AI and Ethics (2024) 4:463–471
https://doi.org/10.1007/s43681-023-00283-8

ORIGINAL RESEARCH

Designing robots that do no harm: understanding the challenges of Ethics *for* Robots

Brian Hutler[1] · Travis N. Rieder[2] · Debra J. H. Mathews[2,3] · David A. Handelman[4] · Ariel M. Greenberg[4]

Received: 24 February 2023 / Accepted: 28 March 2023 / Published online: 17 April 2023

Abstract
This article describes key challenges in creating an ethics "for" robots. Robot ethics is not only a matter of the effects caused by robotic systems or the uses to which they may be put, but also the ethical rules and principles that these systems ought to follow—what we call "Ethics for Robots." We suggest that the Principle of Nonmaleficence, or "do no harm," is one of the basic elements of an ethics for robots—especially robots that will be used in a healthcare setting. We argue, however, that the implementation of even this basic principle will raise significant challenges for robot designers. In addition to technical challenges, such as ensuring that robots are able to detect salient harms and dangers in the environment, designers will need to determine an appropriate sphere of responsibility for robots and to specify which of various types of harms must be avoided or prevented. These challenges are amplified by the fact that the robots we are currently able to design possess a form of semi-autonomy that differs from other more familiar semi-autonomous agents such as animals or young children. In short, robot designers must identify and overcome the key challenges of an ethics for robots before they may ethically utilize robots in practice.

Keywords Machine ethics · Moral-scene assessment · Nonmaleficence · Responsibility · Harm

1 Introduction

Robots and other artificial intelligence (AI) systems already occupy a significant role in our lives, with more applications on the horizon. Autonomous vehicles are on the roads, autonomous weapons systems have been deployed in combat, and learning algorithms are used to evaluate job applicants and to diagnose diseases [1]. The prevalence of robots and AI is likely to increase in the aftermath of the COVID-19 pandemic, as corporations and universities have developed AI systems to fight the pandemic, including supercomputers that are able to predict a pathogen's evolution [2], and robots that can disinfect hospitals [3]. Concerns about future infectious disease outbreaks may also prompt an increased use of automation in a variety of workplaces, such as agriculture, manufacturing, food service, and healthcare [4].

Our increased utilization of and reliance upon robots and AI raises numerous ethical, legal, social, and political challenges. This paper focuses on issues related to robots, especially (but not exclusively) those that are or could be used in healthcare settings, understood broadly to include robots designed to perform surgery, eldercare, childcare, emergency medical services, and search and rescue. We loosely define "robots" to be autonomous or semi-autonomous computer systems that have some ability to move around and to gather information from their environment, and to communicate with humans using verbal communication [5]. We will define "autonomy" in this context to be the ability to carry out tasks and achieve goals to some degree or in certain respects in novel contexts or environments, without complete external oversight or control [6]. "Autonomy" in this sense differs from "automation," which is the ability to perform a pre-specified task in a stable environment

✉ Brian Hutler
brian.hutler@temple.edu

[1] Department of Philosophy, Temple University, 1114 Polett Walk, Philadelphia, PA 19122, USA

[2] Berman Institute of Bioethics, Johns Hopkins University, 1809 Ashland Ave, Baltimore, MD 21205, USA

[3] Department of Genetic Medicine, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA

[4] Applied Physics Laboratory, Johns Hopkins University, 11100 Johns Hopkins Road, Laurel, MD 20723, USA

without oversight or control. Autonomy can be described as the ability to "make decisions," in the sense of the ability to plan courses of action and set proximate goals. Although mental and cognitive concepts like "plan" and "decision" are useful in describing autonomy, we are agnostic about whether autonomy requires "artificial agency," where agency is defined in terms of "intentional mental states" [7].

The use of autonomous robots promises potential benefits in the healthcare context. Autonomous surgical robots, for example, may be able to complete complex tasks less invasively and more efficiently than human surgeons working alone [8]. Likewise, elder care robots designed to function as home health aides can monitor vital signs, look for indications of problems, and serve an alert function, as well as provide direct assistance with, say, movement or medication reminders [9].

These potential uses of autonomous robots prompt a range of moral and ethical questions. The ethical questions about how humans ought to develop and use robots can be described as the Ethics *of* Robots [10]. These questions, while often extremely challenging, are similar to those that arise for other new technologies, in the sense that the target or object of our needed ethical guidelines are the human beings (and their institutions) that plan to use these new technologies. Associated with the Ethics of Robots, however, is a distinct set of ethical questions regarding the rules or principles that ought to guide or govern robots in their own decision-making and behavior when they operate autonomously, that is, when there is not a human providing oversight or feedback "in the loop" or "on the loop." We describe this other set of questions as the Ethics *for* Robots—i.e., what rules, principles, or guidelines we should program, teach, or otherwise instill into a robot's programming so that they make "ethical" decisions and engage in "ethical" behaviors. Isaac Asimov's widely cited "Three Laws of Robotics" are an example of an Ethics for Robots, and they illustrate the distinctive character of these ethical questions [11]. Contemporary theorists such as Susan Leigh and Michael Anderson have also done much to identify what is distinctive about Ethics for Robots, and to propose approaches for arriving at appropriate ethical principles [12].

The questions raised by Ethics for Robots differ from the ethics questions raised by, e.g., the development of nuclear weapons or gene editing techniques, because the target or object of the needed ethical guidelines is not solely for humans or their institutions, but also for the robots themselves. We need to develop these ethical guidelines even if robots are not now or in the near future moral agents analogous to adult human beings. There are many pressing questions about Ethics for Robots that are relevant given our current level of technological sophistication and the robots that are currently in production and in use. Engineers and theorists must take seriously the distinctive challenges raised by creating an ethics that is tailored to an entirely new kind of entity.

This paper argues that Ethics for Robots is not simply a matter of programming human morality into robots. The structure and content of human morality must be rethought and retheorized in order to cohere with the roles, functions, and abilities of currently available robots and those we are likely to produce in the near future. That is, we need to develop and tailor an ethics that is appropriate specifically for the robots that are being built in real life—not for fictional robots that possess capabilities such as general intelligence that are comparable to the full moral agency possessed by adult human beings [13]. In short, the normative content of an Ethics for Robots must be grounded in the actual abilities and limitations of the robots we currently (or will soon) create, while also accounting for the constraints of engineering and programming.

This paper aims to identify challenges that designers and programmers face in creating and implementing an Ethics for Robots that go beyond just the challenges of determining the appropriate ethical content or ethically theory. To illustrate these challenges, we will work from an uncontroversial starting point, namely, whatever the Ethics for Robots in the healthcare context might be—whatever moral rules and principles are needed to fill out its normative content—it will have as one core component the Principle of Nonmaleficence, i.e., the principle of "do no harm." Nonmaleficence will almost certainly not be the sole component of a fully developed Ethics for Robots, and it might not be the most important. But, as we will argue below, it is likely that nonmaleficence (or something like it) will be a component of Ethics for Robots in some form, especially for robots employed in the healthcare setting. We will then use the Principle of Nonmaleficence to illustrate a range of challenges—conceptual, moral, technical, and political—that must be addressed to successfully develop an Ethics for Robots. Although this article focuses on the Principle of Nonmaleficence and the healthcare context, our goal is to open up discussion of the range challenges that we anticipate may occur more generally as robots are developed with greater autonomy and increasing usefulness and incorporation into daily life.

2 The Principle of Nonmaleficence

We start by arguing that one aspect of Ethics for Robots is the Principle of Nonmaleficence or "do no harm." The Principle of Nonmaleficence says that a moral agent should avoid directly causing harm to humans (as well as other relevant creatures) within some specified scope of responsibility. The Principle of Nonmaleficence is importantly different from a Utilitarian calculation of maximum expected utility, because nonmaleficence prioritizes avoiding direct causation of harm. According to nonmaleficence, every individual human being must be treated as an independent entity, and no one individual may be harmed even to protect the interests of many others. For example, the Principle of Nonmaleficence would not allow killing one person in order to save the lives of five others.

The Principle of Nonmaleficence is a familiar feature of the principles of biomedical ethics that govern doctors and other medical professionals, reflected in the famous “first, do no harm” of the traditional Hippocratic Oath as well as many contemporary biomedical ethics treatises and textbooks [14]. Roughly speaking, the Principle of Nonmaleficence says that medical professionals should pursue the treatment option that is all-things-considered safest for their patient—meaning least likely to cause harm—even if other treatment options could provide a higher but more risky upside. Of course, decisions about medical treatment and care are often very difficult, requiring an appreciation of nuance and complexity and an application of professional judgment involving the weighing and balancing of competing principles and considerations [15]. That said, nonmaleficence is often taken to be a core tenet by practicing clinicians, having a special status over the other core tenets of biomedical ethics including beneficence, justice, and patient autonomy. On our view, nonmaleficence sets a baseline or floor of acceptable conduct for members of the medical profession.

The Principle of Nonmaleficence (as we will use the term) also distinguishes *doing* from *allowing*, or *action* from *inaction*, placing greater moral significance on avoiding directly causing harm through an active choice or behavior (i.e., “doing”) versus merely allowing harm to take place through inaction. In general, the Principle of Nonmaleficence prohibits directly or intentionally causing harm through action. But in some cases, the Principle of Nonmaleficence may be consistent with allowing harm to occur through inaction. Here we follow the traditional understanding of the Principle of Nonmaleficence in biomedical ethics, which distinguishes causing harm from “beneficence” or providing aid [12]. Both nonmaleficence and beneficence are important aspects of biomedical ethics, but they are importantly different and may at times be in conflict. For example, a surgeon may be faced with the decision whether to perform a risky surgery to attempt to save a patient’s limb—a surgery which if successful would certainly benefit the patient but which if unsuccessful may be harmful to them. The Principle of Nonmaleficence would suggest avoiding the surgery in such a case. Note that Asimov’s famous First Law of Robotics, which not only prohibits a robot from causing harm, but also from allowing a human to come to harm through inaction, tends to blur the distinction between nonmaleficence and beneficence. This difficulty will be addressed below, in our discussion of robot responsibility.

3 Nonmaleficence and robots in healthcare

The Ethics for Robots will likely include the Principle of Nonmaleficence for robots that are designed to perform medical or healthcare functions, such as robots employed to assist in surgery, to provide home health assistance for elderly patients, or to aid first responders in emergency situations.

In support of this idea, we will argue, first, that human medical professionals who partner with robots to assist them in performing their professional tasks must ensure that employing the robot will not conflict with their professional ethical obligations. That is, medical professionals have an obligation to ensure that employing a robot is consistent with their own ethical obligations. But, as robots are granted a greater share of autonomy in such a partnership, it is difficult to see how the human medical professionals could uphold their own obligations to their patients unless they know (or reasonably believe) that the robot is programmed to adhere to (or at least not to violate) those same obligations. As such, robots designed to work with medical professionals must be programmed to adhere to the same professional obligations that would be relevant for any task to which they may be assigned. Since nonmaleficence is the floor or baseline principle of biomedical ethics, robots employed in healthcare contexts must be designed and programmed to adhere to the Principle of Nonmaleficence.

The argument sketched in the previous paragraph appeals to what we will call the “Partnership Principle,” which says that a human may not partner with an autonomous robot to achieve a task or goal unless the human knows or reasonably believes that the robot will not violate the human’s own moral, ethical, or legal obligations in completing that task. That is, a human who partners with a robot must reasonably believe that the robot’s decisions and actions will be consistent with the moral, ethical, or legal standards according to which those decisions and actions would be evaluated if performed by the human. Or, to put it another way, a human cannot use an autonomous robot to offload or escape the moral, ethical, or legal responsibilities that are relevant to the achievement of a particular goal.

The Partnership Principle is by no means axiomatic, and it may not apply in all contexts. For example, it may not apply in cases where an autonomous robot cannot cause significant harm to persons or their property. Suppose Atticus has a Roomba that bumps into Betty’s toe. Atticus himself has an obligation not to kick Betty in the toe, but his Roomba has arguably not caused Atticus to fail to meet this obligation, at least in part because the physical harm at stake is minimal. If, however, Atticus has an autonomous vehicle (AV) that runs over Betty's foot, the matter may be very different. There is a plausible argument that Atticus's obligation to Betty as an operator of an automobile requires him to ensure that his AV adheres to the same obligations that he would himself be required to uphold. Accidents happen, of course, but typically the driver of a vehicle is responsible for accidents that they cause. According to the Partnership Principle, Atticus cannot plausibly say that the injury to Betty's foot is not his fault because his AV was programmed to adhere to different ethical principles. Currently, U.S. law is unresolved as to whether to hold the driver or the AV manufacturer liable in such a case [16]. But in our view, the driver should be responsible (at least morally) unless the driver could reasonably believe that the AV is programmed to abide by the same moral principles that would govern the driver's own operation of the vehicle—a demand which may set a very high bar in the context.

There are two general reasons to accept the Partnership Principle, one pragmatic and one moral. First, pragmatically, a mismatch of moral rules and principles between a human and a robot with whom they partner could lead to failures of coordination and communication. Kant's famous "murder at the door" example provides an apt illustration of the problems that could occur if the Partnership Principle were not satisfied. In Kant's example, a murderer appears at the door, looking for the owner of the house [17]. The servant, hoping to protect the homeowner, lies to the murderer saying that the homeowner is not home. Hearing the murdering coming, however, the homeowner slips out the backdoor, and is caught by the murderer behind the house. Although presented as an argument against lying, Kant's example could also be understood as an argument that people working together should adhere to the same set of moral or ethical rules, since a mismatch of rules can cause coordination problems even between people who share the same goal. A truth-telling servant working for a Utilitarian homeowner might tell the truth to the murderer, but if the homeowner incorrectly predicted the servant's behavior, the homeowner might still be caught. Similar coordination problems could result if a partner robot were programed to adhere to different moral or ethical rules than the human with whom it is partnered.

A second reason in support of the Partnership Principle is that people should not offload their responsibility for moral wrongs onto a subordinate or partner. The rule-breaking vigilante, committing moral wrongs in order to protect others, is a familiar trope [18]. But whatever we may think about vigilantes in general, it is problematic to intentionally avoid moral responsibility for a morally fraught task by delegating the task to someone who does not adhere to the same moral rules. Likewise, we should not design robots to do our moral dirty work for us, by programming them to adhere to a moral or ethical code that we would be unwilling to adopt as our own.

The Partnership Principle is especially plausible in the healthcare context, where medical professionals must perform a variety of tasks that could potentially cause harm to their human patients. If any such task is assigned to a robot partner, it is plausible that the actions and decisions of the robot are subject to the same ethical duties as would apply to the medical professional. Since the Principle of Nonmaleficence sets a baseline of ethical obligation for medical professionals, it follows that for any task that could cause morally salient harm to a patient, a medical professional may assign such a task to a robot only if the robot will not violate the Principle of Nonmaleficence that applies to the medical professional. And, if so, any use of autonomous robots in the healthcare context would require ensuring that these robots are designed and programmed so as not to violate the Principle of Nonmaleficence.

To be clear, we do not suggest that robots can now, or in the near future, hope to replicate or replace the decision-making capability of human medical professionals, who are tutored by experience and able to perceive the full array of considerations that are relevant to human moral reasoning. Instead, our position is only that the Ethics for Robots that applies to robots used in a healthcare context must ensure the ethical treatment of patients. Tasks that involve ethical complexities beyond what robots are capable of correctly navigating should be left to properly trained human professionals. Our concern here is for the ethical principles that must be programmed into robots who are designed to partner with medical professionals in a healthcare context.

In summary, we have argued that the Principle of Nonmaleficence is a necessary component of the normative content of Ethics for Robots that applies to at least a significant subset of robots, namely those employed in healthcare contexts and assigned to perform some medical tasks. Given that nonmaleficence is a component of Ethics for Robots, we will now identify and describe some of the challenges faced in creating robots that are able to adhere to this principle.

4 Robots as semi-autonomous agents (SAAs)

Nonmaleficence is a simple moral principle on its surface. It does not require complex calculation regarding trade-offs of harms and benefits, nor does it require a sophisticated account of, for instance, human moral agency. In practice, however, designing robots to comply with the Principle of

Nonmaleficence is not so simple. Because robots possess a form of agency or autonomy that is significantly different from that possessed by human moral agents, an Ethics for Robots must recognize the challenges created of placing "old" moral principles—such as nonmaleficence—into the "new" form of agency possessed by the robots we are able to create.

As some of us have elsewhere argued, "giving ethics" to robots must account for the distinctive agential and decision-making capacities that robots currently possess [19]. In particular, current and near-term future robots are "semi-autonomous agents" (or SAAs), which means that they possess a limited form of autonomy that distinguishes them in significant ways from other entities to whom we may try to "teach" morality such as children or pets. We often use examples to teach morality to children as a sort of scaffolding, hoping that children will eventually extrapolate from these examples to understand and appreciate general rules and principles of morality. This process allows children to gradually achieve greater levels of autonomy, independence, and responsibility as they grow into full adults. But unlike human children, robots currently lack the potential to develop full autonomy that is similar to human adults.

Consider an analogy with another sort of semi-autonomous agent to whom we "teach" a version of moral principles—namely domesticated dogs. We would not expect a dog, even a well-trained search-and-rescue dog, to make judgments about which human life to prioritize in a triage-type situation, e.g., whether to prioritize saving the human with a greater chance of short-term survival or the one with greater life expectancy once rescued. A dog could perhaps be trained to adhere to proxies for these complex moral judgements—to prioritize saving children over adults, perhaps—but could not (and arguably should not) be making decisions in real time about whom to save and whom to sacrifice. And despite the enormous processing power possessed by current AI systems, it is doubtful whether robots have a more sophisticated sense of what is morally salient than domestic dogs. Dogs, after all, have co-evolved with humans over millennia and we have developed systematic training techniques tailored to the capabilities and limitations that are distinctive to domesticated dogs.

By analogy, we must work to develop systematic programming or "training" techniques that will allow us to instill moral principles into semi-autonomous robots, techniques that are tailored to the distinctive capabilities and limitations of the robots that we have or could soon develop. The goal of this "training" should be to ensure that robots are able to perform tasks and interact with humans in a morally appropriate way, much like the training that we give to dogs. Of course, robots are not the same as dogs, and indeed, many robots possess computational abilities that outstrip those of dogs. Some robots may possess greater capacities for reasoning and cognition as well. But like dogs, we should not simply teach morality to robots as we would to human beings. We will need to develop a distinctive system of moral training—and likely also distinctive moral concepts and rules—tailored to the specific capabilities and limitations that robots possess.

To make this point concrete, we generally support the efforts of researchers like Susan Leigh and Michael Anderson to use a combination of symbolic and machine learning (ML) programming techniques to gradually "teach" morality to robots utilizing a training set of examples of moral dilemma [20]. But we worry that relying on examples of moral dilemmas that are tailored to adult humans, or that are meant to teach children how to eventually be adults, may cause the robot to learn the wrong lessons about morality and to extrapolate moral principles that are appropriate for human agents but not for robots. For example, borrowing an example developed by the Andersons, it may be morally appropriate (or even required, in some cases) for a human healthcare worker to "challenge a patient's decision if [the decision] is not fully autonomous" [21]. And perhaps this principle should be incorporated into an AI system that is designed to give advice to human healthcare workers, as the Andersons suggest. We may not want a robot, however, to make determinations about a patient's autonomy, let along to challenge their autonomy, whatever that may entail.

In summary, we argued in this section that robots are semi-autonomous agents, which requires us to develop a distinctive approach to moral "training" as well as distinctive moral principles that are tailored to the distinctive form of semi-autonomy they possess.

5 The challenges of nonmaleficence

In this section, we will describe important conceptual, technical, moral, and political challenges that we must face in creating an Ethics for Robots, given that robots are semi-autonomous agents. We will use the Principle of Nonmaleficence as an illustrative example.

The first challenge is *conceptual*. Though it may seem simple on its face, the Principle of Nonmaleficence entails significant complexity. First, the Principle of Nonmaleficence can be applied only once we have described a "sphere of responsibility" within which the robot is "responsible" for the harms that it may cause. Attributing responsibility for potential harms will require limiting our expectation of robot causality to some reasonably foreseeable extent [22]. But the boundaries of this sphere of responsibility will be difficult to determine in advance. For example, an eldercare robot with no stair-climbing ability cannot be "responsible" for preventing harms that may occur to humans on the second floor. But such a robot may be "responsible" for dialing 9-1-1 if a perceivable second-floor harm were to occur. In such a case, its sphere of responsibility will depend on its assigned task as well as its programmed abilities and limitations.

Importantly, the sphere of responsibility for not "causing harm" may include both actions and (some) inactions within the scope of possible actions available to a robot at a given time. This is because once a robot has begun to act, or has taken on a specific task, failing to complete this task is arguably (in some cases) a direct cause of harm, not merely "allowing" a harm to occur. For example, a robot designed to assist in surgery that fails to remove a surgical sponge from the incision is not simply allowing the subsequent harm (e.g., an infection), but is responsible for causing it. As such, the harms for which a robot is causally responsible will depend upon the task to which it is assigned, together with the design features and programming of the robot as well as environmental and contextual factors. In short, the Principle of Nonmaleficence, as we understand it, may require a robot to actively intervene to prevent harm in some but not all situations.

To be clear, we are not suggesting that robots can or should be held morally or legally responsible for actions that fall within its sphere of responsibility. As Andreas Matthias has argued, holding AI systems morally or legally responsible may face serious conceptual difficulties [23]. Our point instead is that simply programming a robot to adhere to the Principle of Nonmaleficence—to "do no harm"—will require a conceptually and often ethically complex specification of the robot's sphere of responsibility. Moreover, because it may not be practical or realistic to hold robots morally or legally responsible, the type of responsibility at stake may differ in important ways from familiar forms of human responsibility. In most human contexts, the concept of "responsibility" is tied to things like blame and anger—what philosophers call "reactive attitudes" [24]—and often grounds obligations of apology, remedy, and repair. But the responsibility that is relevant to robots—i.e., responsibility that only specifies the scope of things for which the robot is causally responsible—is likely to be somewhat thinner, and less interconnected with other thick moral and legal concepts. Indeed, it may be that robot responsibility is an emergent and novel form of responsibility unique to robots. To avoid confusion, we suggest the inelegant shorthand "robot responsibility" to refer to this unique form of responsibility [25].

The second type of challenge is *technical*. To create robots that consistently avoid causing harm to humans, robot designers must first identify and systematically catalog the range and types of harms that humans are vulnerable to, and for which a robot may be "responsible" in a given context. We refer to this systematization of possible harms as a "harm ontology." This harm ontology must distill the relevant features of objects and persons and identify the relationships between these features in order to predict possible sources of harm to the persons (i.e., "hazards"). Depending on their design features, robots can be responsible for many physical effects in humans, ranging from stubbed toes to broken bones to mortal injury. Not all of these effects are "harms," however. For example, a Roomba that bumps into Betty's toe may momentarily inconvenience her, but not in a way that qualifies as a harm. An aisle-cleaning robot at the grocery store that knocks away an elderly person's cane or walker, on the other hand, may cause serious injury [26].

Less obvious, but still important, are other types of harms including mental, emotional, and dignitary harms, often tied to the emotional attachments that humans may form to robots. For example, studies have shown that elderly patients form attachments to robotic pets such as AIBO to the same extent as they form attachments to real animals [27]. While AIBO is perhaps not sophisticated enough to require moral training, we should be sensitive to this potential for emotional attachment as we design more advanced robots designed to interact with humans, e.g., in eldercare or educational settings. Emotional harm is a significant type of harm, and in general, robots that are the object of human emotional attachments must not act in ways that provoke emotional distress to those who have formed such attachments. To avoid such harms, robots that may provoke emotional attachments must, at a certain level, be able to perceive potential emotional harms and avoid inadvertent emotional manipulation. As such, a robust harm ontology must identify the full range of harms relevant to any given robot's intended role and context, including non-physical harms. Once we have developed such a harm ontology, it can be used as the basis for a programming structure that can operate within a robot, allowing the robot to *preemptively* identify and assess potential harms.

One of the authors of this article is developing a harm ontology that is represented as a knowledge graph similar to those used to deduce visual affordances, augmented with the relationships needed to infer the dangerousness of objects in the scene [28]. This approach utilizes computer vision techniques to identify objects and attributes in a scene, which are then connected to potential hazards in the knowledge graph. Computer vision can also identify individual people in the scene and their specific vulnerabilities. These vulnerabilities are then connected to potential hazards in the knowledge graph, allowing the system to identify which people are affected by the potential hazards. According to this approach, dangerousness is not an intrinsic property of an object, but rather an inference relating an entity attribute to a vulnerability, and a human's particular susceptibility to be harmed by exposure to that attribute. This harm ontology will incorporate physical harms as well as non-physical harms, including financial harms, psychological and emotional harms, dignitary and reputational harms that negatively impact an individual's perceived social status, and aesthetic or cultural harms that impact persons or groups who place value in objects of social significance.

The third type of challenge is *moral*. Even once robots are able to represent and identify the full array of possible harms within a given context, we must program them to distinguish which "harms" are morally salient within a given context. For example, surgeons must often make a painful incision in order to operate on an internal organ. The incision is morally salient in a number of respects: the patient must give her consent to the procedure (where possible) and the surgeon is responsible for ensuring that the incision is closed and eventually heals. But in general, the harm of the incision is not an all-things-considered reason to avoid the surgery all together. The localized harm of the incision must be understood and evaluated in its broader medical context.

In a similar way, robots that "do no harm" must be able to correctly identify when it is permissible to directly cause a specific or localized harm when that is part of a larger (permissible) task it is performing. In particular, robots must be able to recognize that a scope restriction is tied to a particular role or task. A surgery robot, for example, must be able to process that the physical damage of an incision—which in another context would constitute a harm—is an acceptable harm in this context, even though *accidently* nicking an artery while performing the surgery is an unacceptable harm. The reasoning required in such a case is not the same as the tradeoff or balancing that may be done by a human diagnostician in a shared decision-making context. Instead, the robot surgeon must be able to recognize that the surgical incision itself is not a harm, but a flawed incision is.

Determining moral salience goes beyond assessing permissible trade-offs. A robot must also be able to determine which possible harms they are robot-responsible for, and so must either avoid this possible harm or call for human attention (i.e., transfer, assign, or "hand off" the decision to a human partner). An ability to recognize robot responsibility for potential harm is an important aspect of "doing no harm." We can take some guidance here from the law of personal injury tort and other related fields (including malpractice and product liability). A robot may be robot-responsible for harms that it does not directly cause—e.g., if it fails to recognize and report facts about a patient that would or should raise valid concerns about a patient's mental health or psychological wellbeing. For example, should an eldercare robot be responsible for calling 9-1-1 in an emergency [29]? Likewise, the Principle of Nonmaleficence may require a robot to intervene to protect a young child from swallowing a small object. On the other hand, because of the importance of allowing humans to make (and to take responsibility for) their own decisions, a robot may not be robot-responsible for harms that could result from the risky behavior of a competent adult human.

Some of the authors of this article are developing a "moral-scene assessment" technology for robots, or "moral vision" for short, that will be able to identify potential harms within a scene and correctly specify which harms ought to be avoided [30]. Completing this project, however, will require marshalling the ethical and practical knowledge of a range of academic fields such as biomedical ethics, neuroscience, law, economics, and philosophy, as well as the expertise and experience of professionals who have experience performing in the role intended to be occupied by the robot. For example, nurses and surgical assistants should help to design robots who will assist in surgical procedures.

The fourth challenge is *political*. The role of robots in our collective lives—and the trade-offs we are willing to countenance—are social and political questions, which ought to be resolved through collective decision-making. Qualitative research that investigates the needs, interests, and values of a cross-section of stakeholders would be a very useful first step. Ultimately, however, some questions of morally appropriate trade-offs (e.g., trade-offs between the use of facial recognition and privacy) must be decided by society as a whole, via established democratic institutions. We are already past needing laws and policies to govern the operations of autonomous robotic systems in our societies. These laws and policies must be well informed and democratically legitimate, but there are and will be a range of permissible policy choices. Ultimately, the question of which trade-offs it is permissible for an autonomous robotic system to make is a political question—analogous to trade-offs we make in implementing a public health policy, e.g., rather than a "moral" question of the sort that individual humans must grapple with in daily life.

Finally, all of the challenges described here are heightened if we include an important design desideratum, namely, that robots be able to explain their decisions. For example, a physical therapy robot should not only avoid breaking its patient's bone, but also explain why it did not do so, either with verbal communication, or at least in ways that are accessible to a technician. Moreover, this explanation should reference, at some level, the harm or potential harm to the patient that the robot sought to avoid. A harm ontology will be needed to provide the basis for explanations that are detailed enough (but not too detailed) to be both comprehensible and to allow for updates to the programming. Designing robots that are capable of explanation ensures that, even if a robot causes harm (or fails to avoid harm), humans can be assured that the harm was not intended or by-design, but instead was an "accident" or aberration. Meeting this challenge of explainability, therefore, requires a systematic incorporation of the Principle of Nonmaleficence and other pertinent ethical principles at the earliest stages of robot programming and design.

6 Conclusion

We have argued that, as a first step towards an Ethics for Robots, we should attempt to implement the Principle of Nonmaleficence, or "do no harm," but to do so we must take seriously the various challenges that must be overcome to create robots that are capable of adhering to this superficially simple principle. In particular, the Principle of Nonmaleficence must be tailored to robots to capture both their range of possible actions and the proper scope of their "responsibility" for preventing harmful outcomes as informed both by scholarship and the expertise and experience of the humans who currently hold the relevant roles. Moreover, creating robots that "do no harm" will require equipping them with the ability to identify morally salient features of the landscape which in turn requires the development of a "harm ontology" that formalizes the relationships between persons, objects, and their attributes. We believe that overcoming these challenges in order to design robots who can "first, do no harm," i.e., can adhere to the Principle of Nonmaleficence, should be a necessary starting point, or a precondition, for producing semi-autonomous robots that are designed to interact with humans.

Acknowledgements This paper is based on the research of an interdisciplinary team—consisting of researchers from the Johns Hopkins University (JHU) Applied Physics Laboratory (APL) and the JHU Berman Institute of Bioethics—and was funded by a JHU Discovery Award.

Data availability We do not analyse or generate any datasets because our research for this paper is theoretical in nature.

Declarations

Conflict of interest B Hutler declares no conflicts of interest. TN Rieder declares no conflicts of interest. DJH Mathews declares no conflicts of interest. DA Handelman declares no conflicts of interest. AM Greenberg declares no conflicts of interest.

References

1. Wiens, J., et al.: Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. **25**(9), 1337–1340 (2019)
2. Broad, W.J.: A.I. Versus the Coronavirus. New York Times, March 26, 2020. https://www.nytimes.com/2020/03/26/science/ai-versus-the-coronavirus.html
3. Diab-El Schahawi, M., Zingg, W., Vos, M., et al.: Ultraviolet disinfection robots to improve hospital cleaning: real promise or just a gimmick? Antimicrob Resist Infect Control **10**, 33 (2021). https://doi.org/10.1186/s13756-020-00878-4
4. Muro, M., Maxim, R., Whiton, J.: The robots are ready as the COVID-19 recession spreads brookings: the avenue, March 24, 2020. https://www.brookings.edu/blog/the-avenue/2020/03/24/the-robots-are-ready-as-the-covid-19-recession-spreads/
5. Tasioulas, J.: First steps towards an ethics of robots and artificial intelligence. J. Pract. Ethics 7 (2019). http://www.jpe.ox.ac.uk/wp-content/uploads/2019/07/Issue7_1-1.pdf#page=65
6. This definition draws, in part, on UNESCO.: Report of COMEST on Robotics Ethics. Paris, UNESCO. (2017). http://unesdoc.unesco.org/images/0025/002539/253952E.pdf
7. Himma, K.E.: Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? Ethics Inf. Technol. **11**(1), 19–29 (2009)
8. Haidegger, T.: Autonomy for surgical robots: concepts and paradigms. IEEE Trans. Med. Robot. Bionics **1**(2), 65–76 (2019)
9. Anderson, M., Anderson, S.L., Berenz, V.: A value-driven eldercare robot: virtual and physical instantiations of a case-supported principle-based behavior paradigm. Proc. IEEE **107**(3), 526–540 (2018)
10. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
11. Asimov, I.: Runaround in I, Robot. Doubleday, New York (1950)
12. Anderson, S.L., Anderson, M.: AI and Ethics. AI and Ethics **1**(1), 27–31 (2021)
13. Baker, J.D., Greenberg, A.G.: Natural vs Artificial Intelligence: AI insights from the cognitive sciences (forthcoming)
14. Beauchamp, T.L., Childress, J.F.: Principles of biomedical ethics. Oxford University Press, Oxford (2001)
15. Veatch, R.M.: Resolving conflicts among principles: ranking, balancing, and specifying. Kennedy Inst. Ethics J. **5**(3), 199–218 (1995)
16. Geistfeld, M.A.: A roadmap for autonomous vehicles: State tort liability, automobile insurance, and federal safety regulation. Cal. L. Rev. **105**, 1611–1694 (2017)
17. Kant, I.: Ethical philosophy, 2nd edn. In: translated by James, W., Ellington, J.W., On the supposed right to lie because of philanthropic concerns, pp. 162–166. Indianapolis, Ind., Hackett Publishing, 1994
18. The Man Who Shot Liberty Valance, Ford, J. dir., Paramount Pictures (1962)
19. Rieder, T.N., Hutler, B., Mathews, D.J.: Artificial intelligence in service of human needs: pragmatic first steps toward an ethics for semi-autonomous agents. AJOB Neurosci. **11**(2), 120–127 (2020)
20. Anderson, M., Anderson, S., Armen, C.: An approach to computing ethics. IEEE Intell. Syst. **21**(4), 56–63 (2006)
21. Anderson, M., Anderson, S.L.: Machine ethics: creating an ethical intelligent agent. AI Magazine **28**(4), 15–26 (2007)
22. Palsgraf V.: Long Island Railroad Co., 248 N.Y. 339, 162 N.E. 99 (1928) (holding that railroad could not be liable for an injury caused as a result of a railroad employee dropping an unmarked box of fireworks, which exploded and startled the plaintiff, causing her to trip).
23. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf. Technol. **6**(3), 175–183 (2004)
24. Strawson, P.F.: Freedom and resentment and other essays. Routledge (2008)
25. Fritz, A., et al.: Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). De Ethica **6**(1), 3–22 (2020)
26. Grossman, D.: Walmart buys over 300 aisle-cleaning robots popular mechanics, Dec 6, 2018. https://www.popularmechanics.com/technology/robots/a25428388/walmart-cleaning-robots/

27. Banks, M.R., Willoughby, L.M., Banks, W.A.: Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs. J Am Med Dir Assoc **9**, 173–177 (2008)
28. Greenberg, A.M., Marble, J.L.: Foundational concepts in person-machine teaming. Front. Phys. **10**, 1–16 (2023)
29. Reeves, J.: Surveillance and communication. The handbook of communication and security. Routledge (2019), p. 368–380.
30. Greenberg, A.M.: Deciding Machines: Moral-Scene Assessment for Intelligent Systems. Human-Machine Shared Contexts. Academic Press (2020), p. 135–160.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.