In R. Capurro and M. Nagenborg (eds), *Ethics and Robotics*, IOS Press, Amsterdam, 2009, pp. 11-22.

Robot Ethics: A View from the Philosophy of Science

Guglielmo Tamburrini[1]
Dipartimento di Scienze Fisiche, Università di Napoli Federico II
tamburrini@na.infn.it

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 the distinctive role that epistemological and methodological reflections on robotics play in roboethical inquiry, focusing in particular on the role that an epistemological appraisal of models of robotic behaviour plays in the analysis of autonomy and responsibility issues.

1. INTRODUCTION

Robots are machines endowed with motor, sensing, and information processing abilities. Information processing in robotic systems takes notably the form of feedback signal processing and control, in addition to perception, reasoning, planning, and

learning. The coordinated exercise of these abilities enables robotic systems to achieve goal-oriented

[1]Dipartimento di Scienze fisiche, Università di Napoli Federico II, e-mail: tamburrini@na.infn.it. I wish to thank Rafael Capurro, Edoardo Datteri, Jürgen Altmann, Michael Nagenborg, and Toyoaki Nishida for helpful comments on an

and adaptive behaviours. Information and 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 in the context of heterogeneous teams of agents that are formed by human, robotic or software agents.

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 robotics 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 the 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.

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. To 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 relationships and linguistic abilities impoverished by extensive interactions with robots replacing humans in assistance and edutainment tasks?
- Can robots be given the licence to kill in the battlefield?
- Are robotic systems, just like human beings, moral agents and bearers of fundamental rights?

These questions span human autonomy protection and promotion, moral responsibility and liability,

---

earlier draft of this chapter.

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 their alternatives, 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, studies on the effects of deceptive simulation of biological systems 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.[2] This chapter examines the distinctive role that epistemological and methodological reflections on robotics play in robot ethics, focusing on the role that an epistemological appraisal of models of robotic behaviour plays in the analysis of autonomy and responsibility issues.

2. Robot ethics and robot-environment interactions

The observable behaviour of a robotic system results from the coordinated working of its bodily parts and their interaction with the environment. The contribution of the environmental context 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 coefficient 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 robot may result from the application of a fairly uniform gait on an uneven and unsteady sandy surface.[3] And variable illumination conditions may hinder or facilitate an outdoor robotic system in the task of identifying and properly reacting to nearby obstacles.

[2]The multidisciplinary character of robot ethics is extensively examined in (Christaller et al. 2001) and (veruggio and Operto 2008).

[3]Herbert Simon used the example of an ant threading on a beach to explore the idea that environmental rather than internal control factors are often the chief source of complexity in the behaviour of natural or artificial systems. See in particular

Isolating the 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 normal operation environments. In particular, evidence must be provided for empirical statements of the following form:

> (S) Any assistant robot of model R, when operating in normal conditions, will not cause serious damage to people, pets or inanimate objects.

Statements of this form are universal statements, which express regularities concerning *every* run *x* of *any* R-robot *y* in *each* normal operation environment *z*. But what is a normal operation environment for such robots? And which environmental factors affect the behaviour of these robots? A relatively clear idea of what are normalcy conditions for robot operation is needed to assign a 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, especially 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 in principle precisely and exhaustively stated: the "disturbing" factors or exceptions to such regularities are expressible by means of a finite list of conditions C_1, ..., C_n. Accordingly, every regularity *P* which admits exceptions (abnormal conditions in which the regularity does not hold) can be, in principle, replaced by an exceptionless regularity of the form "*P* unless C_{p1}, ..., C_{pn}". This view and the epistemic solution it affords have been criticized on the ground that the kinds of exceptions or disturbing factors may be very large or even unbounded. As a consequence, no matter how one amends regularity *P*, this will continue to admit unspecified exceptions. This epistemic predicament is signalled

---

Simon 1996, pp. Xxxx.

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.

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 and circumscribe normal operation environments for robotic systems relies on suitable "closed-world" assumptions (cit.):

> (a) one models robotic behaviours in closed worlds that are predictable in their internal evolution on the basis of known and well-tested regularities;
>
> (b) one attempts to enforce, in the concrete environments in which robots will be actually immersed, the easily predictable conditions of those ideal closed worlds.

The heuristic strategy of enforcing closed-world assumptions and limiting robot-environment interactions without compromising task requirements has been extensively applied in industrial automation. Since human workers are usually a major source of dynamic changes that are difficult to predict in industrial environments, a robot "segregation" policy is conducive to enforce quasi-static and more easily predictable robot environments: either one confines factory workers and robots to different workspaces or else one limits and rigidly regiments their interactions there.

The 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, the unsociable robot policy is a *prima facie* appealing alternative to the segregation policy for the purpose of limiting robot-environment interactions: robots are endowed with, and single-mindedly exercise the capability

to avoid contact with any human-like object. This policy places the entire burden of ensuring safety conditions on the design of the robot control system. Even though the overall rule governing the behaviour of an unsociable robot is relatively easy to state, its actual implementation raises non-trivial theoretical and technological problems. These problems include suitable provisions for 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 increasingly cluttered and dynamic.

The unsociable robot policy is unfit when task specification requires 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, inadequate to simplify the environmental context of service and personal robots, insofar as these robotic systems must be capable of engaging themselves into rich and flexible forms of HRI within environments that are populated by a wide variety of animate and inanimate objects. Accordingly, the modelling of robot-environment interactions tends to become more complicated, both theoretically and practically, as one moves from industrial robotics towards the current frontiers of service and personal robotics. These modelling challenges concern, in particular, the problem of providing precise formulations of normalcy conditions involved in statements of form (S), which express safety conditions for robotic systems, 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.

The modelling of rich robot-environment interactions, possibly involving flexible forms of HRI, raises serious conceptual and empirical challenges. If, due to rich and unregimented robot-environment interactions, the available models are only poorly predictive of robot behaviours, one is hardly in the position of formulating precisely and severely testing safety conditions and other properties of robotic behaviours that are relevant to autonomy, responsibility, and liability issues. This predicament is most

vividly illustrated by reference to envisaged military applications of autonomous robotic systems in the battlefield.

3. Robot soldiers: task requirements and ethics

Present military applications of robotics include the remote-controlled, de-mining robotic system PackBot[4], equipped with cameras, communication equipment, and manipulators. The PackBot has been used to detect and detonate improvised explosive devices (IED) in the second Iraqi war. The Talon SWORDS, another remote-controlled robot deployed in the second Iraqi war, can be equipped with machine guns and grenade launchers. These remote-controlled military robots were 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. gThey don't get hungry, h said Gordon Johnson of the Joint Forces Command at the Pentagon. gThey'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. h The robot soldier is coming.

The meaning of this allegedly apodictic conclusion is not quite clear, pending an answer to a wide variety of hard questions, notably including the following ones:

- What is a robot soldier? In particular does a remote-controlled robotic system, such as the Talon SWORDS, qualify as a robot soldier?
- What does it mean to assert that robot soldiers will globally do a better job than humans?
- What are the significant properties on the basis of which the performances of robot soldiers should be evaluated against the performances of human soldiers?

Let us focus on this latter question. A "good" soldier, whatever it is, must behave in the battlefield in accordance with humanitarian law, and to abide by internationally recognized rules of *jus in bello*, such as those included in the Geneva Conventions. Thus, ethical reflection is needed to clarify what it takes to qualify as a good robotic soldier, and which tests must be passed in order to be recognized as a good robotic soldier.

[4]REF. Even cruise missiles should be counted as military applications of robotics insofar as perceptual feedback is crucially involved in their trajectory control system.

Recently, in an article appearing on the front page of the *International Herald Tribune* on November 26, 2008, it has been suggested that autonomous, intelligent robots will behave in the battlefield better than human soldiers even when evaluated from an ethical perspective.

> "My research hypothesis is that intelligent robots can behave more ethically in the battlefield than humans currently can□h 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.

This guess does not conflict with present scientific knowledge at large. However, 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 bystanders apart from foes. Independently of what are the details of any sensible behavioural test one may conceive for a robot to qualify as a system capable of providing satisfactory solutions to problems (a) and (b), it is quite clear that the problem of implementing a robotic system capable of passing such test is well beyond state-of-art artificial intelligence and cognitive robotics. Indeed, solving these perceptual discrimination problems involves context-dependent disambiguation of surrender gestures, even when these are rendered in unconventional ways, an understanding of emotional expressions, real-time reasoning about deceptive intentions and actions, and so on. Briefly, the contextual information which has to be properly sensed and evaluated in order to make the right decision is extremely dynamic and open-ended, especially when compared to any of the environments in which robotic technologies have been successfully embedded up to the present day. Accordingly, the variety of mental processing abilities and background knowledge that are jointly required to provide human-level solutions to these problems 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, may be aptly called "AI-complete problems".

State-of-art artificial intelligence and cognitive robotics hardly provide any significant cues towards the solution of AI-complete problems, including the solution to problems (a) and (b).[5] And yet, autonomous firing robot soldiers failing to meet the human-level abilities needed to solve (a) and (b) are not less likely to kill the innocent than a human soldier. Accordingly, the possession of these discrimination abilities can be sensibly advanced as a central criterion for an ethical evaluation of prospective military applications of robotics in the battlefield, insofar as the possession of these abilities can make the difference between a robot behaving in accordance with internationally recognized rules of *jus in bello* and a robot and waging massacre against the innocent.

It is worth noting that the proposed criterion makes sense from a variety of prominent ethical theorizing standpoints. To being with, the killing of the innocent is regarded as an absolute prohibition from some teleological perspectives in ethics. Moreover, it is doubtful that the preconditions for a sound application of the double effect principle are satisfied if robot soldiers failing to meet the proposed criterion are knowingly deployed in the battlefield. An appeal to this principle is often made in warfare in order to label as a side effect or "collateral damage" an event which is not permissible to bring about intentionally, such as the killing of children. If one knows that a system deployed in the battlefield is unable to discriminate between the innocent and the enemy, then one no longer has a rational basis for distinguishing between the goals one pursues by deploying these robotic systems in the battlefield and their alleged side effects. Finally, the proposed criterion should be acceptable from consequentialist standpoints in ethics too: while the killing of the innocent may bring short-term advantages in a war, it is likely to induce long-term resentments in the enemy, whose expected consequences should be properly taken into account in order to minimize the loss of human lives, the length of the conflict, and so on. In conclusion, the proposed criterion appears to be acceptable from a

[5]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 environments that are highly dynamic and fairly unpredictable in their evolution. The theoretical and technological tools developed by state-of-art cognitive robotics and artificial intelligence can be hardly sufficient to address the wide variety of

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.

Raising expectations about robotic technology and its promise for improving many aspects of human life 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 artistic explorations of the theme of robotic agency, which transcend 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. Popular reports on cutting edge robotics research fail more than occasionally to draw a clear and responsibly made distinction between present or imminent technological developments on the one hand and these distant scenarios on the other hand. Belittling the real limitations of current robotic technologies prevents one from appreciating the real ethical issues at stake. The reasons underlying this (often but not invariably unintentional) screening effect are a research theme pertaining the sociology of science and technology more than the philosophy of science proper. But the broad methodological and epistemological reflections on robotic modelling which are in the purview of philosophy of science can effectively contrast this screening effect, thereby contributing - among other things - to re-establish a more balanced epistemic basis for ethical reflection on robotics.

4. Learning robots and epistemology

The problem of designing adequate control policies for robotic systems immersed in dynamic environments must be properly addressed in order to develop a wide variety of military and non-military applications of robotic technologies alike. Indeed, it was pointed out above that the environments in which robots are supposed to act become more dynamic and less readily predictable as

---

competitive or cooperative HRI problems arising in those environments.

one progressively moves from industrial robotics towards the current frontiers of service and personal robotics. Nevertheless, some possibly unknown regularities must be present in these dynamic environments, for the successful action of any agent, including service and robotic systems, is based on a wide variety of expectations and assumptions about persisting features of the environment and the causal factors determining changes there.[6] Some of these assumptions are procedurally built into robotic control systems. Other assumptions about environmental regularities are explicitly represented for use in robot deliberative processes. These various assumptions may concern expectations about environment topology (a planar office, say, rather than a 3D uneven terrain), patterns or objects that the robot is likely to detect there, fixed interaction schemata with other agents, expectations about the outcome of one's own action or the action of other agents. In charting a territory, for example, a robotic system usually acts on the hypothesis that map topology does not change too often or too drastically, insofar as previously identified landmarks are relied on for further exploration, map-building, and planning.

Robot designers may not be in the position to isolate, describe in a detailed fashion and furnish a robotic system with some hypotheses about the environment which are needed to achieve reactive and flexible goal-directed behaviour in dynamic HRI conditions. This epistemic limitation provides a strong motivation for endowing service and personal robots 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 cannot take

[6]In artificial intelligence, the general problem of isolating a set of assumptions and expectations enabling an intelligent artifact to cope effectively with its environment is called the frame problem.

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 some built-in "bias" or some marked out repertoire of functions (hypotheses) by means of which to represent the target regularity[7] of the environment Thus, a priori assumptions about the regularities . that have to be discovered play a crucial role in machine learning strategies as well. Learning agents usually rely on additional priori expectations about the unknown target regularity in order to narrow ghtforward example of background conjectural assumption which down their search space. A strai learning agents use to downsize search spaces is expressed in a procedural form by the rule of choosing Successful learning from experience by [8] .atad devresbo htiw elbitapmoc era taht sesehtopyh "elpmis" machine learning methods depends crucially on the correctness of these empirical background .assumptions

Computational learning theory addresses the problem of evaluating the correctness of learning processes in a rigorous mathematical framework, insofar as it aims at establishing the existence of probabilistic bounds on learning errors for given learning problems. One should be careful to note, however, that these mathematical proofs depend on various background assumptions, notably including the hypothesis that one is dealing with a well-defined stochastic phenomenon characterized by a fixed statistical distribution, and that training examples are independently drawn from this background statistical distribution.

In conclusion, the conjecture that a learning processes has been successfully carried out relies, according to both machine learning approaches and mathematical theories of computational learning, on various background hypotheses about the relationship of training datasets to target functions. Thus, the expectation that robotic systems will achieve better performances by learning is contingent on the correctness of these various assumptions. Hence, a poor approximation of the target function on

[7] This sweeping claim is clearly stated and motivated in Cucker and Smale 2001. Mitchell 1997 (p. 42ff.) is also a valid source for a discussion of inductive biases needed by computational learning agents.

unobserved data cannot be excluded with certainty, insofar as a good showing of a learning algorithm at future outings depends on these fallible background hypotheses. This is indeed the point where machine learning and theories of computational learning meet the philosophical problem of induction, which is usually construed in the philosophy of science and the theory of knowledge as the problem of providing justifications, if any, for the background assumptions used by inductive rules.[9]

These epistemological reflections on the fallibility of the empirical background assumptions used in computational learning play a significant role in the analysis of responsibility and autonomy issues arising in human-robot interaction contexts. Let's see.

5. Learning robots: autonomy and responsibility issues

Learning procedures have been examined in the previous section as enabling factors which are conducive to achieve better performances of robotic systems in dynamic HRI environments. Learning robots, if any, will be introduced in those environments in view of their ability to pursue goals that are endorsed by their human users. In particular, human users of robotic assistants will be willing to delegate the execution of some repertoire of actions to these systems as a means to fulfil their intentions. These acts of delegation and transfer of action control to learning robots are, in many prospective applications of service and personal robotics, traded off for greater autonomy by human users. Elderly or disabled people commit to a learning robotic system the execution of actions reflecting their intentions in order to achieve a restored procedural capability to attain goals corresponding to their own desires. But do they?

The above epistemological reflections on computational learning theories and machine learning methods suggests 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. Thus, in particular, if a learning robot was sold in a shop, it is unlikely that user manuals

[8]This rule is just an instance of the methodological maxim known as Ockham's razor.
[9] For discussion, see Tamburrini 2006; for an analysis of early cybernetic reflections on ethics and the use of learning

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.

Some of the epistemic "errors" committed by a learning robot may harm its users or bring about different sorts of damaging events. Under the epistemic predicament affecting programmers, manufacturers, and users of learning robots who are not in the position to predict exactly and certify what these machines will actually do in their intended operation environments, who is responsible for the harmful errors made by a learning robot? This is, in a nutshell, the responsibility ascription problem for learning robots.

The familiar move which enables one to address 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 conceptual and technical tools have been put in place, during the historical development of ethical 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 an ethical and legal perspective, 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 HRI contexts, this inability is meaningfully related to the inability of dog owners to curb their pets in every possible circumstance, and even to the inability of parents to predict and exert full control on the behaviour of their children. Interestingly, the information processing abilities of robotic systems, whose behavioural effects are not fully predictable by their users, suggest the opportunity of assimilating these systems more to biological systems capable of perceiving, planning and acting than to other kinds of machines.

Liability problems do not, in general, allow one to identify in a particular subject the sole or

---

machines, see Cordeschi and Tamburrini 2005.

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, and on account of this fact are regarded as both objectively and morally responsible for offences caused by their young.

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 some action of learning robots, and the distribution of compensation for those damages. These are all *retrospective* responsibility ascription problems, concerning attributions of objective or moral responsibility for past events. But what about *prospective* responsibilities concerning learning robots? Who are the main actors of the process by which one shapes appropriate criteria for deciding whether some learning robots should be ushered in our societies? A partial answer to this question has been outlined in this section: the various disciplinary perspectives which have to be merged into this process must include epistemological reflections on both computational learning frameworks and the background assumptions that are used to solve specific learning problems.

6. Conclusions: robotics, scientific method, and myth

Current developments of robotic technologies raise expectations concerning the extension of human capabilities and the improvement of many aspects of human life, including freedom from repetitive

jobs and fatigue, extension of human capacities for high precision tasks, more effective assistance for elderly and disabled people, new kinds of companions in education and entertainment. These expectations about robotic technologies are coherent with a classical view about the role of technology in general, put forward by Bacon, Descartes, and the Enlightenment thinkers. This view rehearses, in terms that are more acceptable to contemporary sensibilities, the Promethean promise of compensating the deficiencies and extending the powers of human biological endowment by means of technical tools and technological devices. Robotics, however, adds a very distinctive flavour to these mythological views and classical expectations about technology at large.

It was pointed out that the machines that robotics endeavours to build are very special ones, insofar as these machines are endowed with motor, sensing, and information processing abilities, whose coordinated unfolding enables robotic systems to manifest goal-oriented and adaptive behaviours. Until the rise of robotics, behaviours of these sort were mostly or even exclusively manifested, in various degrees, by biological systems. For this reason, the connections between robotics and myth reach out into territories that have gone uncharted so far in the history of technology. In particular, robotics is meaningfully related to the mythical tales concerning the origin of animate beings from inanimate matter, by the assembly of clay, fire, air and other substances. This connection of robotics with mythology and religion was explored by Norbert Wiener in his book *God & Golem*, whose subtitle tellingly recites "A Comment on Certain Points where Cybernetics Impinges on Religion". Robotics demonstrates that human beings can bring about, by the assembly of inanimate matter, entities that are capable of adaptive and intelligent action. By the same token, robotics is a manifestation of human hubris, insofar as it can be construed as a technology allowing human beings to usurp and arrogate to themselves a divine prerogative. Thus, robotics adds a new dimension to the mythical association of sin and burglary with technological progress, which is subversive of the natural order insofar as it attempts to bend the course of nature to human ends.

The punishment delivered by the gods to human beings for their Promethean burglary are the

diseases and the afflictions of the soul flowing out of Pandora's vase. In terms that are more acceptable to contemporary sensibilities, this punishment is more appropriately identified with the human inability to predict completely the behaviour of their technological inventions and to fully control their uses for the benefit of humanity. This epistemic and practical predicament is particularly evident in the case of robotic technologies, insofar as it concerns the problem of understanding, predicting, and controlling the behaviour of artifacts which have an enormous potential for autonomous action and goal oriented behaviours. These epistemic limitations - it was emphasized in the previous sections - may give rise to behaviours that are not congruent with the intentions of robot designers and even conflict with entrenched values and moral attitudes. The image that the Promethean myth about the origin of technologies renders is a an image in which elements of promise and danger are deeply intertwined. To some extent our scientific understanding of technological devices can be used to set these elements apart. Scientific models of technological devices are uncertain and incomplete, but one can hardly doubt that scientific rationality provides us with the best uncertain knowledge that one can rely on in order to promote the promise of technologies for improving the well-being of human beings and to protect them from its dangers. In particular, in this chapter I have argued that epistemic reflections on the science and technology of robotics provide powerful assistance for the purpose of identifying promises and dangers of robotic systems, and for the purpose of evaluating the ethical sustainability of specific applications of robotic technologies.

References

Christaller T....., (2001) *Robotik. Perspektiven für menschliches Handeln in der zukünftingen Gesellschaft*.Berlin: Springer ,

Simon, H. (1996), The Sciences of the Artificial, 3rd ed., Cambridge: MIT Press.

Veruggio G, Operto F. (2008),