
Intelligent Machines and Warfare

Historical Debates and Epistemologically Motivated Concerns

ROBERTO CORDESCHI AND GUGLIELMO TAMBURRINI

ABSTRACT. The early examples of self-directing robots attracted the interest of both scientific and military communities. Biologists regarded these devices as material models of animal tropisms. Engineers envisaged the possibility of turning self-directing robots into new “intelligent” torpedoes during World War I. Starting from World War II, more extensive interactions developed between theoretical inquiry and applied military research on the subject of adaptive and intelligent machinery. Pioneers of Cybernetics were involved in the development of goal-seeking warfare devices. But collaboration occasionally turned into open dissent. Founder of Cybernetics Norbert Wiener, in the aftermath of World War II, argued against military applications of learning machines, by drawing on epistemological appraisals of machine learning techniques. This connection between philosophy of science and techno-ethics is both strengthened and extended here. It is strengthened by an epistemological analysis of contemporary machine learning from examples; it is extended by a reflection on *ceteris paribus* conditions for models of adaptive behaviors.

1 Introduction

The so-called “electric dog”, the ancestor of phototropic self-directing robots, designed about 1912 by engineers John Hammond, Jr. and Benjamin Miessner, graphically illustrates the interest of both scientific and military communities for early self-directing robots. In 1918, biologist Jacques Loeb emphasized the significance of the electric dog as a material model of animal phototropism. He argued that the actual construction of this machine supported his own theoretical model of animal phototropism, insofar as the machine was internally organized as prescribed by the theoretical model and turned out to behave just like heliotropic organisms. Possible applications of this self-directing device as an “intelligent” weapon were enthusiastically proposed in 1915, during World War I. Section 2 reports on both motives of interests for the electric dog.

In 1943, psychologist Kenneth Craik named “synthetic method” the process of testing behavioral theories through machine models. The “synthetic method”, envisaged by Loeb in his reflections on heliotropic machines, has been enjoying increasing popularity in the modelling and explanation of animal and human behavior from Cybernetics up to the present time. And warfare applications flowing from particular implementations of the synthetic method have flourished too. Pioneers of Cybernetics were often involved in both synthetic modelling and military adaptations of their machine models during World War II. Kenneth Craik is a prominent case in point; Norbert Wiener went as far as claiming that World War II was “the deciding factor” for the development of Cybernetics. But Wiener argued against military applications of cybernetic machines in the aftermath of World War II, especially by drawing on epistemological reflections on machine learning techniques. In 1960, dissenting with AI pioneer Arthur Samuel, Wiener envisaged “disastrous consequences” from the action of automatic machines operating faster than human agents, or the action of learning machines abstracting their own behavioral rules from experience. Wiener tapped from his specialized knowledge to make public opinion aware of dangers connected to military applications of adaptive machines, and to undermine intelligent weaponry rhetoric. Section 3 highlights both collaborative and critical attitudes manifested by pioneers of Cybernetics towards warfare application of their system design principles.

Wiener’s arguments vividly illustrate how philosophy of science bears on the implementation of precautionary principles in applied research. This connection between philosophy of science and techno-ethics is strengthened in section 4, on the basis of an epistemological analysis of machine learning from examples. One would like to have a guarantee that a robot will learn to behave as expected most of the time, without bringing about the “disastrous consequences” that Wiener contemplated in awe; but theoretical guarantees of this sort, – it is pointed out by reference to so-called supervised inductive learning –, are very hard to come. Finally, Wiener’s reflections on the connections between philosophy of science and techno-ethics are extended in section 5 by considering *ceteris paribus* conditions for adaptive machine behaviors.

2 From the “electric dog” to the “dog of war”

It was not unusual to read in American newspapers and popular science magazines from about 1915 the description of a machine that looked like little more than a toy but attracted much attention for its unprecedented features as an “orientation mechanism.” This machine, designed in 1912 by two American experts in radio-controlled devices, John Hammond Jr.

and Benjamin Miessner, was actually built by the latter. Two years later, Miessner presented this machine in the *Purdue Engineering Review* under the name of “electric dog,” by which it became popularly known.

Miessner described in some detail the behavior of the electric dog in *Radiodynamics: The Wireless Control of Torpedoes and Other Mechanisms* [Miessner, 1916]. The electric dog orientation mechanism included two selenium cells. These cells, when influenced by light, effect the control of two sensitive relays. These relays, in their turn, control two electromagnetic switches: when one cell or both are illuminated, the current is switched onto the driving motor; when one cell alone is illuminated, an electromagnet is energized and effects the turning of the rear steering wheel. In this case, the turning of the machine brings the shaded cell into light. As soon and as long as both cells are equally illuminated with sufficient intensity, the machine moves in a straight line towards the light source. By turning a switch on, which reverses the driving motor’s connections, the machine can be made to back away from light. When the illumination intensity is so decreased by the increasing distance from the light source that the resistances of the cells approach their dark resistance, the sensitive relays break their respective circuits, and the machine stops.

The self-directing capacity of the electric dog attracted the attention of Jacques Loeb, described by Miessner as “the famed Rockefeller Institute biologist, who had proposed various theories explaining many kinds of tropism.” The explanation of the orientation mechanism, Miessner emphasized, was “very similar to that given by Jacques Loeb, the biologist, of reasons responsible for the flight of moths into a flame.” In particular, the electric dog’s lenses corresponded to “the two sensitive organs of the moth” (p. 196). Miessner carefully noted that “Hammond had been much taken with the writings of Jacques Loeb” (p. 36)¹

Loeb reprinted excerpts from Miessner’s machine description in *Forced Movements, Tropisms, and Animal Conduct* [Loeb, 1918, pp. 68-69], a book documenting his extensive work on lower organism tropisms. In particular, Loeb carefully documented the ways in which the orientation of bilaterally symmetrical lower animals, like the moth, depends on light. These are in fact “*natural* heliotropic machines”. Now, he claimed to have found an instance of “*artificial* heliotropic machine”, as he called it, in the orientation mechanism of the electric dog. His surprise was quite justified. Automata of the earlier mechanistic tradition could not simulate the heliotropic behavior

¹See [Cordeschi, 2001, ch. 1] for more details about the working of electric dog and Loeb’s theory of tropisms. Arguably, the electric dog is a forerunner of Walter Grey Walter light sensitive “tortoises” and Braitenberg’s “vehicles” (see [Walter, 1953; Braitenberg, 1984]).

of biological systems. These automata, based on the concept of clockwork mechanism, were incapable of exchanging information with the environment. In short, what was needed to achieve this kind of simulation was a machine endowed with sense organs, a negative-feedback control device, and motor organs. Hammond and Miessner's automaton was just such a machine, automatically adapting its behavior to the changing conditions of the external environment, and adjusting its movements by means of a negative-feedback control device (the rear steering wheel brought the machine back in the direction of light whenever it went too far off its course). Loeb's keen interest in this machine was motivated on epistemological grounds:

It seems to the writer that the actual construction of a heliotropic machine not only supports the mechanistic conceptions of the volitional and instinctive actions of animals but also the writer's theory of heliotropism, *since this theory served as the basis in the construction of the machine*. We may feel safe in stating that there is no more reason to ascribe the heliotropic reactions of lower animals to any form of sensation, e.g., of brightness or color or pleasure or curiosity, than it is to ascribe the heliotropic reactions of Mr. Hammond's machine to such sensations [Loeb, 1918, pp. 68-69].

This epistemological standpoint was to enjoy increasing popularity in the explanation of animal and human behavior up to our time. According to Loeb, a behavioral theory is supported by the theory-driven construction of a machine that behaves like the living organisms in the domain of the theory. The machine is a material model of biological systems insofar as it *embodies* the assumptions of the behavioral theory serving as a basis for its construction. Loeb regarded Hammond and Miessner's machine as a significant step in the process of eliminating idle hypotheses about purportedly fundamental differences between natural (that is, living or "chemical") machines and artificial (that is, inanimate or "inorganic") machines. In addition, the machine simulation of an organism's heliotropic behavior provided strong evidence that mentalistic language was not needed to predict and explain animal behavior. This simulation showed that the physical principles harnessing the simulating machine suffice to explain the behavior of lower animals in the domain of the biological theory. The elimination of introspective psychology ("speculative" or "metaphysical" psychology, as Loeb called it) from scientific inquiries into animal behavior is coherent with Loeb's purely automatic (mechanical) account of animal reaction to stimuli, and his concomitant refusal to ascribe sensations to lower animals. The moth does not fly towards the flame out of "curiosity," nor is it "attracted by" or "fond of" light, as earlier animal psychologists put it. It is simply "oriented" by the action of light—just like Hammond and Miessner's machine.

Loeb's interest for Hammond and Miessner's machine was epistemologically motivated, insofar as the electric dog enables one to test the empirical adequacy of some biological theory of behaviour. Different motives of interest for this machine soon emerged. At the time, Hammond was well-known for his dirigible torpedoes, - actually remote control radio-directed boats. Since 1910 he had been running a research laboratory in Gloucester, Massachusetts, where he was perfecting several radio-controlled torpedoes. Miessner was one of his main collaborators in the years 1911 and 1912. He wrote a long description of these devices for *Radiodynamics*, in which he mentioned earlier related work, in particular the so-called "teleautomata" or "self-acting automata" built in New York between 1892 and 1897 by Nikola Tesla, another pioneer of radio-controlled systems.

It was Miessner who explained a chief reason of interest for the orientation mechanism: Hammond's dirigible torpedo "is fitted with apparatus similar to that of the electric dog, so that if the enemy turns their search light on it, it will immediately be guided toward that enemy automatically" (p. 198). In the 1915 volume of the *Electrical Experimenter* one finds an enthusiastic description of both Hammond's torpedo and electric dog, jointly considered as a target-seeking automatic system, and prized for effective military applicability. This should not come as a surprise, as Europe was at the time engulfed in World War I.

[...] The performance of Mr. Hammond's truly marvellous radio-mechanical craft [...] seems to inherit superhuman intelligence [...] It bids fair to revolutionize modern warfare methods. The USA Government is seriously considering the purchase of the entire rights in this radio control scheme, as worked out by young Mr. Hammond and his associate scientist and engineers. It would be of inestimable value for the protection of harbors, [...] and it also could be directed from shore directly at or toward any hostile warship it is seen that a very powerful weapon is thus placed in the hands of our coast defense corps. It has been reported of the late that the Japanese Government has been negotiating for the exclusive rights to this invention, but undoubtedly the American naval authorities will be wide awake to the far-reaching merits and properties of such a system [...] Likewise, it has been proved in Mr. Miessner's experiment that the deadly naval torpedo or even an automatic bomb-dropping aeroplane can be manoeuvred in action from ship or shore by the [electric dog]. (*The Electrical Experimenter*, September and June 1915, pp. 211 and 43)

Miessner regarded the automatic orientation mechanism that he designed for the electric dog a significant advance over earlier devices, as it made Hammond's torpedo self-directing. He claimed that its self-directing capacity could be further refined on the basis of some experiments in submarine

detection and defence. And prophetically concluded:

The electric dog, which now is but an uncanny scientific curiosity, may within the very near future become in truth a real ‘dog of war,’ without fear, without heart, without the human element so often susceptible to trickery, with but one purpose: to overtake and slay whatever comes within range of its senses at the will of its master [Miessner, 1916, p. 199].

3 Cybernetics and applied military research: wartime connections

Miessner’s forecast was vindicated by the advent, less than thirty years later, of automatic control systems. It was another conflict, the Second World War, as Norbert Wiener pointed out, “the deciding factor” for cybernetic control systems, based on the mathematics of stochastic processes developed by Wiener himself, and the newborn technology of computing machinery [Wiener, 1961, p. 3]. This was particularly evident in anti-aircraft predictors: as Wiener pointed out, it was “the German prestige in aviation and the defensive position of England” (p. 5) which pushed many scientists towards applied research on these automatic devices. Wiener and Julian Bigelow investigated the theory of the curvilinear prediction of flight, and supervised the construction of self-controlling and computing apparatuses based on this theory (several papers Wiener and Bigelow produced on this topic were either secret or restricted). These apparatuses were designed “to usurp” (p. 6) the human functions of computing and forecasting, at least insofar as forecasting the future position of flying targets was concerned.

The epistemological implications of these wartime investigations were worked out later on, once Wiener became acquainted with Arturo Rosenblueth’s work on self-regulating mechanisms in biological systems, and were presented in papers outlining scope and heuristic principles of cybernetic research programmes (see [Rosenblueth *et al.*, 1943; Rosenblueth and Wiener, 1945]).

Loeb’s view of the epistemological and methodological relationship between his tropism theory and Hammond and Miessner’s phototropic machine is consistent with Rosenblueth and Wiener’s more general analysis of the relationship between theoretical (or formal) models and material models. The latter, in their view, may enable the carrying out of experiments under more favorable conditions than would be available in the original system. This translation presumes that there are reasonable grounds for supposing a similarity between the two situations; it thus presupposes the possession of an adequate formal model, with a structure similar to that of the two material systems. The formal model need not to be thoroughly comprehended; the material model then serves to supplement the formal

one [Rosenblueth and Wiener, 1945, p. 317].

A material model taking the form of a machine may enable the carrying out of suitable tests on the theoretical or formal model, because the latter “served as the basis in the construction of the machine”, as Loeb put it. This epistemological and methodological standpoint is at the core of the cybernetic programme and motivates much AI and robotics research up to present time².

The knowledge flow to and from machine-based investigations into adaptive biological behaviors and applied warfare research is evident in the work of British scientists from the early 1940’s. The work of Cambridge psychologist Kenneth Craik is a significant case in point. His investigations on scanning mechanisms and control systems were a major source of inspiration for epistemological claims made in his book *The nature of explanation* [Craik, 1943].

The scientific activity of Craik and other pioneers of automatic computing and control got a shaft in the arm from military research projects carried out during World War II. Grey Walter’s recollections graphically convey the interconnection of scientific and defence goals in the control mechanism research community in general, and in Craik’s work in particular:

The first notion of constructing a free goal-seeking mechanism goes back to a wartime talk with the psychologist Kenneth Craik, whose untimely death was one of the greatest losses Cambridge has suffered in years [Craik died in 1945 at 31]. When he was engaged on a war job for the Government, he came to get the help of our automatic analyser with some very complicated curves he had obtained, curves relating to the aiming errors of air gunners. Goal-seeking missiles were literally much in the air in those days; so, in our minds, were scanning mechanisms [Walter, 1953, p. 53].

In his 1943 book, Craik stated that thought’s function is “prediction” which, in its turn, involves three steps: “translating” processes of the external world, perceived by means of a sensory apparatus, into an internal, simplified or small-scale model; drawing from this model possible inferences about the world by appropriate machinery; “retranslating” this model into external processes, i.e. acting by means of a motor system (pp. 50-51). According to Craik, both organisms and newly conceived feedback machines are predictive systems, even though the latter are still quite rudimentary in

²For discussion, see [Cordeschi, 2001, ch. 4], and for conceptual connections between cybernetics and contemporary biorobotic modelling, see [Tamburrini and Datteri, forthcoming]. Note however the different conclusions that are drawn from the use of self-regulating machines as models of organisms: for Loeb, this use justifies the *elimination* of mental language in the study of living organisms, for Wiener and co-workers, this use justifies the *introduction* of mental language (under the form of a reinstated “teleological” language) in behavioral inquiries about living organisms [Rosenblueth *et al.*, 1943].

the way of prediction. As an example of such machines, Craik mentioned the anti-aircraft gun with a predictor, so familiar to Wiener and other pioneers of Cybernetics. And he described the human control system as a “chain” that includes a sensory device, a computing and amplifying system, and a response device. This is what Craik called “the engineering statement of man”, whose abstract functional organization was a source of inspiration for his military investigations as well. The concept of man as computing and control system (the engineering statement of man) was admittedly a radical simplification, neglecting many dimensions of human psychology that Craik mentioned in *The Nature of Explanation*. But this simplification served to unveil deep connections across academic subjects: psychology, in Craik’s words, was to bridge “the gaps between physiology, medicine and engineering”, by appeal to the shared functional architecture of computing and control systems.

The development of computer science paved the way to broader functional investigations into adaptive and intelligent behaviors. In particular, the modelling and development of learning systems, capable of improving their performance with experience, became a priority in problem solving, perception, and action planning by machines. Wiener appealed to the early developments of machine learning in order to emphasize the limited control one has, in general, on the outcome of automatic learning procedures, and the “disastrous consequences” that might be expected from this.³ This epistemological appraisal became a major premise in his arguments against military applications of learning machines. Notably, in a 1960 article he criticized the use of learning machines in decisions concerning “push-button wars”:

It is quite in the cards that learning machines will be used to program the pushing of the button in a new push-button war [...] The programming of such a learning machine would have to be based on some sort of war game [...] Here, however, if the rules for victory in a war game do not correspond to what we actually wish [...] such a machine may produce a policy which would win a nominal victory on points at the cost of every interest we have at heart (Wiener 1960: 1357).

³The association “Computer Professionals for Social Responsibility” established a Norbert Wiener Award in 1987. The motivation for naming this award after Wiener mentions the fact that “Wiener was among the first to examine the social and political consequences of computing technology. He devoted much of his energy to writing articles and books that would make the technology understandable to a wide audience.” It is worth recalling, in connection with the techno-ethical issues discussed here, that the Norbert Wiener award was assigned in 2001 to Nira Schwartz and Theodore Postol “For their courageous efforts to disclose misinformation and falsified test results of the proposed National Missile Defense system”. See <http://www.cpsr.org/cpsr/wiener-award.html>

Arthur Samuel, a pioneer of AI investigations into problem solving and machine learning, dismissed Wiener's concern on the ground that machine actions fulfil the intentions of its human programmer or intentions directly derived from these. In Samuel's words, "the 'intentions' which the machine seems to manifest are the intentions of the human programmer, as specified in advance, or they are subsidiary intentions derived from these, following rules specified by the programmer" [Samuel, 1960, p. 741].

Samuel's sweeping "optimism" is not really supported by theoretical knowledge of machines. For one thing, the undecidability results, obtained in the framework of computability theory about 25 years before Samuel's article was written, suffice to show that machines are, in general, unpredictable. For example, the undecidability of the halting problem shows that there is no algorithmic procedure enabling one to decide of every given program and input whether that program will eventually halt with a definite output.⁴ Notice that this epistemological limitation concerns the whole class of algorithmic procedures, independently of whether these are specified by human programmers or not. Even more significantly bearing on the Wiener-Samuel controversy is more recent work on machine learning from examples. This work shows that one has limited control on what a machine actually learns, at least insofar as major supervised learning techniques are concerned. These epistemological reflections, we submit, strengthen Wiener's appraisal of limited human understanding and control of automatic learning procedures, and therefore support the major premise in his arguments for the implementation of precautionary principles in warfare applications of learning machines. Let's see.

4 Learning machines and warfare: epistemologically motivated concerns

A central issue in machine learning is whether a machine which learns from experience and approximates the target function well over a fairly large set of training examples will also approximate the target function well over unobserved examples. The connection between this issue and the classical epistemic problem of induction in both scientific method and practical reasoning was explored by Donald Gillies, who claimed that scepticism towards induction is no longer tenable in the light of recent advances of machine learning in the way of both concept and rule learning [?]. The epistemic problem of induction is the problem whether and what sorts of constraints can be imposed on inductive patterns of inference, so that their conclusions *be reasonable to believe*. In particular, Gillies appealed to ID3-style learning algorithms to support this claim. If Gillies were right, that is, if the

⁴See [Davis *et al.*, 1994, p. 68].

epistemic problem of induction were solved in particular machine learning domains, one would have a guarantee that such learning machines would behave as expected most of the time, thereby defusing Wiener's concerns about the consequences of warfare applications of learning machines. Indeed, Wiener's concerns were motivated just on the ground that one has only limited understanding and control of how learning machines will behave after training.

In contrast with this, we argue that Wiener's concerns are not defused by recent developments of machine learning. More specifically, we argue that a sweeping problem affecting supervised inductive learning in general, and ID3-style learning in particular, jeopardizes the idea that a genuine solution to the epistemic problem of induction is afforded by these learning systems. This is the overfitting of training data, which reminds one that a good approximation to the target concept or rule on training data is not, in itself, diagnostic of a good approximation over the whole instance space of that concept or rule. And the successful performances of machine learning systems are of no avail either in the present context: a familiar regress in epistemological discussions of induction arises as soon as one appeals to past performances of these systems in order to conclude that good showings are to be expected in their future outings as well. Thus, epistemic guarantees about the future behaviors of learning machines are very hard to come. These various problems have to be effectively addressed before one can conclude that Wiener's techno-ethical concerns are put to rest by more recent developments of machine learning.

Let us begin by emphasizing the connection between learning from examples and the epistemic problem of induction. A distinctively inductive assumption is often made about computational systems that learn concepts or rules from examples. Schematically,

(IC) Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over unobserved examples.⁵

Clearly, a critical examination of this broad assumption requires an extensive survey of learning systems that goes well beyond the scope of this paper. Here, we focus on versions of (IC) concerning the inductive decision tree algorithm ID3, for Gillies appealed just to ID3-style learning algorithms to claim that scepticism towards this inductivist claim is no longer tenable [Gillies, 1996].

Let us then consider the following inductive claim:

(IC-ID3) Any hypothesis constructed by ID3 which fits the target function over a sufficiently large set of training examples will approximate the

⁵[Mitchell, 1997, p. 23].

target function well over unobserved examples.

To begin with, let us recall some distinctive features of (the ID3) decision tree learning. Decision trees provide classifications of concept instances in a training set, formed by conjunctions of attribute/value pairs. Each path in the tree represents a classified instance. The terminal node of each path in the tree is labelled with the yes/no classification. The learnt concept description can be read off from the paths which terminate into a “yes” leaf. Such description can be expressed as a disjunction of conjunctions of attribute/value pairs. Concept descriptions that make essential use of relational predicates (such as “ancestor”) cannot be learnt within this framework.

ID3 uses a top-down strategy for constructing decision trees. Each non-terminal node in the tree stands for a test on some attribute, and each branch descending from that node stands for one of the possible values assumed by that attribute. An instance in the training set is classified by starting at the top-most, root node of the tree, testing the attribute associated to this node, selecting the descending branch associated to the value assumed by this attribute in the instance under examination, repeating the test on the successor node along this branch, and so on until one reaches a leaf. Each concept instance in the training set is associated to a path in a tree, which is labelled “yes” or “no” at the terminal node. ID3 places closer to the tree root attributes which better classify positive and negative examples in the training set. This is done by associating to each attribute P mentioned in the training set a measure of how well P alone separates the training examples according to their being positive or negative instances of the target concept. Let us call this preference in tree construction the ID3 “*informational bias*”.

There is another bias characterizing the ID3 construction strategy. ID3 stops expanding a decision tree as soon as a hypothesis accounting for training data is found. In other words, simpler hypotheses (shorter decision trees) are singled out from the set of hypotheses that are consistent with training data, and more complicated ones (longer decision trees) are discarded. On account of this *simplicity bias*,⁶ longer decision trees that are compatible

⁶Simplicity is identified here with the length of decision trees, and the latter is contingent on the choice of primitive attributes. A simplicity bias is introduced in many machine learning algorithms for hypothesis selection [Michalski, 1984, p. 98]: “For any given set of facts, a potentially infinite number of hypotheses can be generated that imply these facts. Background knowledge is therefore necessary to provide the constraints and a preference criterion for reducing the infinite choice to one hypothesis or a few preferable ones. A typical way of defining such a criterion is to specify the preferable properties of the hypothesis, for example, to require that the hypothesis is the shortest or the most economical description consistent with all the facts.”

with the training set are not even generated, and thus no conflict resolution strategy is needed to choose between competing hypotheses.

We are now in the position to state more precisely inductive claim (IC-ID3), by reference to the main background hypotheses used by ID-3 to reduce its hypothesis space:

(IC-ID3: second version): Any hypothesis constructed by ID3 on the basis of its informational and simplicity biases which fits the target function over a sufficiently large set of training examples will also approximate the target function well over unobserved examples.

Scepticism about this claim is fostered by the overfitting problem. A hypothesis $h \in H$ is said to overfit the training set if another hypothesis $h' \in H$ performs better than h on X , even though h' does not fit the training set better than h . Overfitting in ID3 trees commonly occurs when the training set contains an attribute P unrelated to the target concept, which happens to separate well the training instances. In view of this “informational gain” P is placed close to the tree root.

Overfitting is a significant practical difficulty for decision tree learning and many other learning methods. For example, in one experimental study of ID3 involving five different learning tasks with noisy, nondeterministic data,... overfitting was found to decrease the accuracy of learned decision trees by 10-25% on most problems [Mitchell, 1997, p. 68].

Unprincipled expansions of the original training set may not prevent the generation of overfitting trees, for a larger training set may bring about additional noise and coincidental regularities. Accordingly, claim (IC-ID3) is to be further qualified: the “sufficiently large set of training examples” mentioned there must be “sufficiently representative of the target concept” as well. This means that (implicit) assumptions about the representativeness of concept instance collections play a central role in successful ID3 learning. Consider, in this connection, the post-pruning of overfitting decision trees (Mitchell 1997: 67-72). In post-pruning, one constructs a “validation set”, which differs from both training and test sets. The validation set can be used to remove a subtree of the learnt decision tree: this is actually done if the pruned tree performs at least as well as the original tree on the validation set. Expectations of a good performance of the pruned tree on as yet unobserved instances rely on the assumption that the validation set is more representative of the target concept than the training set. Thus, the sceptical challenge directed at (IC-ID3) can be iterated after post-pruning, just by noting the conjectural character of this assumption.

In order to counter this sceptical challenge to (IC-ID3), one should look more closely at the criteria used for judging how representative of the target concept are training and validation examples. But additional problems arise

here. These criteria may vary over concepts, and are not easily stated in explicit form. In expert systems, for example, the introspective limitations of human experts is a major bottleneck in system development. The process of extracting rules from human experts turns out to be an extremely time consuming and often unrewarding task. These subjects can usually pick out significant examples of rules or concepts, but are often unable to state precisely the criteria underlying these judgments.⁷ Accordingly, automatic learning from examples is more likely to be adopted when criteria for selecting significant concept or rule instances are not easily supplied by human experts; and yet an examination of these criteria is just what is needed to support inductive claim (IC-ID3) by appeal to the representativeness of training examples.

Confronted with these various difficulties, which the sceptic consistently interprets as symptoms that inductive claim (IC-ID3) cannot be convincingly argued for, let us try and assume a different perspective on ID3. We have already formed a vague picture of ID3 as a component of a trial and error-elimination cycle: ID3 makes predictions about the classification of concept instances that are not included in the training set, on the basis of assumptions guiding both training set construction and the selection of some concept *c*. If predictions about unseen instances are satisfactory, then one is provisionally entitled to retain concept *c*. Otherwise *c* is discarded, and correction methods (such as post-pruning) come into play, which implicitly modify the original set of assumptions.

To sharpen this description of ID3 processing as a two-layered prediction-test cycle (leading from a falsification of instance classification predictions to a refutation of the conjunction of the various assumptions used to select the falsified hypothesis), one can draw on the above distinction between the *preferences* or biases embedded in ID3 proper (which determine both the language for expressing concepts and the construction of decision trees) on the one hand, and the presuppositions that are used to select training sets on the other hand. In the end, ID3 learning projections will work as long as both kinds of assumptions will turn out to be adequate in the learning environment. But one has no *a priori* guarantee that this adequacy condition is actually satisfied. In other words, there is no guarantee that such machine, which learns from experience and happens to approximate the target function well over a sufficiently large set of training examples, will also approximate the target function well over unobserved examples.⁸

⁷See, for example, the survey of knowledge acquisition methods used in expert system research in [Puppe, 1993].

⁸For more extensive discussion of the relationship between AI and the philosophical problem of induction, see [Tamburrini, forthcoming].

In our opinion, the above epistemological analysis sharpens, in the case of ID3-style learning from examples, the broad motives for Wiener's reservations about warfare applications of learning machines, insofar as the hypotheses underlying successful learning from examples are more precisely identified, and their conjectural character is more clearly brought out. But how significant is this reflection about ID3-style learning for the more general problem Wiener raised about military applications of learning machines? One may reasonably suspect that some of Wiener's concerns can be defused by appeal to some other learning algorithm from examples, for some learning procedures may turn out to be immune from the above sceptical conclusions. In order to effectively address Wiener's concerns, however, one would have to show that the learning procedure in question enables one to accrue reliable information on the approximation or convergence to target functions.

5 Learning machines and normal task environments

The *distribution-free* or *probably approximately correct* learning (pac-learning) [Valiant, 1984] is an approach to machine learning which goes a long way towards meeting the epistemic requirement of reliable control on approximation or convergence to target functions. Pac-learning constraints are meant to ensure that the hypotheses advanced by means of a learning procedure using a reasonable amount of computational resources is most likely correct.

The broad motivations for this approach are informally presented by Valiant in connection with the guarantees one would like to read in the user manual of a newly bought home robot:

... whatever home you take this robot to, after sufficient training on some tasks it will behave as expected most of the time, as long as the general conditions expected there are stable enough. To make this informal statement into a usable criterion, some quantitative constraints are needed in addition. First, the number of training sessions required should be reasonable, as should the amount of computation required of the robot to process each input at each such session, Second, the probability that the robot fails to learn because the training instances were atypical should be small. Lastly, the probability that, even when the training instances were typical, an error is made on a new input should be small. Furthermore, in the last two cases the probability of error should be controllable in the sense that any level of confidence and reliability should be achievable by increasing the number of training instances appropriately [Valiant, 1994, p. 102].

In the domain of concept acquisition, for example, pac-learning addresses the problem of characterizing classes of concepts that one can learn with arbitrarily high probability from randomly drawn training examples using

bounded computational resources.⁹ Conceptually, this is a fairly satisfactory machine learning explication of the intuitive idea of an epistemically justified inductive procedure, as long as the “arbitrarily high probability” of a hypothesis is regarded as a meaningful indication that the learning system will behave as expected most of the time. Moreover, as Valiant emphasizes, hypotheses about the representativeness of training examples are not needed here, for the instances can be randomly drawn.

It turns out that the classes of concepts and rules that are known to be pac-learnable are fairly limited.¹⁰ For example, one of the major open problems in pac-learning is the efficient learning of DNF expressions, that is, the kind of learning problems discussed above in connection with ID3 learning. Moreover, the pac-learning approach is not considered as a definitive framework for practical learnability, but rather as a promising starting point [Turàn, 2000]. Accordingly, the relevance of pac-learnable concepts and rules in the military applications that Wiener was concerned with is not immediately obvious. More generally, in order to provide a satisfactory answer to the problem whether any machine learning approach provides a viable strategy to meet Wiener’s techno-ethical concerns, one has to address subtle epistemological questions concerning our capability to control and reliably estimate convergence to target functions in practically interesting machine learning applications.

An important proviso in Valiant’s vivid illustration of the guarantees one would like to have before buying some home robot has gone unnoticed in our discussion so far: this robot should mostly behave as expected in our homes *as long as the general conditions expected there are stable enough*. This proviso can be reformulated as the requirement that one can expect the robot to manifest a certain behaviour *if the functioning environment is normal*, that is, if no perturbing factors are present in that environment.

The problem of specifying normal functioning conditions for machines is another pervasive epistemological problem, bearing on various techno-ethical issues that arise in AI and robotics, in both learning and non-learning environments. Even assuming that some learning machine has been successfully trained at some task, the machine may still fail to behave as expected because of abnormal usage context. Specifying these normalcy conditions is akin to the inexhaustible problem of specifying the intended range of validity of any scientific law, given that even so-called universal physical laws hold *ceteris paribus*, that is, when perturbation factors are not present. A complete list of boundary conditions characterizing the range of validity of

⁹Computational resources must be polynomially bounded in the parameters expressing the relevant measures of the learning problem.

¹⁰For discussion, see [Mitchell, 1997, pp. 213–214], and references therein.

some scientific law or the environments in which a machine works properly is at best a regulative idea of scientific inquiry: in order to identify *every* causal factor which may disturb the regular behavior of some machine, one should take into account evident constraints (such as, say, “Temperature should not exceed 600 ° C”), examine conditions that are less readily classified as relevant or irrelevant (“No changes in gravitational force”), and pay some attention even to *prima facie* irrelevant conditions (“No Persian cats under the table”). Thorough examination of potentially relevant boundary conditions is nothing but thorough paralysis of scientific inquiry.

Since one cannot circumscribe precisely the class of normal task environments, for an unlimited number of boundary conditions should be taken into account, a more pragmatic attitude is usually adopted. In user manuals, one mentions what are deemed the more consequential or more easily overlooked boundary conditions - concerning, say, temperature, voltage, humidity, and so on - relying on a global commonsense judgment by machine users concerning the absence of any other abnormal usage condition. Similarly, for the purpose of testing in a selective manner whether some candidate boundary condition is actually needed to ensure normalcy, one builds up experimental settings E in which that boundary condition is lifted, and makes the default empirical hypothesis that no other abnormal task condition arises in E . Clearly, when erratic warfare scenarios are substituted for controlled experimental environments E , it is more difficult to support in a responsible way (that is, by severe testing) similar default hypotheses about the absence of disturbing factors, and thus the prediction that the machine will behave as expected in such warfare scenarios, without bringing about the “disastrous consequences” that Wiener contemplated in awe.

6 Concluding remarks

Hammond and Miessner’s self-regulating machine was hailed as a significant innovation in apparently distant, but ever since tightly interacting domains of inquiry. According to Loeb, this kind of machines supported his own behavioral theories in biology. And this very machine, insofar as it was endowed with “superhuman intelligence”, was seen as revolutionizing modern warfare technologies. Arguably, this is the first time that the potential impact of the newly conceived self-regulating machines on both scientific method and military technology is clearly identified. This potential impact became more evident during the cybernetic age. And the dangers arising from unconstrained military applications of cybernetic machines became more evident too. The connection between philosophy of science and techno-ethics suggested by Wiener’s reflections on warfare applications of learning machines has been strengthened here by a reflection

on more recent approaches to supervised inductive learning. And possible extensions of Wiener's reflections have been suggested by reference to the *ceteris paribus* problem for scientific laws and machine proper functioning.

Philosophy of science bears on the implementation of precautionary principles about military applications of AI and robotics in ways that have not been discussed in this paper. Notably, current military research on autonomous robotic agents addresses AI problems whose solution paves the way to the solution of any other problem that AI will ever be confronted with. These problems, by analogy to well-known classifications of computational complexity theory, might be appropriately called "AI-complete problems". As an example, consider the problem of recognizing surrender gestures by the enemy, or the capability of telling bystanders apart from hostile forces. Solving these recognition problems involves context-dependent disambiguation of gestures, understanding of emotional expressions, real-time reasoning about deceptive intentions and actions. However, human-level performances in these tasks, especially in uncontrolled warfare scenarios, are a far cry from current AI and robotics research efforts.

Techno-ethical issues arising from warfare applications of robotic and intelligent information systems are prominent items included into a much broader and rapidly growing list of techno-ethical issues emerging in these scientific and technological domains of inquiry. In the near future, robotic and intelligent information systems are expected to interact ever more closely with human beings, and to enhance human mental, physical, and social capabilities in significant ways. Crucial ethical issues in these areas, over and above responsibilities for (possibly unintended) warfare applications, include the preservation of human identity and integrity, applications of precautionary principles with respect to system autonomy, economic and social discrimination deriving from restricted access to robotic and intelligent information resources, system accountability, nature and impact of human-machine cognitive and affective bonds on individuals and society. Epistemological reflections on the scope and limits of our knowledge about AI and robotic systems are likely to improve our understanding, triaging, monitoring, and overall capability to cope with many of these techno-ethical issues.

Acknowledgements

An earlier version of this paper was presented at the First International Symposium on Roboethics, held at Villa Nobel, Sanremo, Italy, on January 30-31, 2004. We are grateful to the Symposium organizers and to the participants for stimulating comments. Financial support by MIUR (Italian Ministry for Education, University and Research), grant COFIN

#2002112548, is gratefully acknowledged.

Roberto Cordeschi

Dipartimento di Scienze della Comunicazione, Università di Salerno, Italy.

Email: cordeschi@caspur.it

Guglielmo Tamburrini

Dipartimento di Scienze Fisiche, Università di Napoli “Federico II”, Italy.

Email: gugt@inwind.it

BIBLIOGRAPHY

- [Braitenberg, 1984] V. Braitenberg. *Vehicles. Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA, 1984.
- [Cordeschi, 2001] R. Cordeschi. *The Discovery of the Artificial*. Kluwer, Dordrecht, 2001.
- [Craik, 1943] K. J. W. Craik. *The Nature of Experimentation*. Cambridge University Press, Cambridge, 1943.
- [Davis et al., 1994] M. Davis, R. Sigal, and E. Weyuker. *Computability, Complexity, and Languages*. Academic Press, Boston, MA, 1994.
- [Gillies, 1996] D. Gillies. *Artificial Intelligence and Scientific Method*. Oxford University Press, Oxford, 1996.
- [Loeb, 1918] J. Loeb. *Forced Movements, Tropisms, and Animal Conduct*. Lippincott, Philadelphia and London, 1918.
- [Michalski, 1984] R. S. Michalski. A theory of methodology of inductive learning. In R. S. Michalski, J. Carbonell, and T. M. Mitchell, editors, *Machine Learning, An Artificial Intelligence Approach*, pages 83–134, Berlin, 1984. Springer.
- [Miessner, 1916] B. F. Miessner. *Radiodynamics: The Wireless Control of Torpedoes and Other Mechanisms*. Van Nostrand, New York, 1916.
- [Mitchell, 1997] T. M. Mitchell. *Machine Learning*. McGraw Hill, New York, 1997.
- [Puppe, 1993] F. Puppe. *A Systematic Introduction to Expert Systems*. Springer, Berlin, 1993.
- [Rosenblueth and Wiener, 1945] A. Rosenblueth and N. Wiener. The role of models in science. *Philosophy of Science*, 12:316–21, 1945.
- [Rosenblueth et al., 1943] A. Rosenblueth, N. Wiener, and J. Bigelow. Behavior, purpose, and teleology. *Philosophy of Science*, 10:18–24, 1943.
- [Samuel, 1960] A. L. Samuel. Some moral and technical consequences of automation – a refutation. *Science*, 132, September 11:741–42, 1960.
- [Tamburrini and Datteri, forthcoming] G. Tamburrini and E. Datteri. Machine experiments and theoretical modelling: from cybernetics to biorobotics. *Minds and Machines*, forthcoming.
- [Tamburrini, forthcoming] G. Tamburrini. Ai and popper’s solution to the problem of induction. In I. Jarvie, K. Milford, and D. Miller, editors, *Karl Popper: A Centennial Appraisal*, London, forthcoming. Ashgate.
- [Turàn, 2000] G. Turàn. Remarks on computational learning theory. *Annals of Mathematics and Artificial Intelligence*, 28:43–45, 2000.
- [Valiant, 1984] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–42, 1984.
- [Valiant, 1994] L. G. Valiant. *Circuits of the Mind*. Oxford University Press, Oxford, 1994.
- [Walter, 1953] W. G. Walter. *The Living Brain*. Duckworth, London, 1953.

[Wiener, 1961] N. Wiener. *Cybernetics, or Control and Communication in the Animal and the Machine [1948]*. MIT Press, Cambridge, MA, 1961.