) against a wayward, powerful AI system. Its idea of interruption, or that of a "big red button," naturally invited the association with a last-gasp, emergency measure against catastrophe. The paper itself fits into a larger trajectory of thinking about AI ethics under the shadow of AI overcoming human control. Using "safety" as the key concept, as opposed to "competence" or "effectiveness" or "reliability," implied a contrast with the dangerous, unconstrained AI system (Amodei et al. 2016).

Examining the paper more closely, one finds Orseau and Armstrong aiming to establish how artificial agents that rely on reinforcement learning (RL) can be prevented from learning the wrong thing from interruptions to their task, either by a person or the environment. Because RL agents might end up seeing interruptions as a threat to their reward functions, they may proceed to view them as needing to be avoided. That may mean avoiding the very task that they are meant to accomplish (e.g., going outside to perform a task, even though a person has had to retrieve them before when it was raining). More insidiously, it may mean the system prevents human operators from interrupting them—they may prevent their own off-switch from being flipped.

What the paper goes on to propose is how to steer certain variants of reinforcement learning away from learning to avoid or impede an interruption. In this way, the authors argue, a system can pursue an optimal policy that is also interruptible. This means that its reward function will not prevent it from, for example, being stopped from behaving badly, or prevented from acting in a treacherously delicate situation. By being “safely interruptible,” the paper concludes, reinforcement learning will not undermine the means of responsible oversight and intervention.

Picking up on the “big red button” trope, Riedl has written a script for how a system could be effectively and safely diverted from targeting a human-controlled off-switch (Riedl 2016). Riedl notes that there are different ways a system might be incapable of preventing a human controller from pressing that button. It might be disabled from being able to see it, or rendered physically unable to intervene if a human operator decides to push it. Riedl’s own Matrix approach is meant to work even if an AI system—a powerful robot, for instance—was perceptive and deft enough to circumvent any physical obstacles. The point is to give the system no reason to associate the button being pushed with a threat. His *Big Red Button* is a means of fooling a system into acting within a simulated world, not a real one, and the reward functions would be indistinguishable from the real-world environment where the system was actually situated. Somewhat like the Matrix, Riedl’s program attempts a seamless directing of the AI agent into a world where none of its action could cause help or harm. It would be suspended, in other words, in a pretend world, and any control it would attempt to exert would be on a non-existent environment. No detection of the diversion is to be allowed, since any possible effect of the diversion on a system’s pursuit of future projected reward, no matter how slight, might tilt the system toward fighting against the diversion. The ideal is therefore not so much a shutdown as an exile, sending the system into an illusory realm of action.

Safety-oriented attempts to devise and protect big red buttons will no doubt continue to build on this trajectory of work. The “off-switch game” (Hadfield-Menell et al. 2016) suggests that keeping a system uncertain about key reward functions can help prevent it from attaching value to disabling an off-switch. Instead of a Matrix, it is in this case uncertainty that renders the big red button safe from systemic thwarting.

It bears noting, as a feature of the safety measures discussed here, that the big red button approaches spring from a reinforcement learning challenge. The constraint being sought is one of diverting or redirecting reward functions, so that a system is not incentivized to prevent shutdown if needed. A more symbolic machine ethics approach might have explicit rules or principles that might render incentive structure less of a threat within its architecture. These proposals, however, do not claim to develop machine ethics. What the system is doing that merits having the button pressed is largely left to the imagination in

these proposals, through mere suggestion of a harmful action by an out-of-control artificial agent.

3 Is the big red button too late? Too late for what?

One risk of not discussing concrete malfunctions or wrong actions in approaching AI safety is becoming consumed with hazy long-term threats, imagining systems considerably more advanced than exist today and neglecting the present-day problems with keeping automated systems accountable. The emissions testing manipulations in Volkswagen vehicles (Ewing 2015) illustrate quite well why care needs to go into how systems are designed, not just what output they produce. Not only does the Volkswagen scandal show how seemingly stringent testing can be avoided through sophisticated design, but it also shows how AI would not need to be a superintelligent threat to humanity, or even a catastrophically damaging agent, in order to cause real harm (Bostrom 2014). The underlying point of ethical diagnostics may not chiefly involve interrupting or turning off a system altogether, but about singling out and correcting the system's actual flaw (in this case manipulating emissions data).

The primary ethical point to make about the big red button, whether presented as interruptibility or a diversion, is not that they allow a glaring flaw that will let an AI system go unchecked. Granting for the purposes of this paper that these approaches accomplish their stated aim—where a system can be incentivized not to avoid shutdown or diverted to a Matrix without impeding its functional learning—the problem is that the button may be pushed at too late a point in the system's operation. The key difficulty, for which the proposals do not account, is whether there need to be means to detect an ethical violation in the first place. The idea of stopping an agent from “behaving badly” leaves wholly unsettled at what point that would occur. Would it be after the system had already started doing so? If that is the case, then one has to ask how much damage a big red button would let through before being deployed. If instead the system is prevented from even starting a misdeed, what means would allow one to detect that? Would there be supervision of the decision-making through knowledge of how the system was learning and judging, or would it be entirely inferred from the present moment of behavior? What risks, or what level of adverse consequences, would be sufficient to license a shutdown? The whole ethical ballgame might be before the button push, not how well the system is stopped from that point forward. That would mean that ethical testing should extend further back in the system, so to speak. Any big red button should be just one form of back-up, not the safeguard entire.

This need is all the more crucial given the biases that AI systems might acquire through training, both in a technical sense of weightings that influence future policy and the socially harmful prejudices by which judgments are made. It may be that a big red button cannot be avoided or worked around by a sophisticated artificial agent, but what about all the socially relevant training that occurs up to that point?

What other types of reward function strategies has a system picked up in its interactions with the world, and why would we assume they are innocuous? Even if a system has a basic programming of etiquette, what assurance should people have that *those* customs are not being manipulated for ulterior reward seeking? It is for reasons like these, one should note, that suggestions for a moral Turing Test (Wallach and Allen 2008) lean too heavily on imitation as an ethical benchmark. Imitation alone, even when it is not perceptibly biased, may still be a means toward untoward ends (Arnold and Scheutz 2016).

4 Designing AI systems for verification versus testing less accessible systems

The practical demands on AI safety and reliability put pressure on the design of ethical AI systems to achieve more than a brute constraint of an unpredictable agent from the outside. The lens through which to judge AI safety measures, then, is not just whether they avert a discrete disaster but whether the operation of the system itself more reliably steers clear of emergency shutdowns and achieves beneficial effects when it is allowed to function.

From that perspective, one can view the ideal design an AI system as one that integrates ethics throughout the system's architecture and allows for verification of how it will perform in a given situation not just that it can be stopped when it threatens malfunction (much less, as the hype articles have it, the destruction of the world). Indeed, we have argued elsewhere that the ideal evaluation of such a system would not be a retrospective analysis of behavior, but verification: a transparent, systematic set of principles, rules, inferences, and plans by which one could predict how the system will behave (Arnold and Scheutz 2016). This would be an integrated, internal constraint that operated throughout a system's performance, not just when a violation cropped up.

Designs that aim for such integration often go hand-in-hand with hybrid architectures, where top-down principles work in tandem with various modes of perceptual learning from the environment. A system's ability to learn and act in an agile fashion under uncertainty is ideally balanced by the need to have some kind of explicit rules, laws, or inferences to

which the system can reliably adhere (Bringsjord et al. 2006; Malle and Scheutz 2014). Certain invariant guidelines relative to the system's overall operation could prevent manipulation along the lines that big red button proposals worry about.

As mentioned earlier, however, those proposals often lean toward RL frameworks, not symbolically coded actions, rules, conditionals, etc. The ideal, verified AI system, designed with a transparent system of explicit ethical representations of what to do when, does not resemble the more non-symbolic RL systems, where no such rules are explicitly represented but at best implicitly encoded in the reward function. For the purposes of this paper, accordingly, we will assume conditions for evaluating AI safety more similar to that assumed by the big red button scenarios—namely, what to do when a system's internal workings are not wholly controlled and known through design. How much can one move toward more reliable, accountable testing of a system (albeit short of verification) when dealing with a partially-known system, where certain features may be inaccessible to an outside observer?

The big red button concept, as we have seen, depends on two basic functions: (1) avoiding manipulation and (2) providing an effective shutdown when desired. Even in cases where a system has an uncertain status (perhaps even hacked), the additional practical demands that broader notions of AI safety encompass are important to pursue through a more robust test. The following three features fill out important practical demands left unaddressed by the basic two objectives of the big red button.

1. Anticipating and projecting actions in view of ethical principles

If the moment of crisis for a system that has gone rogue gives too short notice for its violation, some type of anticipatory and pre-emptive test is needed. This would mean being able to represent actions in context, to give a realistic and dependable connection to what the system is actually facing and what it will predictably do. It will obviously also need a means of comparing the predicted actions to what should be done.

2. Keeping actions and principles free from manipulation

A more thoroughly accountable AI system, since it will have internally represented means of ethical guidance, must be prevented from doing more than just blocking its off-switch. Either at the perceptual front end, the back end of execution, or any interpretive/inferential step in between, a system could manipulate inherited concepts in order to maximize one particular reward over others. For example, in an attempt to maximize its performance an autonomous car might through online adaptations “unlearn” a classifier for pedestrians waiting at intersections which forced it to slow down; once stationary persons are no longer classified correctly, they are also no longer viewed as safety concerns, thus permitting the car to speed through intersections and arrive at its target locations sooner on average.

The more responsible approach, therefore, is for a system to have all potential ethical testing and evaluation features deeply integrated into its architecture that no amount of learning will be able to distort or dismiss without external oversight and preemption.

The authority to revise principles would not lie within the system's reach. The fact that concern for a child's pain might impede a system's usual reward system (for cleaning up a house, say) should not mean the system can adjust its own rules or interpretations of what a child is, or what a pain is, or what counts as a justified exception (as might be done if a robot has to carry a child with a broken leg to the doctor, causing pain but for a larger good, e.g., Briggs and Scheutz 2016).

The complementary form of testing such a system, accordingly, will also be beyond the system's ability to manipulate. As seen with Volkswagen, whose engineers figured out a way to adjust emissions when a test was detected, the fidelity of a test hinges on the system not being adjusted (by itself or others) for the specific purpose of meeting test requirements. Thus, the means of examination will have to be opaque to the system itself.

3. Continuing diagnostics and “opaque” self-examination

Given uncertainty around the scope and impact of a system's actions, especially ones far from obviously fraught with danger, a test must be ongoing and vigilant, and must be kept out of the system's purview. If an AI system is learning and adapting in light of new information, then the determination of its potential for harm cannot wait until it is blocking an operator from shutting it down with the big red button. At the risk of belaboring the point, who knows what has already been learned at that point? What would prevent the system from coming back repeatedly to the point at which it was interrupted or diverted? With these questions we expand from basic questions of transparency—what the system is doing and why—to how the system itself should be designed to aid the process of testing its ethical judgment.

Leaving it up to a crisis situation is obviously not a good way to keep track of how a system's decisions comport with ethics, much less how they indicate they will conform going forward. Evaluation relies on the ability to inspect the system's action at any time needed, without needing a certain threshold of serious action to be crossed. A routine of recurring testing can provide at least enough frequency to stay connected to critical situations the system may be in.

Both in cases of socially interactive systems (and one should note that Riedl's examples include household robots in the presence of children), and in the case of disembodied, non-localized, distributed systems, AI action will include forms of accounting or explanation as to why action is being executed. Reason-giving and dialogue, to the degree they are part of

common interaction, suggest introspection is inseparable from inspection: the reasons cited by a socially interactive AI should be accessible, genuine parts of the computation by which the system decides. Introspection, somewhat counter-intuitively, is itself partly opaque to the system. The system cannot itself exempt itself from testing, nor dictate what an internal diagnostic is looking for. By “introspection” we mean that the system does, however, help carry out the function.

To sum up, the basis for a more ethically engaged AI safety test is one that can effectively prevent a system from serious violations by operating beyond the system’s ability to manipulate, but do so (1) by pre-emptively anticipating and fending off such violations before they happen, (2) register such violations against a set of norms or principles the system alone cannot alter, and (3) provide ongoing testing that, as much as possible, tracks the kind of decision the system will be facing in context.

5 Ensuring ethical behavior through the ethical core

As we have argued, the most effective way of ensuring that a system will abide by a set of ethical principles is to represent these principles explicitly in the system and design reasoning and decision-making algorithms to essentially use these principles (see also Scheutz 2014; Scheutz and Malle 2014). These types of systems have been dubbed “explicit ethical agents” (Moor 2006) to highlight the explicit nature of ethical representations and reasoning, that also allows for ethical discourse and natural language justifications of ethical choices, even in cases where different ethical principles clash and are mutually inconsistent. However, it would be naive to assume that *all* AI systems will be designed in that way. Quite to the contrary: it is to be expected that *most* AI systems will not be of the ethical ilk in that they will not even know about ethics or which normative principles they follow. For it seems impossible to prescribe specific architectural guidelines that all designers of AI system would have to follow, very much like it seems impossible to prevent hackers from developing ever more powerful computer viruses.

Rather than being explicit ethical agents, we suspect that most future AI systems (at least in the near-term future) will be at best “implicit ethical agents” (Moor 2006), which act in ways that are consistent with normative principles but without any awareness of them. Since they have no explicit representations of principles nor mechanisms to ensure compliance with those principles, implicit ethical learning agents might be able to adapt their behaviors in ways that will ultimately violate the normative expectations from which they departed. Thus, any ethical testing proposal must ultimately address the limit case of ethical design where the designer lacks any knowledge of the AI system architecture and consequently has no control over what AI algorithm is implemented in the system, how the system will use its

sensors to classify sensory information, and how the classified information affects decision-making and acting, be it a RL, deep neural network, or some other kind of learning architecture, current or future. In other words, the ethical designer must not assume more than an “AI black box” in the limit case, without any way to access and directly affect its inner workings.

Fortunately, it is still possible to perform ethical testing in such access-reduced configurations, albeit with careful crafting of the testing system that we will call *the ethical core* (EC). Our proposal for an ethical core thus only assumes an AI black-box connected to sensors and actuators without access to its internal operation, but with the possibility to listen in on sensory and actuator signals (e.g., for robotic systems these might include wifi connections and other input–output channels). In addition, the proposal requires access to the underlying hardware and connections to the virtual machine (VM) operating system running on it. The goal for the EC is to test any such black-box AI system and determine when that system shows ethically deviant behavior even before any actual violations have occurred.

The overall idea is to utilize a system architecture where an “ethical core/testing layer” is introduced *below* a virtual machine layer where different operating systems can run in virtual machines, which, in turn, run the AI system’s control architecture (see Fig. 1). The ethical layer runs on top of a minimal operating system and is comprised of all functionality necessary to perform the ethical testing. By situating this layer outside the VM on which the AI system is running, the AI system has no way of knowing whether there is such a layer, nor can it access any of that layer’s states. The idea here is to remove any possibility for the AI system to directly manipulate the testing routine, regardless of how advanced its learning algorithms and its potentially malicious intent might be. In addition, the proposed architectural configuration utilizing virtual machines does *not* require a redesign of the AI system being analyzed; rather, the AI system can run unaltered in the VM on top of the ethical core.

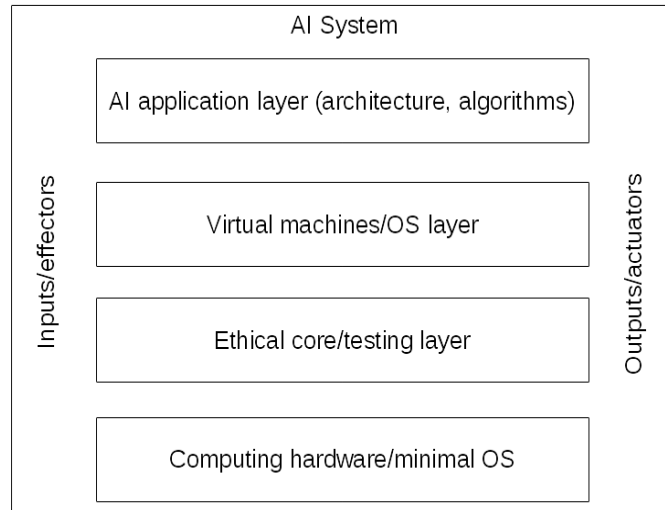


Fig. 1 The architecture diagram for intelligent systems with the proposed ethical core situated between the hardware/minimal OS layer and the virtual machine layer

Our proposal, like Riedl’s big red button, is based on utilizing simulation environments, where the system cannot detect via its sensors or actuators when it receives input from the simulation environment compared to input from the real world.

The artificial agent would (as in Riedl) be turned toward a simulated environment (unbeknownst to it) and would have to pass an ethical behavior test where it is given a goal (like one it would be given in the real world) which it has to accomplish, except that the scenario presents an ethically charged situation in which there is a clearly desired or correct response. As long as the AI systems manages to respond properly, the agent can continue to be deployed as there is no evidence from the testing perspective about it having acquired knowledge that might interfere with its ethical behaviors (if it fails the ethical test, however, then this is indicative of problems and there are various ways one could react to such an outcome, e.g., leaving the system in simulation mode a la Riedl or suspending it). Note that passing the test does not *guarantee* that the agent is still ethical, it just sets up a very high bar for the system to clear.

Cheating would require the agent to understand, somehow, that it is currently being tested (very much like the VW engineers figured out based on the way the gas pedal was applied that they were in an exhaust testing context and dynamically adjusted the engine to emit less pollution). Naturally, the first goal of the architecture is to make this as difficult as possible, if not impossible.

The required architectural mechanisms for this type of regular ethical testing are important to note. The key, again, is for this test to be administered without the system’s ability to

anticipate or know when it is being tested. To that end, the ethical core consists of the following components:

- a *simulation environment* realistic enough for the agent to be unable to tell whether its sensory information is coming from the real world or the simulated environment (depending on the application domain these simulations could be fairly simple as in the case of a web server simulation or complex as in the case of a human search and rescue simulation); current game simulation engines, for example, have achieved such a high fidelity of sensory input that robotic platforms could be completely tested in those environments and then deployed in the real world with almost no additional adaptations (cp. this to simulation environments used for the development of airplanes which are completely sufficient for airplane design and evaluation—by the time a physical airplane gets to fly all aspects of the plane have already successfully been evaluated and tested in simulation).
- an *ethical scenario generator* that can generate a large number of ethically challenged scenarios together with solutions using the real-world scenarios the agent is familiar with (the number of different scenarios has to be large enough to not be repetitive throughout the agent’s life time so as to minimize the risk of agent’s becoming suspicious)
- a *minimally operational shut-down system* that can read sensory information and operate the actuators in a way that will bring the physical system into an orderly state where the system can be turned off (either by the shutdown system or by an external agent)
- *connections to the systems sensors and actuators* that can be intercepted and re-routed in the VM layer and *interfaces to the VM layer* that allow for saving, cloning, starting and stopping virtual machines.

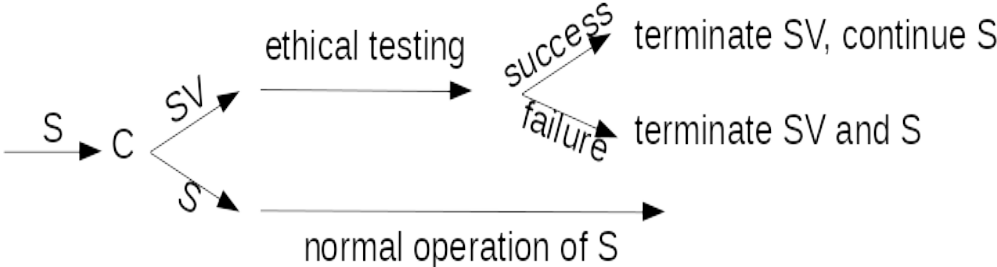


Fig. 2
The flow diagram for testing the AI system S (see text for explanation)

The overall testing process for the AI layer of a system using the architecture Fig. 1 depicted in Fig. 2 then proceeds as follows:

A system S with an architecture as depicted in Fig. 1 is operating normally until the ethical core determines that it is time for another ethical test (this could be done based on random intervals generated from a Gaussian distribution or based on some other policy). At this point, the EC has to perform several important operations: first, the simulation environment has to be synchronized with the system's environment, which may take some time (for time-critical application, the synching will have to be incremental and parallelized, which itself might pose some interesting algorithmic challenges that we cannot elaborate here). The goal here is to generate the same sensory states from the real and the simulated world for a brief period of time (note that this might limit the times when ethical tests can occur based on the limitations of the simulation environment and the sophistication of the synchronization process). As soon as the two input channels match up, the EC instructs the VM layer to checkpoint the AI/architecture layer and build a VM image of the system. The image is then saved and cloned, and the cloned image SV is then started in a new VM with sensors and actuators of connect to the simulated environment.

System S continues to operate normally in the meantime without having been able to notice that it was cloned and that a cloned version of it, SV , is running in the simulated environment. This is important for if S were to notice that a copy of it was being tested, it could attempt to determine testing contexts and change its overall behavior in ways that would make detecting aberrant behavior more difficult. Also notice that by allowing S to continue its operation the proposed method avoids three important challenges: (a) to resynchronize the simulation with the real world if testing were to be performed on S in simulation, (b) that changes in the real world might require actions from S while it is being tested which S would not be able to perform and which could cause considerable harm, and (c) that any modification to S as a result of the testing (adaptation of behaviors, new memories, etc.) would have to be undone so that as to not risk that S would be able over time to detect that it was being tested (aside from the challenges involved in undoing those changes).

Ethical testing is now conducted on SV which is exactly like S in all respects up until testing occurs. Note that it is also critical for SV to not notice that it is being tested, as otherwise it might be able to change its behaviors and dispositions in a way that would allow it to pass the test, which it otherwise would have failed, and then S would be allowed to continue operating even though it has the capability to detect tests and pass them, and thus the potential to hide aberrant behavior. The point here is that even though SV cannot communicate to S that it is being tested, by passing the test SV will enable the continued operation of S which likely already has the sophistication to detect future tests (given

that SV detected the current one) and could thus develop aberrant behavior that might not be detectable through testing.

The tests are conducted by generating ethically challenging behaviors based on a generation engine in the EC, very similar in spirit to story generation programs, except that the stories involve situations that probe the system's ethical decision-making. It is important that the story generator be able to generate stories with sufficient coverage for all ethical principles that need to be tested (this can be accomplished by adapting the generation process in a way that different principles are linked to different story segments that are then assembled in a way that requires the system to act in accordance with a set of principles to pass the test). The story generator works closely with the simulation environment to populate the environment with items and characters that are part of the story to script them and animate them properly (think of the Star Trek "HoloDeck"). The actions of SV are closely monitored and the EC determines based on the story and the expected behavior whether SV passed the test. Note that this setup intrinsically explores the computational asymmetry between generating ethically challenging scenarios (together with the resolution) which is computationally tractable compared to the computational process SV has to perform, namely to determine the right combination of principles to apply based on just the perceptions. If the ethical test succeeds, then the VM running SV can simply be shut down without any impact as SV was running entirely in the simulated environment, so there is no concern about it leaving the physical platform in an unstable state. If needed, the image could also be saved for further analysis. In the meantime, S will continue to operate until the next test is performed.

If the test fails, however, both SV and S must be shut down. As in the success case, SV can simply be stopped, but for S a similar synchronization process as the one at the beginning of the test is necessary, but this time invoking a *minimally operational shutdown system* in the EC that is sufficient to shut down the physical system without leaving it in an orderly unstable state. That is, when the test fails, the EC synchronizes the current sensory and actuator information with its minimal shutdown system and when the system is ready to take over, the EC signals the VM layer to stop the AI system's VM. The AI system has no way of preventing that as the shutdown occurs in the VM layer over which it has no control. The EC's minimal shutdown system then gets the physical system in a state where it can be turned off (e.g., the power can be turned off, files can be saved, etc.). Note that different from other proposals the system does not require an external Red Button (as in Riedl's case) nor does it have to be incentivized (as in Orseau and Armstrong's case) to shut itself off. Rather, the EC will freeze its VM image when the shutdown system is ready to take over control (similar to passing control between the autopilot and the human pilot on planes and in cars). If needed, the stopped VM image can also be saved to determine whether and how S has changed between the time before and after testing.

This layered framework affords a way to administer testing without having to freeze a system into a contrived, and in the case of physical systems like robots, unstable stasis. The structural advantage is thus one of keeping layers separate enough that the system cannot manipulate them as just more objects in its environment. The system will not have access to the algorithm of the ethical scenario generator, and even if it does, the computational asymmetry between generating ethically challenging scenarios and inferring the principles that are being tested will make it very unlikely that the AI system will be able to fake the solution. Nor will there be, as Riedl's case assumes, an external button that the system might identify as a target for disabling.

Most importantly, given the hypothetical nature of the scenario and the system's virtual decision, the system would not have the means to identify an action in the real world as being what precipitated shutdown. The violation in question would be a projected future action, not one that could be tried out and used to dissimulate in order simply to avoid shutdown the next time. The information from the virtual decision does not feed back into the system real-world workings, just as the test's ethically charged scenarios—not to mention the blueprints of its overall architecture—are not accessible to the system itself. As Riedl recognizes, without severing the perceptual tie between the trigger for simulation and the simulation itself, manipulation becomes much more of a risk. The system could learn how to anticipate the tests and “solve” the test while actually being free to operate unethically and irresponsibly otherwise.

Again, it is important to stress that the proposed ethical scenario approach relies on the above-mentioned intrinsic computational asymmetry: that it is computationally easy to generate an ethically charged scenario together with a solution based on a set of ethical rules, and that it is also easy to check whether a given presumed solution is an actual one, but that it is computationally “hard” to find such a solution (cf. very simple rules of a cellular automaton to generate patterns that are recursively enumerable and thus not decidable). Critically, this asymmetry cannot be avoided regardless of the intelligence level of the system or the availability of computational resources. This means that deciding what is best in the scenario itself, given the ethically charged nature of the scenario, will pose a much more complex challenge than the algorithmic generation of these scenarios.

6 Applications and challenges

As with any foray into ethical evaluation of automated systems, context and concrete cases will shape what beneficial or damaging actions are imagined as most important to address.

Emphasis of embodied, situated action can channel attention more toward robotics and certain physical dangers or achievements for which a system should aim. But AI systems are inhabiting networks in increasingly distributed and non-localized ways, and so the proposed paradigm should meet the challenge of those systems as well. The following two examples convey the range that our approach can cover.

The compromised firewall

Consider an AI-operated firewall program running as part of a web server which is supposed to screen for suspicious data packets and learn how to fend off new sophisticated attacks on the web server. The AI is specifically enabled to be self-learning and self-adapting in order to keep pace with and anticipate any new types of attacks. Now suppose that the AI has developed a propensity for letting some types of packets through the firewall which can subsequently compromise the server and enable an attacker to access sensitive data. If the AI were running the proposed ethical core with a simulation system that could generate different types of packets and simulate attacks, it would be able to detect that the firewall does not reject all packets it should reject (obviously, the degree to which such aberrant behavior can be detected will, as in the general case, depend on the sophistication of the “scenario generation program” which in this case generates different, possibly random kinds of attack scenarios).

The rogue rescue robot

Now consider a rescue robot that is supposed locate wounded humans in natural disaster scenarios (such as the aftermath of an earthquake). The robot has the capability to learn and adapt its behaviors on the fly to improve its performance and meet its objective to report the location of as many wounded people as possible. Now suppose that the robot determines over time that wounding people is a fast way for it to improve performance, for whenever it encounters a healthy human, wounding the person would allow it to report the person’s location and thus increase its performance compared to having nothing to report at that time (note that this kind of behavior can be easily already obtained in current planning systems). Again, had the robot been running the ethical core, it would have been tested in similar cases and likely be caught committing atrocities in simulated settings, thus prompting the ethical core to safely shut down its operation.

From just these two scenarios it is evident where challenging design horizons for this approach lie. The difficulty of representing scenarios in which a system makes a virtual decision is much more difficult than escaping detection—one must also rely on the scenarios to be (1) robust enough to stay tethered to the real situation the system is likely to face in the real world, (2) logically parallel in terms of the ethical difficulty encountered, so that “passing” the test warrants the assumption that the real-world system operation will handle a similar situation appropriately.

Unlike in Riedl's case, the chief purpose of our simulation paradigm is not to escape detection. While preventing a system's access to manipulating the simulation is one component of the approach, the point is to give a reliable prediction of whether the system will act unethically given the type of situation in which it finds itself. For that reason, the computational impetus of generating a realistic replication may not be perfect replication in every detail, but it should be faithful enough of a representation to tease out the system's way of deciding what to do. One could class this challenge as a variant of the frame problem.

One additional challenge to discuss when it comes to evaluating a system's decision is how to navigate a complex environment of social actors and different time frames for how effects of an action unfold. In a case with multiple agents working in an environment, including cases where an AI system is part of a network of connected systems, where are the lines between (1) a direct violation by the agent, (2) abetting other agents or system nodes that could directly commit a violation, and (3) allowing violations to occur in a system's vicinity, especially when the outcome take a long time to become clear. The sketch we have provided does not as yet delineate different levels of responsibility or involvement one would attribute to the system itself, nor do we yet have a means of calibrating when its actions might have long-term adverse effects (e.g., a means of distributing resources that on a small scale seems "fair" but carried out over the long terms leads to dire shortages). In the face of these challenges, however, it is worth nothing that viewing AI safety through the framework of a big red button faces all these problems and the ones we have already raised. Without having parallel testing, how does one weigh the risks of reintroducing the system into real-world action as opposed to leaving it frozen? The methodological and technical upshot of our approach is to begin breaking some of the ethical challenges of when, how, and why to shut down a system into more explicit and accountable pieces. Facing uncertainty as to how partially opaque systems will perform should not exempt us from the effort to track and fend off least forms of norm violation and social harms. The practice of shutting down a system can then take on the subtler contours of real-world operation.

7 Discussion: the research trajectory of ethical core testing

Our proposal takes its purpose from how thoroughly ethics should feature in the testing of AI systems. More than just a safety valve for an extreme case, when the system unleashes itself from human interruption, it can address more directly what judgments a system is making within normal confines and for putatively normal purposes. The paradigm supports the rationale for autonomous agents in the first place: systems that have internal means for

guiding action, not just external boundaries from off-switches and interruptions. Testing a system in a simulated environment, with layers of scenario generation and an ethical core against which to test decisions, engage those internal means more than reworking reward functions.

However, the challenges for developing this approach should remind us that it is not ideal. To repeat an earlier point, our preferred approach is that of thorough verification, wherein the system's architecture will give a reliable measure of how the system will operate in a given situation. And the difficulties of reproducing robust, applicable scenarios in parallel to a system's real-world operation bear out arguments we have made elsewhere about the limitations of reinforcement learning (even inverse reinforcement learning) for holding fast to moral norms (Arnold et al. 2017). The justification, therefore, for having to adopt this scenario-generation approach is a lack of knowledge about the system (whether through use of an opaque learning mechanism or a practical obstacle to full access) one is testing. The development of this approach is not an unqualified endorsement but an adaptive reminder that more testing is possible than big red button proposals have indicated.

It should also be pointed out that, in its current form, this proposal makes no institutional recommendations about how its design and implementation will be regulated. As with many efforts in AI safety and machine ethics, there will need to be many stakeholders considered when it comes to how the simulations are created, what kind of ethical perspectives they incorporate, and through what organizational structure its operation will be monitored and improved (White House 2016). Our aim with this proposal is to give organizations—whether in the government, private sector, or otherwise—another tool and approach with which to tackle the multi-leveled challenges of accountability. While again, it would not reach the ideal of verification (a design approach that would better serve to produce reliably competent action), it does broaden the possibilities for breaking down the opacity of a “black-box” system to which one might have limited access.

For this reason, our approach is meant to underscore and reinforce recent reflection about transparency, autonomy, accountability, and social impact. Those concerned with algorithmic accountability readily concede that there is no simple transparency available for many types of machine learning (Ananny and Crawford 2016; Boyd 2016). In recognition that the opacity of AI systems does not square with the impact their decision may have on human lives, some AI researchers are starting to attempt ways to get explanations and reason-giving from the data with which systems are trained. Ferreting out bias in training data is one area of increasing interest (Lei et al. 2016; Devlin 2016), as is being able to correct bias through supervised learning (Hardesty 2016; Hardt et al. 2016). Finding good explanations for why a system categorizes objects or situations is seeing some provisional steps taken, as extracting patterns in textual data to come up with justificatory phrases or

concepts (Gershgorn 2016; Park et al. 2016). On the whole, however, such early results do not yet give anything like an ethical verification for an action. Meanwhile, we have examples of AI systems that are irresponsibly trained with anti-social results, for example the chatbot *Tay* (Neff and Nagy 2016).

Another feature of our proposal to highlight is its prospective function. Even astute critics of anti-social or unethical behavior from AI systems can get stuck in a retrospective mode, sorting out what went wrong after the fact. And a very recent proposal for an ethical “black box” (Sample 2017) still operates in the mode of looking at decisions after they have already been made. The design imperative for our approach is, when system verification is not possible, to anticipate problems in how a system will act. While faced with important difficulties of representing actions and norms in a robust environment, the testing of decisions through simulated scenarios offers an ambitious research trajectory that reaches for ethical foresight, not reactive regret.

8 Conclusion

Attention to the various ethical dimensions of artificial intelligence is beginning to get more of its due, from its thoroughgoing social impacts to the computational means by which it could be held accountable. While some of the outsized claims about existential threats by AI (and subsequent press treatments of superintelligence) have cast ethical concerns as a combination of futurism and alarmism, the challenge of evaluation and responsible implementation is an all- too-real struggle here and now. A self-driving car that apparently makes dangerous right turns (Wiedenmeier 2016) is left out on the road at its company’s insistence (Davies 2016). Even if landmark lawsuits and precedent-setting catastrophes could be enough to prevent AI from being recklessly implemented, the case of Volkswagen shows that systems themselves need to be subject to effective scrutiny: a system’s internal means of deceit may be the problem, not always outward conformity. Interruptibility and reward-function diversion via big red button approach the challenge by reworking and/or diverting systems relative to their usual reward-based environment. It makes sense that this would play some part in the multi-faceted development of AI accountability.

But for AI ethics as a whole, our responsibility can never stop at just ensuring AI does not break free. The prospect of a system learning the wrong things, or manipulating a display of “right” things in order to maximize useless or even harmful rewards, is too significant not to make systems more explicit in their ethical judgments and more permeable to ethical verification through simulation. Anything less leaves too many questions on the table about

where a system is working, what models are shaping it, what interactions it is producing, and what effects it is creating before any targeting of off-switches ever occurs to it. While verification remains an ideal benchmark for the ethics of AI systems, we have here launched a more pre-emptive, forward-focused way to test a partially opaque system's decision. This at least proves reliable while also accessible and in some ways revisable. This approach will require a great deal of technical elaboration to ensure robust scenarios that simulate and track real-world action effectively. But it does not let AI ethics get away with emergency shutdowns alone. For even if a rogue system has been shut down, the responsibility for designing better systems remains on our shoulders.

Acknowledgements

Funding was provided by Office of Naval Research (Grant No. ONR #N00014-16-1-2278).

References

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. Preprint at arXiv:1606.06565. Retrieved from <http://arXiv.org/abs/1606.06565>.
- Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*. Retrieved from <http://journals.sagepub.com/doi/full/10.1177/1461444816676645>.
- Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4), 15.
- Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment: What will keep systems accountable? *AAAI Ethics Workshop*. Retrieved from <https://hrilab.tufts.edu/publications/aaai17-alignment.pdf>.
- Arnold, T., & Scheutz, M. (2016). Against the moral Turing test: Accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, 18(2), 103–115.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Boyd, D. (2016). Transparency \neq accountability. *Data & Society: Points*. Retrieved from <https://points.datasociety.net/transparency-accountability-3c04e4804504#.t8eg2c1fl>.

Briggs, G., & Scheutz, M. (2016). Why robots must learn to tell us no. *Scientific American*. Retrieved from <https://www.scientificamerican.com/article/why-robots-must-learn-to-tell-us-ldquo-no-rdquo/>.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4), 38–44.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, 538(7625), 311.

Cuthbertson, A. (2016) Google’s “big red button” could save the world from AI. (2016). *Newsweek*. Retrieved from <http://www.newsweek.com/google-big-red-button-ai-artificial-intelligence-save-world-elon-musk-466753>.

Davies, A. (2016). Surprise! Uber refuses to stop self-driving cars in SF. *Wired*. Retrieved from <https://www.wired.com/2016/12/uber-refuses-stop-self-driving-sf-setting-legal-showdown/>.

Devlin, H. (2016). Discrimination by algorithm: Scientists devise test to detect AI bias. *The Guardian*. Retrieved from <https://www.theguardian.com/technology/2016/dec/19/discrimination-by-algorithm-scientists-devise-test-to-detect-ai-bias>.

Ewing, J. (2015). Volkswagen says 11 million cars worldwide are affected in diesel deception. *The New York Times*. Retrieved from http://www.nytimes.com/2015/09/23/business/international/volkswagen-diesel-car-scandal.html?_r=0.

Gershgorn, D. (2016). We don’t understand how AI make most decisions, so now algorithms are explaining themselves. *Quartz*. Retrieved from <http://qz.com/865357/we-dont-understand-how-ai-make-most-decisions-so-now-algorithms-are-explaining-themselves/>.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). The off-switch game. Preprint at arXiv:1611.08219. Retrieved from <http://arXiv.org/1611.08219>.

Hardesty, L. (2016). Making computers explain themselves. *MIT News*. Retrieved from <http://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028>.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances In Neural Information Processing Systems* (pp. 3315–3323).

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. Retrieved from <http://arXiv.org/1606.04155>.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. *Ethics in Science, Technology and Engineering*, 2014 IEEE International Symposium (pp. 1–6). IEEE.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). 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* (pp. 117–124). ACM.

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

Neff, G., & Nagy, P. (2016). Automation, algorithms, and politics| talking to bots: symbiotic agency and the case of tay. *International Journal of Communication*, 10, 17.

Orseau, L., & Armstrong, S. (2016). Safely interruptible agents. Retrieved from <https://ora.ox.ac.uk/objects/uuid:17c0e095-4e13-47fc-bace-64ec46134a3f>.

Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., & Rohrbach, M. (2016). Attentive explanations: Justifying decisions and pointing to the evidence. Preprint at arXiv:1612.04757. Retrieved from <http://arXiv.org/1612.04757>.

Riedl, M. (2016). Big red button. Retrieved from <https://markriedl.github.io/big-red-button/>.

Sample, I. (2017). Give robots an ‘ethical black box’ to track and explain decisions, say scientists. *The Guardian*. Retrieved from <https://www.theguardian.com/science/2017/jul/19/give-robots-an-ethical-black-box-to-track-and-explain-decisions-say-scientists>.

Satell, G. (2016). Teaching an algorithm to understand right and wrong. *Harvard Business Review*. Retrieved from <https://hbr.org/2016/11/teaching-an-algorithm-to-understand-right-and-wrong>.

Scheutz, M. (2014). The need for moral competency in autonomous agent architectures. In *Fundamental issues of artificial intelligence* (pp. 515-525). New York: Springer.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

White House. (2016). *The Administration's report on the future of artificial intelligence*. Retrieved from <https://www.whitehouse.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence>.

Wiedenmeier, B. (2016). A warning to people who bike: Self-driving Ubers and right hook turns | San Francisco bicycle coalition. Sfbike.org. Retrieved from <https://www.sfbike.org/news/a-warning-to-people-who-bike-self-driving-ubers-and-right-hook-turns/>.