

Robot Ethics: A View from the Philosophy of Science

GUGLIELMO TAMBURRINI^{a,1}

^a*Dipartimento di Scienze Fisiche, Università di Napoli Federico II*

Abstract. Robot ethics is a branch of applied ethics which endeavours to isolate and analyse ethical issues arising in connection with present and prospective uses of robots. These issues span human autonomy protection and promotion, moral responsibility and liability, privacy, fair access to technological resources, social and cultural discrimination, in addition to the ethical dimensions of personhood and agentivity. This chapter examines distinctive roles that epistemological and methodological reflections on robotics play in ethical inquiries on robotics. In particular, reflections of this sort on models of robot-environment interactions and on models of learning robotic systems are brought to bear on the analysis of autonomy and responsibility issues in robot ethics.

Keywords. Ethics, autonomy, responsibility, field robotics, service robotics, military robots, learning robots, prediction of robotic behaviours.

Introduction

Robots are machines endowed with sensing, information processing, and motor abilities. Information processing in robotic systems takes notably the form of perception, reasoning, planning, and learning, in addition to feedback signal processing and control. The coordinated exercise of these abilities enables robotic systems to achieve goal-oriented and adaptive behaviours. Communication technologies enable robots to access networks of software agents hosted by other robotic and computer systems. New generations of robots are becoming increasingly proficient in coordinating their behaviours and pursuing shared goals with heterogeneous teams of agents which include other robots, humans, and software systems.

During the last decades of the last century, robots were mostly confined to industrial environments, and rigid protocols severely limited human-robot interaction (HRI) there. The rapidly growing research areas of field and service robotics² are now paving the way to more extensive and versatile uses of robots in non-industrial environments, which range from the extreme scenarios of space missions, deep sea explorations, and rescue operations to the more conventional human habitats of workshops, homes, offices, hospitals, museums, and schools. In particular, research in a special area of service robotics called personal robotics is expected to enable richer and more flexible forms of HRI in the near future, bringing robots closer to humans in a variety of healthcare, training, education, and entertainment contexts.

¹ Guglielmo Tamburrini, Dipartimento di Scienze Fisiche, Università di Napoli Federico II, Complesso Universitario Monte S. Angelo, Via Cintia, I-80126 Napoli, Italy. E-mail: tamburrini@na.infn.it

² See the informative chapters on field and service robotics in Siciliano & Khatib (2008).

Robot ethics is a branch of applied ethics which endeavours to isolate and analyse ethical issues arising in connection with present and prospective uses of robots. The following questions vividly illustrate the range of issues falling in the purview of robot ethics.

- Who is responsible for damages caused by service and personal robots?
- Are there ethical constraints on the design of control hierarchies for mixed human-robot cooperative teams?
- Is the right to privacy threatened by personal robots accessing the internet?
- Are human linguistic abilities and culture impoverished by extensive interactions with robots which are linguistically less proficient than human beings?
- Can military robots be granted the licence to kill in the battlefield?
- Should one regard robots, just like human beings, as moral agents and bearers of fundamental rights?

These questions span human autonomy protection and promotion, moral responsibility and liability, privacy, fair access to technological resources, social and cultural discrimination, in addition to the ethical dimensions of personhood and agentivity. The conceptual and policy shaping challenges for applied ethics which arise from these questions call for an effective merging of multiple disciplinary perspectives. Notably, one has to take into account comparative cost-benefit analyses of robotic technologies with respect to alternative technologies, the projected impact of robotic technologies on the job market, psychological and sociological investigations on how HRI affects human conceptual structures, emotional bonds, and intercultural relationships,³ studies on the effects of deceptive simulation of intentional behaviours and emotions by edutainment and therapy-oriented robots, an understanding of delegation and trust relationships in human-robot cooperative decision making and action, in addition to analyses of the prospective impact of robotic technologies on developing countries and technological divides.⁴ This chapter examines distinctive roles that epistemological and methodological reflections on robotics play in ethical inquiries on robotics. In particular, methodological reflections on models of robot-environment interactions and on models of learning robotic systems are brought to bear on the analysis of autonomy and responsibility issues in robot ethics.

1. Robot Ethics And Models Of Robot-Environment Interactions

The observable behaviour of a robotic system results from the coordinated working of its bodily parts and their interactions with the environment. The contribution of the environment in shaping robotic behaviours can be hardly overestimated. The trajectory of a personal robot negotiating a living room surface can be significantly affected by changes of frictional coefficients only, such as those introduced by a Persian carpet or a glass of water spilled on the floor. The twisted paths traced on a beach by an insect-like

³ See the chapter by Toyooki Nishida in this book for a discussion of allegedly different cultural attitudes towards robots in Eastern and Western countries.

⁴ The multidisciplinary character of robot ethics is emphasized and extensively discussed by Christaller et al. (2001) and Veruggio & Operto (2008).

robot may result from the application of a uniform gait on the uneven and unsteady terrain.⁵ And dimming illumination conditions may hinder a robotic system in the task of sensing and properly reacting to nearby obstacles.

The problem of whether and how is one able to isolate environmental factors affecting robotic behaviours is crucial for robot ethics too. To illustrate, consider the prospective use of mobile robots as assistants to elderly or disabled people in their homes. In order to grant selling permissions and to shape suitable responsibility and liability guidelines, manufacturers of assistant robots will be asked to supply proper evidence that these robots can be safely operated in their intended operation environments. In particular, evidence must be provided for empirical statements of the form:

(S) Any exemplar of such-and-such robot model, when operating in normal conditions, will not cause serious damage to people, pets or inanimate objects.

Statements like (S) are universal statements, which express regularities concerning *every* run x of *any* exemplar y of some specified robot model in *each* normal operation environment z . But what is a normal operation environment for such robots? And which environmental factors may disrupt the intended behaviour of these robots? A relatively clear idea of what are normalcy conditions for robot operation is needed to assign precise meaning and testable empirical content to regularities that are expressed by means of statements of form (S).

The problem of identifying normalcy conditions for a regularity to hold has been extensively discussed in the philosophy of science, notably in connection with the formulation of regularities in biology and other areas of empirical inquiry falling outside the scope of fundamental physics. According to one prominent view, normalcy conditions for a regularity to hold can be, at least in principle, precisely and exhaustively stated: the “disturbing” factors and exceptions to such regularities are expressible by means of a finite list of conditions

$$C_1, \dots, C_n.$$

Accordingly, every regularity P which admits exceptions (abnormal conditions in which the regularity does not hold) can be replaced, in principle, by an exceptionless regularity of the form:

$$P \text{ unless } C_{p^1}, \dots, C_{p^n}$$

According to a different view, exceptions or disturbing factors may come in such large or even unbounded numbers, that one is not always in the position of completing P and turning it into an exceptionless regularity. This inability may depend on contingent reasons (exceptions come in large numbers) or on essential reasons (there are unboundedly many kinds of exceptions). This epistemic predicament is acknowledged

⁵ Herbert Simon used the example of an ant threading on a beach in order to convey the idea that environmental rather than internal control factors are often the chief source of behavioural complexity in natural or artificial systems (Simon, 1996, pp. 51-52).

by adding a *ceteris paribus* clause to P : “*Ceteris paribus, P*” means that P holds in the absence of unspecified (and possibly unspecifiable) disturbing factors.⁶

Robotic engineers are well aware of the theoretical and practical difficulties surrounding the epistemic problem of identifying environmental disturbing factors which jeopardize the normal working of robotic systems. A heuristic strategy which is often applied to address this epistemic problem is to make concrete environments in which robotic systems will be immersed as similar as possible to ideal situations in which suitable behavioural regularities are believed to hold. This heuristic strategy relies on the fact that any consistent (non-contradictory) set of statements T admits a model, in the usual sense of the word 'model' which is adopted in mathematical logic. More specifically, if T is a consistent set of sentences, then there is an interpretation of T , relative to some domain of objects and relations, which makes true all the sentences of T . This heuristic strategy can be rationally reconstructed as a process involving two main steps.

(a) Idealized domains of objects and relations are introduced, such that the desired regularities concerning robotic behaviours are true when interpreted in those domains;

(b) the concrete environments in which robots will be actually immersed are modified in order to make them as similar as possible to the idealized domains introduced in the previous step.

In industrial automation, one extensively applies the above strategy in order to enforce ideal-world conditions which exclude “disturbing factors” while preserving task compliance conditions. In particular, since human workers are a major source of dynamic changes and disturbing factors impinging on industrial robot behaviours, a robot “segregation” policy is usually pursued to achieve quasi-static and more easily predictable robot environments: factory workers and robots are often confined to different workspaces, and their mutual interactions are severely limited or altogether excluded.

A robot segregation policy becomes increasingly difficult to pursue as one moves from industrial robotics to applications of service and personal robotics in environments that are specifically designed for human activities. The task of delivering letters or documents to different rooms of an office building does not require, in principle, any purposive interaction with human beings. However, a mobile robot negotiating the corridors of an office floor is likely to encounter employees or visitors on its way to offices and mail delivery rooms. In these circumstances, an “unsociable” robot policy is a *prima facie* appealing alternative to the segregation policy for the purpose of limiting robot-environment interactions. Indeed, an unsociable robot is endowed with, and single-mindedly exercises the capability to avoid contact with any human-like object.

The unsociable robot policy places the entire burden of ensuring safety conditions on the robot control system. Even though the overall rule governing the behaviour of an unsociable robot is relatively easy to state, its actual design and implementation raises non-trivial theoretical and technological problems, notably including provisions for

⁶ For a discussion of *ceteris paribus* clauses and related predictive and explanatory issues concerning regularities which admit exceptions, see Earman, Glymour & Mitchell (2002).

real-time reactivity and motion planning in high-dimensional configuration spaces. Furthermore, the effectiveness of the proposed solutions tends to decline sharply as the environment becomes an increasingly cluttered and dynamic one. Finally, one should be careful to note that the unsociable robot policy is inapplicable when task specifications require extensive forms of HRI. Indeed, a robot which is programmed to avoid contact with any human being is unfit to rescue people or assist elderly people in their homes.

Both segregation and unsociable robot approaches are, in general, unsuitable for the operational contexts of service and personal robotics, insofar as many service and personal robots must be capable of rich and flexible forms of HRI in environments inhabited by a wide variety of animate and inanimate objects. On account of this fact, the modelling of robot-environment interactions tends to become more complicated as one moves from industrial robotics towards the current frontiers of field and service robotics. Major modelling challenges in field and service robotics concern the problem of providing precise formulations of normalcy conditions involved in statements of form (S), the problem of assigning a precise empirical content to the regularities expressed by means of those statements, and the problem of submitting these regularities to severe empirical tests.

Epistemological problems concerning the modelling of rich robot-environment interactions significantly bear on an ethical assessment of robotic systems. Notably, if robot-environment interactions involve rich forms of HRI, and the available models are only poorly predictive of actual robot behaviours, then one is hardly in the position of formulating precisely and severely testing properties of robotic behaviours that are relevant to autonomy, responsibility, and liability issues. This predicament is exemplified in the next section by reference to envisaged military applications of autonomous robotic systems in the battlefield.

2. Robot Soldiers: Task Requirements and Ethics

Thousands of military robotic systems have been deployed in Afghanistan and Iraq. These robots include the remote-controlled PackBot system, which enables one to detect and detonate improvised explosive devices (IED), and the Talon SWORDS, another remote-controlled robot deployed in the second Iraq war, which can be equipped with machine guns and grenade launchers (Forster-Miller 2009).⁷ The latter kind of military robot was presented as a significant step towards the development of robot soldiers in an article appearing on the front page of the *New York Times* on February 16, 2005:

The American military is working on a new generation of soldiers, far different from the army it has. “They don’t get hungry,” said Gordon Johnson of the Joint Forces Command at the Pentagon. “They’re not afraid. They don’t forget their orders. They don’t care if the guy next to them has just been shot. Will they do a better job than humans? Yes” The robot soldier is coming.

⁷ For ethical reflections on unmanned robotic vehicles developed for military purposes, see the chapter by Jürgen Altmann in this book.

The meaning and the truth-conditions of this allegedly apodictic conclusion are rather obscure, pending an answer to the following questions:

1. What is a robot soldier? Does a remote-controlled robotic system, such as the Talon SWORDS, qualify as a robot soldier?
2. What does it take for a robot soldier to do a better job than a human soldier? Which behavioural features can be sensibly used as a basis to compare and rank the performances of robot soldiers and human soldiers?

In connection with question 1, one should be careful to note that a Talon SWORDS robot is a remote-controlled system, and therefore all firing decisions are taken by its human controller. Accordingly, if one requires that any soldier be capable of taking autonomous firing decisions, then no remote-controlled robot qualifies as a robotic soldier.

In connection with question 2, it is taken for granted here that a “good” soldier, whatever it is, must behave in the battlefield in accordance with international humanitarian law, including internationally recognized treaties such as the Geneva and the Hague Conventions, in addition to the “code of ethics”, if any, and the rules of engagement (ROE) adopted by its own army. This broad requirement suggests that ethical reflection is needed to understand what it takes to be a good robotic soldier, and which behavioural tests must be passed to qualify as a good robotic soldier.

In an article appearing on the front page of the *International Herald Tribune* on November 26, 2008, some researchers in robotics suggested that intelligent robots which are *autonomous* in their firing decisions – that is, robots that are not controlled by external agents as far as firing decisions are concerned – will eventually behave “more ethically” than human soldiers in the battlefield.

“My research hypothesis is that intelligent robots can behave more ethically in the battlefield than humans currently can”, said Ronald Harkin, a professor at the Georgia Institute of Technology who is designing software for battlefield robots under contract from the U.S. Army.⁸

The guess that robot soldiers will outperform in the battlefield human soldiers as far as their moral behaviour is concerned does not conflict with present scientific knowledge at large. However, one should be careful to note that the process of turning this guess into a serious technological possibility requires substantial – and presently unwarranted – technological advances in robotics. To illustrate, consider the capabilities of (a) recognizing surrender gestures and (b) telling bystander or friend from foe, which are often discussed in connection with the prospects of robotic warfare (Sharkey 2008). In order to make a good showing in any behavioural test designed to control whether one is capable of providing satisfactory (albeit not infallible) human-level solutions to problems (a) and (b), a robotic system must possess a wide variety of perceptual and deliberative capabilities that are well beyond state-of-art artificial intelligence and cognitive robotics. Human-level solutions to problems (a) and (b) are issued on the basis of context-dependent disambiguation of surrender gestures, even when these are rendered in unconventional ways, understanding of emotional expressions, real-time reasoning about deceptive intentions and actions. Accordingly, the perceptual

information which has to be extracted from sensory data and contextually evaluated in order to make the right decision is extremely varied and open-ended. And the stereotyped sensing, information processing, and acting conditions that have been enforced in successful applications of robotic technologies up to the present day cannot be enforced in these unstructured perceptual and decision-making environments. Indeed, the variety of cognitive processing abilities and background knowledge that are jointly required to provide human-level solutions to problems (a) and (b) are such that their implementation in a robotic system will pave the way to solving any other problem that artificial intelligence and cognitive robotics will ever be confronted with. These problems, by analogy to familiar classifications of computational complexity theory, are aptly called “AI-complete problems” (Cordeschi & Tamburrini, 2006).

State-of-art artificial intelligence and cognitive robotics hardly provide any significant cue towards the solution of AI-complete problems in general, and towards the solution of problems (a) and (b) in particular.⁹ But the possession of abilities enabling one to solve AI-complete problems can make all the difference between a robot behaving in accordance with internationally recognized rules of *jus in bello* and a robot waging massacre against the innocent. Thus, the possession of abilities that are well beyond state-of-art robotics and artificial intelligence is a central requirement for a “good” autonomous firing robot. And clearly, no “good” autonomous firing robot is in the purview of imminent developments of robotic technologies.

It is worth noting that the proposed requirement of affording human-level, albeit non-infallible solutions to problems (a) and (b) makes sense from both teleological and consequentialist ethical theorizing standpoints. To being with, let us note that the intentional killing of the innocent is usually regarded as an absolute prohibition from teleological theorizing perspectives. Is this prohibition violated if an autonomous robot which failed tests for human-level solutions to (a) and (b) kills the innocent in the battlefield? If the killing of the innocent is not an intentional consequence of the decision to deploy such robots in the battlefield, then one can make appeal to the principle of double effect in order to classify as mere “collateral damages” events that are morally blameworthy when brought about intentionally. However, it is doubtful that the required precondition for a sound application of the double effect principle is satisfied in this circumstance, insofar as both programmers and military commanders *knew* in advance that robot soldiers deployed in the battlefield failed to provide human-level solutions to problems (a) and (b). Since human soldiers are less likely to kill the innocent than these robot soldiers, then there are available precautions, which have not been taken, to avoid the killing of the innocent, and to achieve one’s own military goals in accordance with recognized rules of *jus in bello*.

The proposed requirement should be acceptable from consequentialist standpoints in ethics too, as soon as a sufficiently broad temporal horizon for action consequences is duly taken into account. While the use of robot soldiers may bring about short-term advantages in warfare, especially by reducing the number of casualties on one’s own side, the killing of the innocent by robots is likely to induce long-term resentments in

⁹ The robotic systems one can build on the basis of state-of-art robotics and its foreseeable developments are not better combatants than human soldiers on many other accounts as well, insofar as these systems do not possess the real-time reactivity, motion planning, and goal-seeking abilities that human soldiers are trained to develop in highly dynamic and fairly unpredictable environments. The theoretical and technological tools developed by state-of-art cognitive robotics and artificial intelligence are insufficient to address the wide variety of competitive and cooperative HRI problems arising in those environments.

the opposite side. These long-term consequences should be taken into account in order to reduce expected losses of human lives, the length of the conflict, and the insurgence of new motivations for another conflict. On the whole, the proposed requirement appears to be acceptable from a variety of prominent ethical theorizing perspectives, and no robotic system in the purview of foreseeable technological developments is likely to be positively judged by their light.

Increasing expectations about the promise of robotic technologies have been fuelled by rapid developments of robotic research during the last two decades. Robotic technology, however, is not a panacea for the humankind. There are imagined robotic scenarios, masterly illustrated in literary and cinematographic explorations of the theme of robotic agency,¹⁰ which transcend both the horizon of state-of-art robotics and its foreseeable developments. The process of turning these scenarios into serious technological possibilities requires substantial – and presently unwarranted – technological advances in robotics. An autonomous firing robot which is capable of behaving in accordance with humanitarian law and rules of engagement, at least to the extent of matching the performances of good human soldiers, is not coming soon. More than occasionally, popular reports on cutting edge robotics research fail to draw a responsible distinction between imminent technological developments on the one hand and distant imagined scenarios on the other hand.

The identification and analysis of ethical issues concerning robotic technologies and systems that are more likely to have an impact on our lives presupposes some understanding of what current robotic technologies are (not) likely to deliver in the near future. A failure to draw this distinction results into an (often but not invariably unintentional) screening effect on social and ethical issues concerning imminent technological developments, whereby public opinion is induced to believe that ethically crucial issues have been solved, thus requiring no further reflection and democratic awareness. In this section, this screening effect was contrasted by a confluence of ethical reflection with epistemological appraisals of robotic models. Robotics is a long way from solving the problem of designing an autonomous firing robot whose behaviour complies with humanitarian law at least as well as the behaviour of a good human soldier. Accordingly, the prospect of introducing in our armies autonomous firing robots must be responsibly rejected, at least until the involved AI-complete problem will have been properly solved.

Let us now turn to explore in the next section another facet of the potential contribution of philosophy of science to ethical reflections on robotics, by drawing on the epistemological analysis of inductive learning processes for robotic systems.

3. Autonomy And Responsibility Issues For Learning Robots

It was pointed out that the environments in which robots are supposed to act become more dynamic as one progressively moves from industrial robotics towards the current frontiers of field and service robotics. It turns out that robot designers are not always in the position to identify, describe in sufficient detail, and implement control policies that are adequate to achieve reactive and sufficiently flexible robotic behaviours in environmental conditions which differ to such a great extent from standard industrial

¹⁰ For a recent comparative discussion of various kinds of agency, including robotic agency, see Capurro (2009).

robotics environments. This epistemic limitation provides a strong motivation for endowing service robots in general, and personal robots in particular, with the capability of learning from their experience, insofar as learning is a powerful source of adaptation in dynamic environments. Thus, instead of furnishing robots with detailed information about regularities present in their operation environments, robot designers endow robots with computational rules enabling one to discover these regularities.

Without loss of generality, a computational agent that learns from its experience can be viewed as an algorithm that looks for regularities into a representative (input) dataset, and subsequently uses these regularities to improve its performances at some task. Learning of this kind does not take place in a vacuum: any attempt to identify regularities that are possibly present into a dataset must rely on some pre-existing “structure” on the part of the computational agent. Such structure may involve the use of built-in preferences or “biases” concerning the class of functions (hypotheses) from which the target regularity must be selected. Learning agents usually rely on additional priori expectations about the unknown target regularity in order to narrow down their search space. A straightforward example of background conjectural assumption which learning agents use to downsize search spaces is expressed in a procedural form by the rule of choosing “simpler” hypotheses that are compatible with observed data.¹¹ Therefore, various priori assumptions about the regularities that have to be discovered in the environment play a crucial role in machine learning strategies.

The evaluation of the correctness of learning processes carried out in the mathematical framework of computational learning theory relies on various background assumptions too. Notably, computational learning theory aims at establishing the existence of probabilistic bounds on learning errors for given learning problems under the empirical assumptions that (a) one is dealing with a well-defined stochastic phenomenon characterized by a fixed statistical distribution, and that (b) training examples are independently drawn from this fixed statistical distribution.

According to both machine learning approaches and mathematical theories of computational learning, the conjecture that a learning processes has been successfully carried out by a computational learning agent relies on various background hypotheses about the relationship of training data to target functions. Since these background hypotheses are fallible, one cannot exclude with certainty that the learning agent will perform poorly on as yet unobserved data. This is the point where machine learning and theories of computational learning meet the philosophical problem of induction, which is usually construed in philosophy of science and theory of knowledge as the problem of providing a justification, if any, for the background assumptions used in inductive reasoning.¹² As we shall presently see, these epistemological reflections on artificial learning agents bears in distinctive ways on robot ethics.

Human users of personal robotic assistants will delegate the execution of some repertoire of actions to these systems as a means to fulfil their intentions. Delegation and transfer of action control to learning robots will be, in many prospective applications of personal robotics, traded off for greater autonomy by their human users. For example, elderly or disabled people will commit to a learning robotic system the execution of actions in order to achieve a restored procedural capability to fulfil their desires. Is this expectation going to be *invariably* fulfilled? One can reasonably doubt

¹¹ This rule is just a special instance of the methodological maxim known as Ockham’s razor.

¹² For discussion, see Tamburrini (2006); for an analysis of early cybernetic reflections on ethics and the use of learning machines, see Cordeschi & Tamburrini (2006).

that this is going to be the case, in view of the fact that programmers, manufacturers, and users of learning robots may not be in the position to predict exactly and certify what these machines will actually do in their intended operation environments. If a learning robot was sold in a shop, it is unlikely that its user manual will contain a statement to the effect that the robot is guaranteed to behave so-and-so if normal operational conditions are fulfilled. Since one cannot be sure that the actions undertaken by a learning robot invariably correspond to the intentions of its users, there are conceivable circumstances in which the autonomy of these users is jeopardized and their intentions betrayed. Moreover, some of the epistemic “errors” committed by a learning robot may harm its users and bring about additional sorts of damaging events. Under the epistemic predicament affecting programmers, manufacturers, and users of learning robots, who is responsible for the harmful errors made by a learning robot? This is, in a nutshell, the responsibility ascription problem for learning robots.

An initial move which contributes to set up an appropriate conceptual framework for addressing problems of this kind is to distinguish between liability or objective responsibility on the hand, and moral responsibility on the other hand. A variety of technical tools have been put in place, during the historical development of moral doctrines and legal systems, to deal with similar objective or liability ascription problems. Our inability to predict exactly and control the behaviour of learning robots is closely related, from ethical and juridical perspectives, to the inability of legal owners of factories to prevent every possible damage caused to or by factory workers. Moreover, in view of the fact that training and learning are involved in these robot-environment interaction contexts, this inability in the way of prediction and control is meaningfully related to the inability of controlling the behaviour of learning biological systems – say, the inability of dog owners to curb their pets in every possible circumstance, and even the inability of parents to predict and exert full control on the behaviour of their children. The information processing abilities of learning robotic systems, whose behavioural effects are not fully predictable by their users, suggest juridical and ethical analogies between learning robots and learning biological systems that are capable of perceiving, planning and acting. Interestingly, these analogies appear to be more stringent than juridical analogies with responsibility and liability problems arising in connection with the use of other kinds of machinery.

In view of these informative analogies, only relatively small adjustments of extant ethical and legal frameworks appear to be needed in order to cope with moral responsibility and liability ascription problems arising in connection with damages caused by actions of learning robots. Liability problems do not, in general, allow one to identify in a particular subject the sole or even just the main origin of causal chains leading to a damaging event. Thus, in addressing and solving these problems, one cannot rely uniquely on such things as the existence of a clear causal chain or the awareness of and control over the consequences of actions. In some cases, ascribing responsibility for damages caused by the actions of a learning robot, and identifying fair compensation for those damages may require an approach which combines moral responsibility and liability considerations. Producers or programmers who fail to comply with acknowledged learning standards, if any, in setting up their learning procedures are morally responsible for damages caused by their robots. This is quite similar to the situation of factory owners who fail to comply with safety regulations or, more controversially, with the situation of parents and tutors who fail to provide adequate education, care, or surveillance. These parents and tutors are regarded, on

account of their negligent behaviour, as both objectively and morally responsible for offences caused by their young.

The responsibility problems examined in this section are all *retrospective* responsibility attribution problems - insofar as the attribution of objective or moral responsibility for past events only is involved here. But what about the *prospective* responsibility that we, as a society, should assume with respect to learning robots? Why do we want to introduce learning robots in, say, our homes and offices? The epistemological reflections presented in this section contribute to outline a general methodological framework for the cost-benefit analysis which is required to address and answer prospective responsibility issues concerning learning robots. This general framework must be adapted to each specific context in which the use of a learning robot is envisaged, especially by considering whether and how harmful consequences may arise from epistemic errors made by this learning robot.

4. Robotic myths, ethics, and scientific method

Current developments of robotic technologies raise great expectations about the extension of human capabilities and the improvement of many aspects of human life, including freedom from repetitive jobs and fatigue, more precise and less invasive surgery, effective assistance for elderly and disabled people, new forms of support in education and entertainment. These expectations about robotic technologies fall squarely in the wake of the classical view of technological progress put forward by Bacon and Descartes. This view rehearses, in terms that are more acceptable to modern sensibilities, the Promethean promise of compensating the deficiencies and extending the powers of human biological endowment by means of man-made tools and devices.

Robotics adds a very distinctive flavour to classical myths and modern age expectations about technology. Robots are very special kinds of machines: the coordinated unfolding of their motor, sensing, and information processing abilities enables one to achieve goal-oriented and adaptive behaviours. Until the rise of robotics, goal-oriented and adaptive behaviours were, by and large, an exclusive prerogative of biological systems. For this reason, the human enterprise of robotics is more closely related than previous technological undertakings to mythical tales concerning the origin of animate beings from inanimate matter.¹³ The first human beings were assembled by divine entities from clay, fire, air, and other material constituents. Human beings are now able to assemble from inanimate matter entities manifesting adaptive and intelligent action. This manifestation of human ingenuity through robotic technologies and systems is, at the same time, a manifestation of human hubris, insofar as robotics allows human beings to usurp and arrogate to themselves a divine prerogative. Thus, robotics adds a new dimension to the Promethean association of sin and burglary to technological progress, insofar as robotics enables one to alter in distinctive ways the “natural order” and “deflect” natural processes towards the achievement of human goals.

The punishment delivered by gods to human beings for their Promethean burglary are the afflictions flowing from Pandora’s vase. In terms that are more palatable to

¹³ This special connection of robotics with mythology and religion was explored early on by Norbert Wiener in his book *God & Golem*, whose subtitle tellingly reads *A Comment on Certain Points where Cybernetics Impinges on Religion* (Wiener 1964).

contemporary sensibilities, this punishment is more appropriately identified with the human inability to predict completely the behaviour of technological devices, and to exert full control on their uses for the benefit of humanity. The image that Promethean myths provide us with is an image in which elements of promise and threat are deeply entangled in technological progress. Human reason is the chief - albeit admittedly imperfect and fallible - instrument that human beings can rely on to disentangle these various elements, and to govern the applications of technologies in the light of ethical reflection and moral deliberation. The methodological and epistemological reflections presented here suggest distinctive ways in which the philosophical analysis of science and technology contributes to disentangle some of these elements, for the purpose of promoting and protecting fundamental human rights in connection with the use of robotic systems.

Acknowledgments

I wish to thank Rafael Capurro, Edoardo Datteri, Jürgen Altmann, Michael Nagenborg, Toyooki Nishida, and Matteo Santoro for helpful comments and discussions on an earlier draft of this chapter. Research for this chapter was partially supported by the European Commission, under the VI. FP project “Ethicbots”.

Reference List

- Altmann, J. (2009). **TITLE, This book.**
- Arkin, R. (2007). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. *Technical Report GIT-GVU-07-11*, Georgia Institute of Technology.
- Capurro, R. (2009). *Towards a Comparative Theory of Agents*. Retrieved May 31, 2009, from: <http://www.capurro.de/agents.html>.
- Christaller, T., Decker, M., Gilsbach, J.-M., Hirzinger, G., Lauterbach, K., Schweighofer, E., Schweitzer, G., & Sturma, D. (2001). *Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft*. Berlin: Springer.
- Cordeschi, R., & Tamburrini G. (2006). Intelligent Machines and Warfare: Historical Debates and Epistemologically Motivated Concerns. In Magnani, L., & Dossena, R. (Eds.). *Computing, Philosophy and Cognition*. London: College Publications, pp. 1-20.
- Earman, J., Glymour, C., & Mitchell, S. (Eds.) (2002). *Ceteris Paribus Laws*. Dordrecht: Kluwer.
- Forster-Miller (2009). Products & Services. TALON Family of Military, Tactical, EOD, MAARS, CBRNE, Hazmat, SWAT and Dragon Runner Robots. Retrieved on January 24th, 2009, from: <http://www.foster-miller.com/lemming.htm>.
- Nishida T. (2009). **TITLE, This book.**
- Santoro, M., Marino, D., & Tamburrini, G. (2008). Learning robots interacting with humans: from epistemic risk to responsibility. *AI and Society* 22(3), pp. 301-314.
- Sharkey, N. (2008). Cassandra or False Prophet of Doom: AI Robots and War. *IEEE Intelligent Systems* 23(4), pp. 14-17.
- Siciliano, B., & Khatib, O. (Eds.) (2008). *Handbook of Robotics*. Berlin: Springer.
- Simon, H. (1996). *The Sciences of the Artificial* (3rd edition). Cambridge: MIT Press.
- Tamburrini, G. (2006). Artificial intelligence and Popper's solution to the problem of induction. In Jarvie, I.; Milford, K., & Miller D. (eds.). *Karl Popper: A Centenary Assessment. Metaphysics and Epistemology* (Vol. 2.). London: Ashgate, pp. 265-284.
- Veruggio, G., & Operto, F. (2008). Roboethics. In: Siciliano & Khatib (2008), pp. 1499-1524.
- Wiener, N. (1964). *God and Golem Inc*. Cambridge: MIT Press.