Responsible Robotics and Responsibility Attribution

Aimee van Wynsberghe

Contents

Introduction	240
The “Laws of Responsible Robotics”	240
Beyond the “Laws of Responsible Robotics”	241
Responsibility	242
Forward Looking–Backward Looking Responsibility	242
Moral vs. Legal Responsibility	243
Individual vs. Collective Responsibility	243
Responsibility and Accountability	244
Responsibility and Technology	244
Responsible Robotics Through Responsibility Attribution	245
Framework for Identifying Responsible Individuals	245
What’s Missing in This Framework: Necessary But Not Sufficient for Labelling “Responsible Robotics”	247
Concluding Remarks: Why We Need a Framework for Responsible Robotics	247
References	248

Abstract

This paper stresses the centrality of human responsibility as the necessary foundation for establishing clear robotics policies and regulations; responsibility not on the part of a robot’s hardware or software, but on the part of the humans behind the machines—those researching and developing robotics. Simply put, we need *responsible robotics*. Responsible robotics is a term that has recently ‘come into vogue’, yet an understanding of what *respon-*

sible robotics means is still in development. In light of both the complexity of development (i.e. the many hands involved) and the newness of robot development (i.e. few regulatory boards established to ensure accountability), there is a need to establish procedures to assign future responsibilities among the actors involved in a robot’s development and implementation. The three alternative laws of responsible robotics by Murphy and Wood play a formidable contribution to the discussion; however, they repeat the difficulty that Asimov introduced, that is, laws in general, whether they are for the robot or for the robot-cist, are incomplete when put into practice. The proposal here is to extend the three alternative laws of responsible robotics into a more robust framework for responsibility

A. van Wynsberghe (✉)
Department of Philosophy, Technical University of Delft, Delft,
The Netherlands
e-mail: aimeevanrobot@protonmail.com

attribution as part of the responsible robotics goal. This framework requires making explicit various factors: the type of robot, the stage of robot development, the intended sector of use, and the manner of robot acquisition. With this in mind, one must carefully consider the scope of the ethical issue in question and determine the kind of responsibility attributed to said actor(s).

Keywords

Responsibility · Robotics · Development · Implementation · Regulation

Introduction

The responsible development and use of robotics may have incredible benefits for humanity, from replacing humans in dangerous, life-threatening tasks with search and rescue robots (Murphy 2014) to the last mile delivery of lifesaving resources in humanitarian contexts (Gilman and Easton 2014; Chow 2012). Despite the success and efficiency that robots promise to bring, however, there are societal and ethical issues that need to be addressed. For the last 20 years, robot ethicists have flagged some of the ethical concerns related to robots, for example: the dehumanization and de-skilling of care workers, care receivers, and care practices when robots are used in care contexts (Sharkey 2014; Sparrow and Sparrow 2006; Vallor 2011; van Wynsberghe 2012); the loss of contextual learning necessary for understanding the detailed needs of others when robots replace humans in surgical or humanitarian care (van Wynsberghe and Comes 2019; van Wynsberghe and Gastmans 2008); and a risk of deceiving children when using robots in the classroom (Sharkey 2016), to name a few. To exacerbate these issues, there is growing concern regarding private organizations proving themselves unworthy of society's trust by demonstrating a lack of concern for safety, e.g., the fatal crashes of self-driving cars. If robotics is truly to succeed in making our world a better place, the public must be able to place their trust in the designers, developers, implementers, and regulators of robot technologies. To do this, the many hands in robot development must engage in responsible innovation and implementation, what I will refer to here as *responsible robotics*.

Responsible robotics is a term that has recently “come into vogue” just as similar terms like *responsible research and innovation* (European Commission 2012, 2014; van de Poel and Sand 2018; van den Hoven 2013), *value sensitive design* (Friedman et al. 2015; Friedman 1996; van den Hoven 2013), and other forms of innovation that take societal values, such as privacy, safety, and security, explicitly into account

in the design of a product. In recent years, research courses¹ and articles (Murphy and Woods 2009) have been dedicated to the topic, and a not-for-profit established with the aim of promoting responsible robotics.² Yet, the understanding of what *responsible robotics* means is still in development. In light of both the complexity (i.e., the many hands involved) and newness of robot development (i.e., few regulatory boards established to ensure accountability), there is a need to establish procedures to assign future responsibilities among the actors involved in a robot's development and implementation. This paper starts with an analysis of the three alternative laws of responsible robotics proposed by Murphy and Woods aimed at shifting the discussion away from robot responsibility (in Asimov's stories) to human responsibility. While acknowledging the incredible benefit that these alternative laws bring to the field, the paper presented here will introduce several shortcomings, namely, the need for: a more nuanced understanding of responsibility; recognition of the entire development process of a robot; and recognition of the robot's impact extending beyond the human-robot interaction alone. The paper proceeds by showing the complexity of the concept of “responsibility” and what it might mean in a discussion of responsible robotics. Finally, I suggest a preliminary responsibility attribution framework—a way in which robot applications should be broken down into the various stages of development, sectors, and patterns of acquisition (or procurement) so as to identify the individuals responsible for ensuring that responsibility to tackle ethical issues of prospective robots are addressed proactively.

The “Laws of Responsible Robotics”

In 1942, science fiction writer Isaac Asimov published the short story “Runaround” in which the three laws of robotics first appeared. These laws would become a “literary tool” to guide many of his future robot stories, illustrating the difficulty for robots to embody the same ability for situational judgment as humans. As such, “although the robots usually behaved ‘logically’, they often failed to do the ‘right’ thing” (Murphy and Woods 2009: 14). The three laws were formulated as follows:

“One, A robot may not injure a human being or, through inaction, allow a human being to come to harm . . .

Two, a robot must obey the orders given it by human beings except where such orders would conflict with the First Law . . .

¹IDEA League summer school on responsible robotics, see <http://idealeague.org/blog-summer-school-responsible-robotics-and-ai/> (retrieved Jan 14, 2020).

²See responsiblerobotics.org (retrieved Jan 8, 2010).

Three, a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.” (Asimov 2004: 37).

In 2009 came the first scholarly work to outline the three *alternative* laws of “responsible robotics,” an alternative to the three unworkable laws of robotics found in the Asimov stories (Murphy and Woods 2009). These “alternative three laws of responsible robotics” are stated as follows:

1. A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics.
2. A robot must respond to a human as appropriate for their roles.
3. A robot must be endowed with sufficient situated autonomy to protect its own existence as long as such protection provides smooth transfer of control to other agents consistent with the first and second laws (Murphy and Woods 2009: 19).

To be sure, these alternative laws are monumental for moving robotics research forward in a responsible way. Particularly noteworthy is their emphasis on the centrality of the researcher’s responsibility for following ethical and professional design protocols, on maintaining safety in the robot architecture, and on ensuring the robot is capable of smooth transfer of control between states of human-in-command (or human operated) and full autonomy (or acting without direct real time human interaction). In line with the pivotal role of human responsibility embedded in these alternative laws is a recognition of the difference between robot agency, as presented in Asimov’s stories, and human agency, i.e., the type of agency presumed (and prescribed) for the alternative laws; this difference “illustrates why the robotics community should resist public pressure to frame current human–robot interaction in terms of Asimov’s laws” (Murphy and Woods 2009: 19). Thus, the three alternative laws redirect the world’s attention to the responsibilities of the human actors, e.g., the researchers, in terms of safe human–robot interactions.

Beyond the “Laws of Responsible Robotics”

Without diminishing the value of the contribution that these alternative laws provide, they do fall short in five critical ways. First, although the alternative laws place an emphasis on the centrality of human responsibility, the authors fall short of capturing the ethical nuance in the term “responsibility” outside of a colloquial usage. Responsibility may, in some instances, refer to individual accounts while in other instances may refer to collective accounts: the roboticist is

responsible for high-level planning, while the company is responsible for having a code of conduct for employees. Responsibility may be forward looking (for future consequences) or backward looking (for past consequences). Giving voice to the nuances of responsibility helps to shape a more robust account of responsible robotics: who is responsible for what in the development, deployment, and implementation of robots.

Second, while the alternative laws do refer to the “user” in the human–robot interaction, the focus does not acknowledge the size and scale of today’s robot development process and the many actors involved in bringing a robot from idea to deployment. For some of the more complex technologies of today, there is a recent phenomenon used to describe the difficulty (or inability) to assign any particular person as “responsible” because of the number of actors involved. This is known as “the problem of many hands” (van de Poel et al. 2012). There are “many hands” (Johnson 2015; van de Poel et al. 2012) involved in the development chain of robots from robot developers to producers to implementers to regulators, each with differing roles and responsibilities. Understanding this network of actors, alongside the variety of kinds of responsibility, is also important for identifying who in this network is responsible for what and the form (or kind) of responsibility (e.g., legal vs. moral responsibility).

Third, each of the alternative laws makes reference to the workings of the robot within the human–robot interaction, yet it is important to acknowledge that the impact of the robot will extend far beyond the direct human–robot interaction (van Wynsberghe and Li 2019). Consider surgical robots in the healthcare system as an example; understanding the responsible use of these robots includes understanding that the robot does more than interact safely with surgeon and patient. These robots have changed: the manner of education for medical students (they must also be trained on a robot as well as conventional and/or laparoscopy) (van Koughnett et al. 2009), the allocation of funding and other resources within a hospital (surgical robots costs upwards of \$1 million dollars), and the standard for care (e.g., surgical robots are thought to provide the highest standard of care in many instances and patients may demand this level of care). Responsible robotics should recognize that the introduction of a robot will be felt across an entire system, rather than strictly within the human–robot interaction.

Fourth, the alternative laws do not acknowledge the different stages and/or contexts of robot design, and as such, the various kinds of decision-making (and attributed responsibilities for said decision-making) that occur across these different stages. Being responsible for safe human–robot interactions in a healthcare context in which one must work together with FDA standards differs from ensuring safety for

robots in edutainment contexts, where few standards exist regarding data privacy.³

Fifth, the formulation of the alternative laws hints at an assumption that it is already known what values a robot should be programmed for, e.g., safety. Also, the alternative laws imply that the obstacles facing roboticists are already known and understood. Yet it should be acknowledged that robotics, whether done in the academic space or made in a corporate space and sold to consumers, is very much an experiment, a social experiment when used in society (van de Poel 2013). Society has little experience with robots in personal and professional spaces so it should not be assumed that there is an understanding of the kinds of obstacles or ethical issues to take into account; these insights will only come with more experience of humans and robots interacting in various contexts.

Based on these five elaborations on the three alternative laws of responsible robotics—a recognition of the many hands involved in robot development, the various kinds of responsibility attributed to these hands, the various stages of robot development and deployment, the experimental nature of robots in society, and the impact of the robot extending beyond the human—robot interaction alone—it becomes paramount to take a closer look at the three alternative laws in order to provide clear, complete guidelines for *responsible robotics*. What is needed for robot designers, developers, implementers, and regulators, is a tool to assign (a type of) responsibility to various (groups of) actors at the various stages of robot development and deployment in an academic (or any other) context. One may suggest an extension of these three alternative laws so as to incorporate various conceptions of responsibility or the various actors involved. But perhaps it is, instead, necessary to question the utility of strict laws or rules governing robotics in the first place. Given the complexity of robotics (e.g., different actors produce the hardware from the actors who create something with it and these again differ from the actors who will implement the robot), the experimental nature of robotics (whether the robot is being tested in an academic setting or in the wild) and the oftentimes lack of policy or regulations to guide the various actors (e.g., few Universities have ethical review boards dedicated to reviewing robotics experiments at Universities) perhaps a more flexible approach is needed to conceptualize responsible robotics. Specifically, what is missing from the alternative laws, and sorely needed, is a robust framework for responsibility attribution. Such a framework should be tasked with outlining the responsible actor or organization at the various stages in a robot's development. Even if an organization buys into the idea of *responsible robotics* in

theory, without designating a specific actor as responsible for these processes, they may never actually take place. The responsible actor(s) are tasked with the development process of the robot itself, but should also seek to go beyond these values through engaging with an ethical technology assessment (Palm and Hansson 2006), as an example, in order to capture the range of ethical issues in need of addressing.

With these thoughts in mind, the next section of this paper will take a closer look at the concept of responsibility to explore some of the distinctions in the term that can add help to shape a responsibility attribution framework for responsible robotics.

Responsibility

There is a wealth of literature on the concept of responsibility aimed at differentiating the different meanings of the term (Feinberg 1988; Johnson 2015; van de Poel and Sand 2018). Some works focus on distinctions between: normative and descriptive notions of responsibility (van de Poel et al. 2012; van de Poel and Sand 2018); senses of responsibility (Hart 2008); and/or between temporal accounts of responsibility, e.g., forward-looking and backward-looking responsibility (van de Poel et al. 2012; van de Poel and Sand 2018). Added to this, there are also heated discussions about the “traditional precondition of responsibility,” namely that individuals who can be held responsible must meet the conditions of: “capacity, causality, knowledge, freedom, and wrong-doing” (van de Poel et al. 2012: 53). It would be impossible and unnecessary to cover all aspects of responsibility in this in this paper; however, for the purposes here—sketching a framework to prospectively assign responsibilities in the development of future robots—it is necessary to identify certain salient distinctions that help shape the concept of responsible robotics. The purpose of presenting certain distinctions in the next section is therefore twofold: first, to highlight the need for more granular discussions when using a phrase like “responsible robotics”; and, second, to identify the conceptualizations of responsibility necessary to establish a framework for assigning responsibility among the various actors in a robot's development and implementation.

Forward Looking–Backward Looking Responsibility

One of the first distinctions to address is that between forward-looking responsibility and backward-looking responsibility. Backward-looking responsibility relates to “things that have happened in the past and usually involves an evaluation of these actions and the attribution of blame or praise to the agent” (van de Poel and Sand 2018: 5; see also

³To be sure, reference to few privacy standards outside of a European Context where the new General Data Protection Regulation exists to ensure data privacy of consumers and citizens across Europe.

Smith 2007; Watson 2004). In robotics, backward-looking responsibility may refer to identifying who is responsible for a robot malfunctioning.

Forward-looking responsibility, on the other hand, refers to “things that have not yet occurred” (van de Poel et al. 2012: 51; van de Poel and Sand 2018). Forward-looking responsibility, therefore, can be understood as the responsibility to prevent harm from happening and/or possessing the character trait of being responsible (van de Poel and Sand 2018: 6). In robotics, forward-looking responsibility may refer to having an individual (or team) in a robotics company tasked with uncovering novel unintended consequences arising from new robot capabilities, and working towards their mitigation.

Of course, it is not possible to entirely separate backward- and forward-looking responsibility; if one establishes forward-looking responsibilities as moral obligations, then at a certain moment they could be breached, and as such, a discussion of backward-looking responsibility occurs. While it is necessary to consider both temporal dimensions in the assignment of responsibility, let us consider, for this paper, the prospective robots of the future, the robots that are currently in the stage of idea generation and will be designed, developed, and deployed over the next 5–10 years. For many of these robots there are no policy structures to guide researchers or developers. In such experimental situations of development, it is necessary to establish norms and expectations about who is responsible for what. One specific sense of forward-looking responsibility is understood as responsibility-as-obligation and refers to instances in which “one has to see to it that a certain desirable state-of-affairs is obtained, although one is free in how this state-of-affairs is to be brought about” (van de Poel et al. 2012: 52). Therefore, in this paper, let us consider what type of responsibility attribution framework is needed in the situations where we must establish prospective rules of engagement. Furthermore, in the creation of the responsibility attribution framework here, our goal is to identify the key players who should hold forward-looking responsibility-as-obligation to realize responsible robotics procedures and products.

Moral vs. Legal Responsibility

Another interesting distinction to be made in a discussion of responsibility is that between moral responsibility and legal responsibility. Moral responsibility has been defined as “responsibility that is attributed on moral grounds rather than on basis of the law or organizational rules” (van de Poel et al. 2012: 51). Legal responsibility may be considered more descriptive than moral. Considering the alternative law of responsible robotics #1 (i.e., A human may not deploy a robot without the human–robot work system meeting the highest

legal and professional standards of safety and ethics), it may not be as simple as saying that humans “must meet the highest legal and ethical standards” given that the legal standards of the organization in which one works may be in conflict with the moral standards of the individual roboticist. One may be morally responsible to follow his/her own ethical principles, for instance voicing unethical practices (aka whistleblowing) within a company where he or she works even when he or she has a contract and maybe even an NDA (nondisclosure agreement) that imposes a legal responsibility for him/her not to speak out (Lenk and Maring 2001).

To be sure, keeping to one’s contract is also a type of moral responsibility (in addition to a legal one) and creates a conflict when deciding which moral responsibility to prioritize above the other. There are many situations throughout history when we see employees speaking out against the company for which they work; for example, in the 1986 *Challenger* explosion, the project leader warned against the scheduled space travel to NASA and was ignored, resulting in the *Challenger* explosion killing seven astronauts (Lenk and Maring 2001).

Of late, we hear more and more stories of “whistle-blowers,” individuals breaking NDAs to speak out about ethically problematic treatment of employees or company practices in large tech companies. For example, there was a recent case in which Facebook content moderators broke their NDAs to speak about the psychological and physical suffering resulting from their jobs (Newton 2019). A roboticist working in a social robotics company may develop certain moral misgivings about their work after more sophisticated prototypes verge on deceiving the human users who interact with it into believing it can form emotional bonds. Roboticists may question their legal responsibility to their employer when it conflicts with their own moral responsibility to maintain expectations and/or be truthful with the public.

The distinction between moral and legal responsibility highlights once again the need for a more granular understanding of responsibility in a discussion of responsible robotics; perhaps following a code of conduct legally dictated by a company may impede following one’s own moral responsibility. What such a distinction also points towards is a difference, and possible conflict, between individual and collective forms of responsibility.

Individual vs. Collective Responsibility

Not only are there temporal distinctions in a discussion of responsibility but also distinctions about the agents that can bear responsibility. For some, when discussing moral responsibility “only human beings can be held responsible and not collective entities” (Miller 2006). Moreover, “. . . social

groups and organizations, have collective moral responsibility only in the sense that the individual human persons who constitute such entities have individual moral responsibility, either individually or jointly” (Miller 2006: 176).

To put this in more concrete terms, a robotics team may be collectively (or jointly) responsible for creating a robot prototype (i.e., the end), but the realization of this collective end results from the individual activities on team members in, for example, navigation, haptics, high-level planning, mechatronics, ethics, and so on. Thus, we might say that companies have a collective end of creating robots in a responsible way. However, it is not through a collective responsibility but through individual responsibilities for specific employees that acts of responsibility occur, including establishing an ethics review board to assess workflows, hire a privacy expert to ensure data privacy, or a sustainability expert to ensure selection of sustainable materials, and so on. Many organizations can have the same collective end of responsible robotics but the responsibilities of the individuals within said organization will differ depending on the type of organization, for example—a private company will have to answer to shareholders, while a University will have to answer to a public board. This distinction is important for understanding that it will not be possible to simply state that a company has the collective end of responsible robotics; rather, it will be necessary to identify what the individual responsibilities are within said company that are needed in order to realize the collective end.

Responsibility and Accountability

Thus far, the discussion has focused on responsibility—who is responsible for any ethical lapses that have occurred, or that may occur in the future, in relation to a particular robot in development. However, roboticists working in a company developing commercially available products may also feel *accountable* to the public at large for developing products that contribute to the well-being of individuals, or at the very least do not introduce new forms of harm. “Accountability-responsibility is embedded in relationships that involve norms and expectations . . . In accountability relationships those who are accountable believe they have an obligation to a forum, e.g., a community, the public, a particular individual or group of individuals. Members of the forum believe that they are owed an explanation; they expect that those who are accountable will answer (provide an account) when they fail to adhere to appropriate norms, i.e., fail to live up to expectations” (Johnson 2015: 713). Accountability differs from responsibility in that in the case of accountability, the responsible party must justify their decisions and actions to an outside entity. For example, many companies may feel accountable to the general public, who

may have certain expectations of robot makers (although it is worth noting that these expectations may or may not be based in reality vs. taken from what they see in movies or hear from the press).

While the norms and expectations up for discussion here are usually known to companies attempting to meet those expectations (e.g., a University advisory board may provide written, public reports), they may be established in either a formal or an informal manner, according to Johnson (2015). They can be formally enshrined in legal obligations or codes of conduct of engineers or they may be informally held by a group based on experience and/or public communication. The general public, for example, may have certain expectations of robot capabilities based on stories they see in the press that give certain expectations of robot capabilities—robots capable of falling in love (Levy 2008), of moral agency (Anderson and Anderson 2011), or of having consciousness (Himma 2009). It may be roboticists (i.e., academic and corporate) who are accountable for recalibrating expectations of the public.

Responsibility and Technology

The question concerning responsibility and robots is further complicated by the tangled relationship that responsibility shares with technology in general and robotics (or autonomous agents) in particular. An added concern for the field of robotics is the level of autonomy a robot may achieve if embedded with artificial intelligence that can allow the robot to learn and function in real time without direct human intervention. Such sophisticated levels of autonomy have led some scholars to raise the question of whether or not a responsibility gap will ensue: “we will not be able to control or predict how [highly specialized artificial agents] will behave” (Johnson 2015: 709).

There have been various rejections of the so-called responsibility gap. Some authors refer to accountability-responsibility relations and also perhaps the legal responsibilities of engineers and remind us that “engineers would be held responsible for the behavior of artificial agents even if they can’t control them, on grounds of professional responsibility” (Nagenborg et al. 2008). Other scholars respond to the idea of a responsibility gap by suggesting that robots themselves could be held responsible. While once again bracketing whether or not a robot could meet the conditions of human responsibility the argument raised in such situations “rests finally on the tendency of humans to assign responsibility to computers and robots rather than something that would justify the attribution of responsibility” (Johnson 2015: 705).

Whereas Murphy and Woods address the impossibility of robots being responsible in their discussion of the three alter-

native laws by stressing the technical impossibility [“robots cannot infallibly recognize humans, perceive their intent, or reliably interpret contextualized scenes” (Murphy and Woods 2009: 15)], Johnson appeals to the wishes of society, that society would not flourish in a state in which no humans were accountable or responsible for the consequences of robot actions, or the malfunction of robot products; “that the human actors involved would decide to create, release, and accept technologies that are incomprehensible and out of the control of humans” (Johnson 2015: 712).

Let me suggest, in accordance with Johnson, Murphy, and Woods among others, that the concept of responsibility in the phrase *responsible robotics* should be a label attributed not to the robots themselves but to the humans acting to make, study, use, regulate, or take apart robot products and services. Therefore, responsible robotics ultimately needs to refer to the kinds of choices made by the humans involved in a robot’s design, development, deployment, and regulation: how were decisions calculated, what other options were explored, what kinds of assessments were done to understand and minimize (or mitigate) negative consequences, and what kind of transparency developers and researchers provided to users/customers.

Responsible Robotics Through Responsibility Attribution

Let us suggest that the collective goal of an organization is to develop responsible robotics—to have procedures in place for establishing procedural trust, and for creating products that are considered to be responsibly developed. As discussed so far, we would not consider these organizations to be morally responsible as a whole, but we would expect the organization to designate responsible individuals, and we would then consider the individuals within said organization to be morally responsible for the collective end of developing responsible robotics.

At this moment, *responsible robotics* requires the execution of (at least) two steps or phases: a phase in which ethical issues are uncovered within an organization, and a second phase in which responsibility (in a forward-looking responsibility-as-obligation sense) for solving said issues is attributed to an individual or group of individuals.

For a first step, one could suggest that each organization involved in creating robots that wants to do it responsibly should be accountable for addressing ethical issues in the research and design (R & D) of their robot. An organization could rely on some of the more well-known ethical issues, e.g., privacy, sustainability, safety, and security, and translate these issues to their context and robot prototype. Meaning, designers, and implementers of a robot for a hospital context may design for the value of privacy and interpret privacy

as both corporeal privacy of patient bodies and privacy of personal (medical) data that the robot has access to.

While a substantive conversation about ethical issues is an important first step towards engaging in responsible robotics, there are formalized processes available to make sure that such conversations are as in-depth, comprehensive, and ultimately as effective as possible. Specifically, an organization could engage in a more in-depth assessment to uncover a greater range of the possible ethical issues they may encounter in R & D, for example, methods such as ethical Technology Assessment (eTA) (Palm and Hansson 2006), Care Centred Value Sensitive Design (CCVSD) (van Wynsberghe 2012, 2013), constructive technology assessment (Schot and Rip 1997), and/or ethicist as designer (van Wynsberghe and Robbins 2014), among others. Each of these approaches differs in scope, but is able to produce a list of ethical and societal concerns related to the development process and/or the resulting artifact. An eTA, for example, may provide information concerning privacy issues related to a certain robot application. A CCVSD approach alternatively will produce a list of ethical concerns related to the impact of a robot prototype on the expression of care values. Ideally, the choice of framework for identifying ethical issues should be informed by the context of use. A care robot to be designed for, or used in, a healthcare institution would benefit from a CCVSD approach whereas the developers of a personal robot assistant could benefit from a more generic eTA approach.

In either case, once an issue (past or future) has been identified, one must assign an agent responsible for mitigating or preventing said issue. However, in situations with so “many hands,” how can this be done, and more importantly how can this be done in a systematic way to create a level playing field for all robotic companies, organizations, and institutions? This is precisely where we are in need of a framework to help solve the attribution of responsibility to individuals. Such a framework must be broad enough that it can capture the range of variables and stakeholders involved while at the same time specific enough that it allows one to appropriately assign responsibility at a more granular level (making distinctions in kinds of responsibility). As mentioned earlier, in practice this should be possible to do in both a forward and/or backward-looking sense, but for this paper we will consider a forward-looking sense predominantly.

Framework for Identifying Responsible Individuals

The components of the framework, when taken together, provide an analysis of a particular robot prototype or class of robots in various stages of development (Fig. 1). For example, the daVinci surgical robot is a prototype, while

Fig. 1 Framework for responsibility attribution. Source: created by the author

the surgical robots are a class of robots (and consist of a variety of commercial products including the daVinci, Zeus, etc.). To do this, we begin with identifying the type, or class, of robot. If speaking about one robot on the market, for example, a DGI drone, versus a class of robots on the market (e.g., drones or surgical robots), then the scope of responsible individuals is narrowed down. Moreover, this criterion is also about specifying the type of robot in terms of its tasks, goals, and capabilities. A greeter robot in hospitals without the capability to collect or store data on individuals in its path will raise a different set of privacy questions compared to a greeter robot in a store designed to collect and store data on customers it interacts with. Thus, the type of robot also alerts one to the range of ethical issues at stake (perhaps in addition to those identified in the eTA or CCVSD analysis).

Next, one must specify the moment in the robot's life cycle, for example, idea generation, first robot prototype, prototype following numerous iterations, large-scale development, distribution, implementation, or regulation. The idea behind this step is to introduce the understanding that the responsible individuals are not just the robot developers, but also the people who will be purchasing, implementing or even regulating, robots. Accordingly, if we consider the surgical robot daVinci, a robot that is produced on a large scale by a company called Intuitive Surgical but is sometimes distributed around the world by other intermediary companies, there will be certain responsibilities that the producers will have (e.g., choice of materials, meeting of ISO standards), while other responsibilities will fall on distributors (e.g., safe delivery and assistance with implementation), and even other responsibilities will fall on hospital administrators to ensure responsible deployment in the hospital (e.g., safety protocols, training of surgeons and other medical professionals).

Third, following identification of the maturity of the robot's development, the next step is to identify the context in which the robot is intended to be used, e.g., agriculture, defense, or healthcare. Such a specification also acts to define (and in some instances limit) the scope of actors able to take responsibility. For example, robots bought and used in a healthcare context may demand responsibility from individuals related to hospital administration, insurance companies, the FDA/EMA, and/or other related institutions. In another context of use—for example agriculture—different stakeholders, like farmers, may be introduced. If the robot is still early on in the experimental stages and is

not yet commercially available, companies may be working closely with farmers to understand their specific needs and farmers may have a moral (not a legal) responsibility to assist in this process. Once the robot is commercially available, as in the case with robots for milking cows, feeding livestock, or cleaning stalls, farmers may no longer be responsible for providing insights into their daily needs, but they may be responsible for providing feedback on the efficacy, safety, or data security measures of the robot.

Fourth, one should identify the robot's mode of acquisition. Many robots used in an academic context are off-the-shelf products made and tested in industry, e.g., the Nao and Pepper platforms of Softbank. If one were to consider the ethical concern of e-waste, it seems beyond the limits of academics to be held responsible for this when they are not themselves making the robot and, even worse, have no environmentally sustainable alternative. In such instances, it would seem adequate to hold industry responsible for sustainability issues related to the robots made for study within academia. If, however, there were multiple robotics platforms available, one of which is a sustainably sourced one, then it seems appropriate to hold academics responsible for the purchasing of sustainable products. The same considerations hold for companies or NGOs who purchase off-the-shelf robots for their own use, e.g., grocery stores that are buying robots to implement in their warehouses and distribution centers. In the case of companies making and distributing their own robots to consumers/users, these companies have a responsibility to search for sustainable avenues of doing so (e.g., percentage of recycled plastics, minerals sourced for the batteries coming from certified mines).

In short, the framework presented here is meant to create a procedure for assigning responsibilities to the variety of actors involved in robot R & D, purchasing, implementation, and regulation of robots. At each stage of the framework one is called upon to list all possible responsible individuals for the ethical issue in question and to refine this list at each of the subsequent stages. By engaging in this framework in this way, one is then able to define the scope of individuals within organizations who bear responsibility for the consequences of the ethical issue at stake (whether in the design, development, use, implementation, and/or regulation). Depending on the ethical issue, the sector, context, and type of robot, the individual who should be assigned responsibility may change. In some instances, it may be developers of

robots that bear responsibilities (e.g., companies making robots responsible for reducing e-waste concerns in their production or procurement officers responsible for providing alternatives to small start-up companies) whereas in other instances it may be the ones purchasing the robots that bear responsibilities (e.g., consumers or companies using robots in their retail processes may be responsible for purchasing sustainable robots). In other instances, it may be the policy makers responsible for developing governance tools to facilitate sustainable development of robots through the creation of subsidies to companies, laws to prevent the breaking of International e-waste treaties, and guidelines to show support for the responsible development of robots.

What's Missing in This Framework: Necessary But Not Sufficient for Labelling "Responsible Robotics"

There are two main criticisms of the work I have presented here: first, distinguishing between "should" questions and "how should" questions. The second being the need for broader policy and infrastructure questions. Neither one of these criticisms takes away from the content presented but each should be acknowledged nonetheless.

First, each of the alternative laws of responsible robotics, along with the responsibility attribution framework presented here, presume that the robot in question *should* be made and the challenge lies in teasing out *how* the robot should be made or in assigning a responsible individual for mitigating an ethical concern. However, responsible robotics must also be about questioning the very application of a robot. A robot for babysitting children may fit all ISO/FDA standards, it may achieve smooth transfer of autonomy, it may be designed with clear expectations of accountability and chains of responsibility to avoid the problem of many hands, and yet these criteria do not address the question of whether or not the robot *should* be made and/or used in the first place. Such a "babysitting" robot could erode bonds between parents and children over time or drastically change the socialization of children and for these reasons be considered unethical to develop and/or use no matter the production process. Thus, the framework presented here may be considered a necessary criterion of responsible robotics but by no means sufficient to claim responsible robotics.

Second, a big question missing here is: who does the assigning of responsible individuals? This is precisely where ethics meets policy as the responsibility infrastructure I am envisioning here will extend far outside an organization and into the realm of developer, implementer, and user. Who will determine the responsible individual (and the determinants of such a role) and who will make the responsible party do

the work of being responsible are questions inviting answers from the policy making space.

Concluding Remarks: Why We Need a Framework for Responsible Robotics

Robotics has come a long way in the last decades. Improvements in technology allow already innovative robots to do even more than they once could. For example, robots now work in factories alongside humans (i.e., cobots) rather than behind cages as they originally did. Financially speaking, robot sales reportedly increase every year. According to the International Federation of Robotics "in 2017, [industrial] robot sales increased by 30% to 381,335 units, a new peak for the fifth year in a row" (IFR [n.d.-a](#): 13) and "the total number of professional service robots sold in 2017 rose considerably by 85% to 109,543 units up from 59,269 in 2016. The sales value increased by 39% to US\$ 6.6bn" (IFR [n.d.-b](#): 11). Globally speaking, "Robotics investments in December 2018 totaled at least \$652.7 million worldwide with a total of 17 verified transactions" (Crowe [2019](#)). With continued investments in robotics, it seems likely that such trends will continue. The question to ask is how to design and develop this technology in a way that pays tribute to societal values, resists the urge to exacerbate existing ethical problems (such as environmental sustainability), and proceeds in a resilient manner to navigate unknown ethical issues as they are revealed.

If robotics is truly to succeed in making our world a better place, the public must be able to place their trust in the designers, developers, implementers, and regulators of robot technologies. To do this, we must engage in the responsible research and innovation of robot development processes and the robots that result from these processes; we need *responsible robotics*. The three alternative laws of responsible robotics by Murphy and Wood play a formidable contribution to the discussion on responsible robotics; however, they repeat the difficulty that Asimov introduced, that is, laws in general in the robotics space, whether they are for the robot or for the roboticist (or any other actor in the design process), are incomplete when put into practice. The proposal here is to extend the three alternative laws of responsible robotics into a more robust framework for responsibility attribution as part of the responsible robotics goal. Such a framework is meant to draw attention to the network of actors involved in robot design and development, and to the differences in kinds of responsibility that each of these actors (either individuals or organizations) may have.

The responsibility attribution framework requires identification of various factors: the type of robot, the stage of robot development, the intended sector of use, and the manner of

robot acquisition. Identifying these details in a step-by-step manner allows one to land on the stakeholder deserving of responsibility. With this in mind, one must carefully consider the kind of ethical issue (or societal value) in question and determine the kind of responsibility attributed to said actor(s). Such a framework rests on four starting assumptions related to the definition and concept of responsible robotics: (1) Responsibility of humans (and not robots) involved in the creation and use of robots; (2) Responsibility understood as a vast concept with various distinctions; (3) Robotics understood as a process with various stages of development for which different actors will bear different responsibilities; (4) An understanding of the impact of robots on systems rather than ending at the human–robot interaction.

The question of how to make robotics in a responsible way and what such products would look like is colossal—impossible to answer in one paper. This paper was meant to open the door for the discussion of a framework to encourage analysis allowing one to arrive at a decision concerning *who* is responsible for mitigating or solving a particular ethical/societal issue. In short, the phrase *responsible robotics* follows from the recognition of robotics as a social experiment and is meant to convey that the robotics experiment be done responsibly. It is directed at the people who design, develop, regulate, implement, and use the entire range of robotics products. It is, furthermore, about ensuring that those people are responsible for proactively assessing and taking actions, which ensures that robotics products respect important societal values.

References

- Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge: Cambridge UP.
- Asimov, I. (2004). Runaround. In I. Asimov (Ed.), *I, Robot* (pp. 25–45). New York: Random House.
- Chow, J. C. (2012). *The case for humanitarian drones*. Available via OpenCanada. Retrieved February 28, 2020, from <https://www.opencanada.org/features/the-case-for-humanitarian-drones/>
- Crowe, S. (2019). *Robotics investments recap: December 2018*. Available via The Robot Report. Retrieved May 7, 2019, from <https://www.therobotreport.com/robotics-investments-recap-december/>
- European Commission. (2012). *Responsible research and innovation: Europe's ability to respond to societal challenges*. London: E Union.
- European Commission. (2014). *Rome declaration on responsible research and innovation in Europe*. London: E Union. Available via European Commission. Retrieved February 28, 2020, from https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf
- Feinberg, J. (1988). Responsibility for the future. *Philosophy Research Archives*, 14, 93–113. <https://doi.org/10.5840/prar1988/19891427>
- Friedman, B. (1996). Value-sensitive design. *Interactions*, 3(6), 16–23. <https://doi.org/10.1145/242485.242493>
- Friedman, B., Hendry, D., Hultdtgren, A., Jonker, C., van den Hoven, J., & van Wynsberghe, A. (2015). Charting the next decade for value sensitive design. *Aarhus Series on Human Centered Computing*, 1(1), 4. <https://doi.org/10.7146/aaahcc.v1i1.21619>
- Gilman, D., & Easton, M. (2014). *Unmanned aerial vehicles in humanitarian response*. Occasional Policy Paper 010. Available via United Nations Office for the Coordination of Humanitarian Affairs. Retrieved February 28, 2020, from <https://docs.unocha.org/sites/dms/Documents/Unmanned%20Aerial%20Vehicles%20in%20Humanitarian%20Response%20OCHA%20July%202014.pdf>
- Hart, H. L. A. (2008). *Punishment and responsibility: Essays in the philosophy of law*. Oxford: Oxford UP.
- Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11(1), 19–29. <https://doi.org/10.1007/s10676-008-9167-5>
- IFR. (n.d.-a). *Executive summary world robotics 2018 industrial robots*. Available via IFR. Retrieved May 7, 2019, from https://ifr.org/downloads/press2018/Executive_Summary_WR_2018_Industrial_Robots.pdf
- IFR. (n.d.-b). *Executive summary world robotics 2018 service robots*. Available via IFR. Retrieved May 7, 2019, from https://ifr.org/downloads/press2018/Executive_Summary_WR_Service_Robots_2018.pdf
- Johnson, D. G. (2015). Technology with no human responsibility? *Journal of Business Ethics*, 127(4), 707–715. <https://doi.org/10.1007/s10551-014-2180-1>
- Lenk, H., & Maring, M. (2001). Responsibility and technology. In A. E. Auhagen & H.-W. Bierhoff (Eds.), *Responsibility: The many faces of a social phenomenon* (pp. 93–108). London: Routledge.
- Levy, D. (2008). *Love and sex with robots: The evolution of human-robot relationships*. New York: Harper Perennial.
- Miller, S. (2006). Collective moral responsibility: An individualist account. *Midwest Studies in Philosophy*, 30(1), 176–193. <https://doi.org/10.1111/j.1475-4975.2006.00134.x>
- Murphy, R. (2014). *Disaster robotics*. Cambridge, MA: MIT Press.
- Murphy, R., & Woods, D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems*, 24(4), 14–20. <https://doi.org/10.1109/MIS.2009.69>
- Nagenborg, M., Capurro, R., Weber, J., & Pingel, C. (2008). Ethical regulations on robotics in Europe. *AI & Society*, 22(3), 349–366. <https://doi.org/10.1007/s00146-007-0153-y>
- Newton, C. (2019). *Bodies in Seats: At Facebook's worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives*. Available via The Verge. Retrieved April 30, 2020, from <https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa>
- Palm, E., & Hansson, S. O. (2006). The case for ethical technology assessment (eTA). *Technological Forecasting and Social Change*, 73(5), 543–558. <https://doi.org/10.1016/j.techfore.2005.06.002>
- Schot, J., & Rip, A. (1997). The past and future of constructive technology assessment. *Technological Forecasting and Social Change*, 54(2–3), 251–268. [https://doi.org/10.1016/S0040-1625\(96\)00180-1](https://doi.org/10.1016/S0040-1625(96)00180-1)
- Sharkey, A. (2014). Robots and human dignity: A consideration of the effects of robot care on the dignity of older people. *Ethics and Information Technology*, 16(1), 63–75. <https://doi.org/10.1007/s10676-014-9338-5>
- Sharkey, A. (2016). Should we welcome robot teachers? *Ethics and Information Technology*, 18, 1–15. <https://doi.org/10.1007/s10676-016-9387-z>
- Smith, A. M. (2007). On being responsible and holding responsible. *The Journal of Ethics*, 11(4), 465–484. <https://doi.org/10.1007/s10892-005-7989-5>
- Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, 16(2), 141–161. <https://doi.org/10.1007/s11023-006-9030-6>

- Vallor, S. (2011). Carebots and caregivers: Sustaining the ethical ideal of care in the twenty-first century. *Philosophy and Technology*, 24(3), 251–268. <https://doi.org/10.1007/s13347-011-0015-x>.
- van de Poel, I. (2013). Why new technologies should be conceived as social experiments. *Ethics, Policy & Environment*, 16(3), 352–355. <https://doi.org/10.1080/21550085.2013.844575>.
- van de Poel, I., & Sand, M. (2018). Varieties of responsibility: Two problems of responsible innovation. *Synthese*. <https://doi.org/10.1007/s11229-018-01951-7>.
- van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics*, 18(1), 49–67. <https://doi.org/10.1007/s11948-011-9276-0>.
- van den Hoven, J. (2013). Value sensitive design and responsible innovation. In R. Owen, J. Bessant, & M. Heintz (Eds.), *Responsible innovation: Managing the responsible emergence of science and innovation in society* (pp. 75–84). London: Wiley.
- van Koughnett, J., Jayaraman, S., Eagleson, R., Quan, D., van Wynsberghe, A., & Schlachta, C. (2009). Are there advantages to robotic-assisted surgery over laparoscopy from the surgeon's perspective? *Journal of Robotic Surgery*, 3(2), 79–82. <https://doi.org/10.1007/s11701-009-0144-8>.
- van Wynsberghe, A. (2012). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics*, 19(2), 407–433. <https://doi.org/10.1007/s11948-011-9343-6>.
- van Wynsberghe, A. (2013). A method for integrating ethics into the design of robots. *Industrial Robot*, 40(5), 433–440. <https://doi.org/10.1108/IR-12-2012-451>.
- van Wynsberghe, A., & Comes, T. (2019). Drones in humanitarian contexts, robot ethics, and the human–robot interaction. *Ethics and Information Technology*, 22, 43. <https://doi.org/10.1007/s10676-019-09514-1>.
- van Wynsberghe, A., & Gastmans, C. (2008). Telesurgery: An ethical appraisal. *Journal of Medical Ethics*, 34(10), e22. <https://doi.org/10.1136/jme.2007.023952>.
- van Wynsberghe, A., & Li, S. (2019). A paradigm shift for robot ethics: From HRI to human–robot–system interaction (HRSI). *Medicolegal and Bioethics*, 9, 11–20. <https://doi.org/10.2147/MB.S160348>.
- van Wynsberghe, A., & Robbins, S. (2014). Ethicist as designer: A pragmatic approach to ethics in the lab. *Science and Engineering Ethics*, 20(4), 947–961. <https://doi.org/10.1007/s11948-013-9498-4>.
- Watson, G. (2004). Reasons and responsibility. In G. Watson (Ed.), *Agency and answerability: Selected essays* (pp. 289–317). Oxford: Oxford UP.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.