Dante Marino and Guglielmo Tamburrini:

# Learning robots and human responsibility

**Abstract:**

Epistemic limitations concerning prediction and explanation of the behaviour of robots that learn from experience are selectively examined by reference to machine learning methods and computational theories of supervised inductive learning. Moral responsibility and liability ascription problems concerning damages caused by learning robot actions are discussed in the light of these epistemic limitations. In shaping responsibility ascription policies one has to take into account the fact that robots and softbots – by combining learning with autonomy, pro-activity, reasoning, and planning – can enter cognitive interactions that human beings have not experienced with any other non-human system.

**Agenda**



**Authors:**

Dr. Dante Marino:

- Administrative officer, ISIS “Francesco De Sanctis”, 80122 Napoli, Italy
- Telephone and email: ☎ + 39 - 081 - 7618942, ✉ dante.marino@tiscali.it
- Relevant publications:
  - D. Marino, G. Tamburrini, Interazioni uomo-macchina. Riflessioni tecnoetiche su robotica, bionica e intelligenza artificiale, L’arco di Giano 44 (2005), pp. 77-88.

Prof. Dr. Guglielmo Tamburrini

- Dipartimento di Scienze Fisiche, Università di Napoli Federico II, Via Cintia, 80146 Napoli, Italy:
- ☎ +39-081-676817, ✉ tamburrini@na.infn.it, 💻 http://ethicbots.na.infn.it/tamburrini/index.htm
- Relevant publications:
  - R. Cordeschi, G. Tamburrini, Cordeschi R. and Tamburrini G. (2005), “Intelligent machinery and warfare: historical debates and epistemologically motivated concerns” in Magnani L. and Dossena R. (eds.), *Computing, Philosophy, and Cognition*, London, King’s College Publications, 1-20.
  - Tamburrini, G., Datteri, E. (2005), "Machine Experiments and Theoretical Modelling: from Cybernetic Methodology to Neuro-Robotics", *Minds and Machines* , Vol. 15, No. 3-4, pp. 335-358.
  - Tamburrini, G. (2006), “AI and Popper’s solution to the problem of induction”, in I. Jarvie, K. Milford, D. Miller (eds.), *Karl Popper: A Centennial Assessment*, *vol. 2, Metaphysics and Epistemology*, London, Ashgate.
  - Tamburrini G., Datteri E. (eds.), *Ethics of Human Interaction with Robotic, Bionic, and AI Systems, Workshop Book of Abstracts*, Istituto Italiano per gli Studi Filosofici, Naples, Italy.
  - Datteri, E., Tamburrini, G., "Biorobotic Experiments for the Discovery of Biological Mechanisms", forthcoming in *Philosophy of Science*.


ISSN 1614-1687

Dante Marino and Guglielmo Tamburrini:

# Learning robots and human responsibility

## The responsibility ascription problem for learning robots

In the near future, robots are expected to cooperate extensively with humans in homes, offices, and other environments that are specifically designed for human activities. It is likely that robots have to be endowed with the capability to learn general rules of behaviour from experience in order to meet task assignments in those highly variable environments. One would like to find in user manuals of learning robots statements to the effect that the robot is guaranteed to behave so-and-so if normal operational conditions are fulfilled. But an epistemological reflection on computational learning theories and machine learning methods suggests that programmers and manufacturers of learning robots may not be in the position to predict exactly what these machines will actually do in their intended operation environments. Under these circumstances, who is responsible for damages caused by a learning robot? This is, in a nutshell, the responsibility ascription problem for learning robots.

The present interest for this responsibility ascription problem is grounded in recent developments of robotics and artificial intelligence (AI). Sustained research programmes for bringing robots to operate in environments that are specifically designed for humans suggest that moral and legal aspects of the responsibility ascription problem for learning robots may soon become practically significant issues. Moreover, an analysis of this problem bears on the responsibility ascription problem for learning software agents too, insofar as the learning methods that are applied in robotics are often used in AI to improve the performance of intelligent softbots. Finally, an examination of these responsibility ascription problems may contribute to shed light on related applied ethics problems concerning learning software agents and robots. Problems of delegacy and trust in multi-agent systems are significant cases in point, which become more acute when learning is combined with additional features of intelligent artificial agents: human subjects may not be in the position to oversee, predict or react properly to the behaviour of artificial agents that are endowed with forms of autonomy, pro-activity, reasoning, planning, and learning; robotic and software agents can perform complicated planning and inferencing operations before any human observer is in the position to understand what is going on; agent autonomy and pro-activity towards human users may extend as far as to make conjectures about what a user wants, even when the user herself does not know or is unable to state her desires and preferences.

In addition to suggesting the present interest of an inquiry into the responsibility ascription problem for learning robots, these observations point to epistemic limitations that fuel this particular problem. It is these limitations that we turn now to discuss.

## Machine learning meets the epistemological problem of induction

The study of machine learning from experience is a broad and complex enterprise, which is based on a wide variety of theoretical and experimental approaches. A major theoretical approach is PAC (Probably Approximately Correct) learning. Distinctive features of this approach are briefly discussed here, and compared with more experimentally oriented approaches to machine learning, – with the overall aim of isolating epistemic limitations which contribute to shape the responsibility ascription problem for learning robots.

PAC-learning is a theoretical framework for the computational analysis of learning problems which sets relatively demanding criteria for successful learning. PAC-learning inquiries aim at identifying classes of learning problems whose correct solutions can (or alternatively cannot) be approximated with arbitrarily small error and with arbitrarily high probability by some computational agent, when the agent is allowed to receive as inputs training examples of the target function that are drawn from some fixed probability distribution and is allowed to use “reasonable” amounts of computational resources only (that is, resources that are polynomially bounded in the parameters expressing characteristic features of the learning problem; for a precise definition of PAC-learnability, and examples of functions that are (not) PAC-learnable, see Mitchell 1997, pp. 203-214).

Does the PAC-learning paradigm put robot manufacturers and programmers in the position to certify that a robot will manifest with some high probability a behaviour which closely approximates a correct use of that concept or rule? One should be careful to note that such certifications may not be forth-


ISSN 1614-1687

coming in cases that are relevant to the responsibility ascription problem for learning robots, either in view of negative results (concerning problems that turn out to be not PAC-learnable) or in view of the difficulty of imposing the idealized PAC model of learning on concrete learning problems. Moreover, one should not fail to observe that these certifications do not put one in the position to understand or predict the practical consequences of the (unlikely) departures of PAC-learners from their target behaviour.

In connection with the limited applicability of PAC-learning methods, let us note that various classes of learning problems which admit a relatively simple logical formulation are provably not PAC-learnable. For example, the class of concepts that are expressible as the disjunction of two conjunctions of Boolean variables (Pitt and Valiant 1988) is not PAC-learnable. Moreover, the possibility of PAC-learning several other interesting classes of learning problems is still an open question. Finally, let us notice that one may not be in the position to verify background assumptions that are needed to apply the PAC model of learning to concrete learning problems. For example, the class of concepts or rules from which the computational learning system picks out its learning hypothesis is assumed to contain arbitrarily close approximations of the target concept or rule. But what is the target function and how can one identify its approximations, when the learning task is to recognize tigers on the basis of a training set formed by pictures of tigers and non-tigers? Another assumption of the PAC-learning model which is often unrealistic is that the training set always provides noise-free, correct information (so that misclassifications of, say, tigers and non-tigers do not occur in the training set).

In connection with the evaluation of the occasional departures from target behaviour that a PAC-learner is allowed to exhibit, one has to notice that the PAC-learning paradigm does not guarantee that these unlikely departures from target behaviour will not be particularly disastrous. Thus, the PAC-learnability of some concept or rule does not make available crucial information which is needed to understand and evaluate contextually the practical consequences of learning robot actions.

PAC-learning relieves instructors from the problem of selecting "suitable" training examples, insofar as a function can be PAC-learnt from randomly chosen examples. In contrast with this, the machine learning methods for supervised inductive learning that are applied in many cases of practical interest must rely on the background hypothesis that the selected training and test examples are "representative" examples of the target function. (The ID3 decision tree learning method is a pertinent case in point; see Mitchell 1997, ch. 3, for presentation and extensive analysis of this method.) The success of a supervised inductive learning process is usually assessed, when training is completed, by evaluating system performance on a test set, that is, on a set of examples that are not contained in the training set. If the observed performance on the test examples is at least as good as it is on the training set, this result is adduced as evidence that the machine will approximate well the target function over all unobserved examples (Mitchell 1997, p. 23). However, a poor approximation of the target function on unobserved data cannot be excluded on the basis of these positive test results, in view of the *overfitting* of both training and test data, which is a relatively common outcome of supervised inductive learning processes.[1] Overfitting gives rise to sceptical doubts about the soundness of inductive learning procedures, insofar as a good showing of an inductive learning algorithm at future outings depends on the fallible background hypothesis that the data used for training and test are sufficiently representative of the learning problem. This is the point where machine learning meets the epistemological problem of induction, insofar as the problem of justifying this background hypothesis about inductive learning procedures appears to be as difficult as the problem of justifying the conclusions of inductive inferences by human learners and scientists (for discussion, see Tamburrini 2006; for an analysis of early cybernetic reflections on the use of learning machines, see Cordeschi and Tamburrini 2005).

[1] Roughly speaking, a hypothesis *h* about some concept or rule from class *H* is said to overfit the training set if there is another hypothesis *h'* in *H* which does not fit the training set better than *h* but performs better than *h* on the whole set of concept or rule instances. "Overfitting is a significant practical difficulty for decision tree learning and other learning methods. For example, in one experimental study of ID3 involving five different tasks with noisy, non-deterministic data,... overfitting was found to decrease the accuracy of learned decision trees by 10-25% on most problems." (Mitchell 1997, p. 68).

## Is there a responsibility gap?

Epistemic limitations concerning knowledge of what a learning machine will do in normal operating situations have been appealed to in order to argue for a responsibility gap concerning the consequences of learning systems actions. Andreas Matthias put forward the following argument (Matthias 2004):

- Programmers, manufacturers, and users may not be in the position to predict what a learning robot will do in normal operating environments, and to select an appropriate course of action on the basis of this prediction;
- thus, none of them is able to exert full control on the causal chains which originate in the construction and deployment of a learning robot, and may eventually result into a damage for another party;
- but a person can be held responsible for something only if that person has control over it; therefore, one cannot attribute programmers, manufacturers or users responsibility for damages caused by learning machines;
- since no one else can be held responsible, one is facing "a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription".

A distinctive feature of traditional concepts of responsibility which, in Matthias's view, give rise to this responsibility gap is the following "control requirement" (CR) for correct responsibility ascription: a person is responsible for *x only if* the person has control over *x*. Thus, the lack of control by programmers, manufacturers or users entails that none of them is responsible for damages resulting from the actions of learning robots. (CR) is usually endorsed and used as a premise in arguments for *moral* responsibility ascription. (But one should be careful to note that different interpretations of the notion of control are possible and prove crucial to determine the scope of someone's moral duties.) Matthias claims that (CR) is to be more extensively applied – indeed, to all situations which call for a responsibility ascription *in accordance with our sense of justice*.

> *For a person to be rightly held responsible, that is, in accordance with our sense of justice, she must have control over her behaviour and the resulting consequences "in a suitable sense". That means that the agent can be considered responsible only if he knows the particular facts surrounding his action, and if he is able to freely form a decision to act, and to select one of a suitable set of available alternative actions based on these facts. (Matthias 2004, p. 175).*

Here the scope of (CR) is overstretched. In general, the possibility of ascribing responsibility according to familiar conceptions of justice and right is not jeopardized in situations in which no one can be held morally responsible in view of a lack of control. (CR) is not necessary for responsibility ascriptions, and the alleged responsibility gap depending on it concerns moral responsibility ascriptions only. Indeed, the epistemological reflections reported in the previous section suggest that the responsibility ascription problems concerning possible applications of machine learning investigations are a recent acquisition of a broad and extensively analyzed class of *liability* problems, where the causal chain leading to a damage is not clearly recognizable, and no one is clearly identifiable as blameworthy. Traditional concepts of responsibility ascription exist for these problems and have been routinely applied in the exercise of justice. Accordingly, a shift from moral responsibility to another - but nonetheless quite traditional - concept of responsibility, which has to be adapted and applied to a newly emerging casuistry, enables one to "bridge" the alleged responsibility gap concerning the actions of learning robots.

Responsibility problems falling under this broad category concern children's parents or tutors, pet owners, legal owners of factories for damages caused by workers, and more generally cases in which it is difficult to identify in a particular subject the origin of the causal chain leading to the damaging event. Parents and tutors who fail to provide adequate education, care and surveillance are, in certain circumstances, held responsible for damages caused by their young, even though there is no clear causal chain connecting them to the damaging events. Producers of goods are held responsible on the basis of even less direct causal connections, which are aptly summarized in a principle such as *ubi commoda ibi incommoda*. In these cases, expected producer profit is taken to provide an adequate basis for ascribing responsibility with regard to safety and health of workers or damages to consumers and society at large.

In addressing and solving these responsibility ascription problems, one does not start from such things as the existence of a clear causal chain or the

awareness of and control over the consequences of actions. The crucial decisions to be made concern the *identification of possible damages*, their *social sustainability*, and how *compensation* for these damages is to be distributed. Epistemological reflections on machine learning suggest that many learning robot responsibility ascription problems belong to this class. And epistemological reflections will also prove crucial to address the cost-benefit, risk assessment, damage identification, and compensation problems that are needed to license a sensible use of learning robots in homes, offices, and other specifically human habitats.

## Responsibility ascription policies: science, technology, and society

The responsibility ascription problems mentioned above are aptly classified as retrospective, that is, concerning past events or outcomes. In view of the above remarks, retrospective responsibility ascriptions for the actions of learning robots may flow from some conception of moral agency or from a legal system or from both of these. But what about prospective responsibilities concerning learning robots? In particular, who are the main actors of the process by which one introduces into a legal system suitable rules for ascribing responsibility for the actions of learning robots? These rules should enable one to identify possible damages that are deemed to be socially sustainable, and should specify criteria according to which compensation for these damages is to be distributed. Computer scientists, roboticists, and their professional organizations can play a crucial role in the identification of such rules and criteria. In addition to acting as whistleblowers, scientists, engineers, and their professional organizations can provide systematic evaluations of risks and benefits flowing from specific uses of learning robots, and may contribute to shape scientific research programmes towards the improvement of learning methods. However, wider groups of stakeholders must be involved too. An examination of issues which transcend purely scientific and technological discourses is needed to evaluate costs and benefits of learning robots in society, and to identify suitable liability and responsibility policies: For the benefit of whom learning robots are deployed? Is it possible to guarantee fair access to these technological resources? Do learning robots create opportunities for the promotion of human values and rights, such as the right to live a life of independence and participation in social and cultural activities? Are specific issues of potential violation of human rights connected to the use of learning robots? What kind of social conflicts, power relations, economic and military interests motivate or are triggered by the production and use of learning robots? (Capurro *et al.* 2006)

No responsibility gaps and no conceptual vacua are to be faced in ascribing responsibility for the action of learning robots. At the same time, however, one should not belittle the novelty of this problem and the difficulty of adapting known liability criteria and procedures to the newly emerging casuistry. The fact that this responsibility ascription problem concerns a very special kind of machines is aptly illustrated by its assimilation, in the above discussion, to responsibility and liability problems concerning parents and pet owners, that is, problems concerning the consequences of flexible and intelligent sensorimotor behaviours of biological systems. Moreover, when learning is combined in a robot with additional features of intelligent artificial agents - such as autonomy, pro-activity, reasoning, and planning - human beings are likely to enter cognitive interactions with robots that have not been experienced with any other non-human biological system. Sustained epistemological reflections will be needed to explore and address the novel applied ethics issues that take their origin in these cognitive interactions.

### References

*Capurro, R., Nagenborg M., Weber J., Pingel C. (2006), "Methodological issues in the ethics of human-robot interaction", in G. Tamburrini, E. Datteri (eds.), Ethics of Human Interaction with Robotic, Bionic, and AI Systems, Workshop Book of Abstracts, Napoli, Istituto Italiano per gli Studi Filosofici, p. 9.*

*Cordeschi R. and Tamburrini G. (2005), "Intelligent machinery and warfare: historical debates and epistemologically motivated concerns" in Magnani L. and Dossena R. (eds.), Computing, Philosophy, and Cognition, London, King's College Publications, pp. 1-20.*

*Matthias A. (2004), "The responsibility gap: Ascribing responsibility for the actions of learning automata", Ethics and Information Technology* ***6****, pp. 175-183.*

*Mitchell, T.M. (1997), Machine Learning, New York, McGraw Hill.*

*Pitt, L. and Valiant L. (1988). "Computational limitations on learning from examples", Journal of the ACM* ***35****, pp. 965-984.*

*Tamburrini, G. (2006), "AI and Popper's solution to the problem of induction", in I. Jarvie, K. Mil-*

*ford, D. Miller (eds.), Karl Popper: A Centennial Assessment, vol. 2, Metaphysics and Epistemology, London, Ashgate, pp. 265-282.*