PETER ASARO*

ROBERTO CORDESCHI ON CYBERNETICS AND AUTONOMOUS WEAPONS: REFLECTIONS AND RESPONSES

1. Introduction[1]

Roberto Cordeschi was a dear friend as well as an intellectual guide and companion. My aim in this paper is to present the arguments that I wish I could have made to Roberto in person, while sipping wine on the citrus-lined veranda of his Raito home, though it was not to be. I feel compelled to do so because of my great respect for Roberto and his work, as well as his friendship. So it is for myself that I must ask, «How can someone who shares such a deep and nuanced view of the nature of machine autonomy, its historical origins, and its internal tensions, arrive at the conclusion that a ban on autonomous weapons is neither necessary nor desirable?».

Indeed, I have encountered others who have followed logical steps similar to those Cordeschi presents in these two pieces, and who also arrived at similar conclusions as he did. It is, therefore, my hope that other friends, colleagues and the unconvinced might still learn from this disagreement between us. And so I shall try to point out what I see as the missteps and misunderstandings in his otherwise careful analysis and powerful insights into the fundamental problem of autonomy and machine reliability, and its implications for autonomous weapons.

I further believe that there is value in revisiting his analysis of the cybernetic origins of machine autonomy and its associated network of concepts. This is because these concepts still underlie the metaphysics and epistemology of our current cybernetic era, and continue to inform our thinking about autonomous systems. In order to better understand the nature of agency and autonomy in complex socio-technical systems, there is great benefit in returning to these fundamental concepts, and revisiting the arguments that shaped their introduction and articulation. Two especially significant papers, Cordeschi and Tamburrini (2005) and Cordeschi (2013), lay out a view of autonomous weapons that draws attention to their relationship with the fundamental concepts developed in the early years of the

* The New School, School of Media Studies, Princeton University, Center for Information Technology Policy; asaro@newschool.edu.

[1] Research for this paper was supported by a grant from the Future of Life Institute Fund, and a fellowship at Princeton University's Center for Information Technology Policy.

cybernetics movement. By historicizing autonomous weapons in this way, these papers reveal some of the deeper concepts underlying what autonomous weapons are, why we may find them objectionable or undesirable, and how we might, or might not, "tame" them with engineering.

In this paper I will begin by reviewing Cordeschi's (2013) analysis and development of the arguments made by Norbert Wiener on the reliability of autonomous systems, and the potential dangers stemming from their unreliability. This will include a history of the problem leading up to Wiener's (1960) paper. It will also consider the crises within the conceptual framework of cybernetics itself, which led ultimately to the development of second-order cybernetics. I will draw upon those insights to help resolve the conceptual and semantic problem at the heart of purposive machines– namely that they have two purposes: purposes from their design and environmental interaction, and purposes stemming from those who operate and control them. Of course, this challenges Wiener's earlier ideas about purposive machines in Rosenblueth, Wiener and Bigelow (1943) and Rosenblueth and Wiener (1950), and how we understand the "purpose" of a machine.

In the third section, I will focus on Cordeschi's response to calls for a ban on autonomous weapons. In particular, Cordeschi explored how the precautionary principle might be applied to argue for such a ban. While Cordeschi found the precautionary principle wanting, and a ban on autonomous weapons unworkable, his analysis of these questions reveals why many people might share these conclusions.

I conclude in the fourth section by drawing out some lessons from the previous sections. In particular, I argue that a clearer understanding of the different types of purpose, design purpose and operator purpose, is an important distinction to keep in mind when discussing autonomous weapons. Moreover, it provides some conceptual and substantive guidance for thinking about and articulating "meaningful human control" in the context of international law regulating autonomous weapons (UNIDIR, 2014).

I must admit from the start that Cordeschi's analysis of the reliability problem facing autonomous weapons, and autonomous agents more generally, is a major contribution to the discussion of the issue. I largely agree with this analysis, and chose to draw out a technical disagreement for primarily pedagogical purposes. I maintain a friendly disagreement, however, both with Cordeschi's assessment of the goals and benefits of a ban on autonomous weapons, and with his analysis of the logic motivating such a ban. I believe that his assessment of the ban proposal is overly broad in construing it as a prohibition of all scientific and technological research into autonomous systems and autonomous weapons – this is simply not what

is being called for. As a result, his critique of a ban is focused on its presumed impact on scientific research, rather than its impacts on military use by states. In particular, he interpreted the primary justification for such a ban as being based in the precautionary principle. However, the precautionary principle is not the only, or necessarily the primary, motive for banning autonomous weapons. Regardless, he saw the precautionary principle as stemming from the intrinsic unreliability of autonomous weapons, and so it is to an analysis of that unreliability that I now turn.

2. The tension between machine autonomy and reliability

Wiener (1960) lays out a set of problems facing the reliability and predictability of autonomous machines that largely remain unsettled and unchanged today. In particular, as the behavior of automated systems becomes more complex, and more dependent on inputs from environmental sensors and external data sources, the less predictable they become (Marino and Tamburrini, 2006). While this is true of any artifact or system, it presents a particular challenge for those who devise systems which are design to predict and adapt to environmental variables – *i.e.* cybernetic machines.

Much of the value of revisiting Wiener's arguments is to sort out the ideas and language which continue to shape today's debates on topics such as the definition and nature of autonomous weapons (UNIDIR, 2014). Much of the power of Wiener's, and cybernetics', ideas and language came from analogizing machines to persons (Hayles, 1999; Asaro, 2008; Asaro, 2011). It is thus important to understand what is at stake in the application of terms such as "purpose" to a machine. The various epistemic and metaphysical tensions that resulted from this move eventually led to what is now known as second-order cybernetics (Hayles, 1999; Asaro, 2007; Asaro, 2008; Pickering, 2010). Consequently, I will employ some insights from second-order cybernetics to hopefully unravel some of the confusion that persists over the nature of purposive machines and their responsible development.

Wiener was very much aware of the application of purposive systems in warfare. Indeed, his work in this area was inspired by his efforts to build a predictive anti-aircraft gun targeting system (based in time-series analysis) during World War II, and resulted in his 1943 paper with Rosenblueth and Bigelow (Galison, 1994). In his 1960 paper, he recognized that learning systems would also be employed, which raised the question of both how to find the appropriate means and methods of training them, and what the implications of their errors and unreliability might be for humanity:

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. Here we are considering a field in which automata of a non-learning character are probably already in use. It is quite out of the question to program these machines on the basis of actual experience in real war. For one thing, a sufficient experience to give an adequate programming would probably see humanity already wiped out (Wiener, 1960, p. 1357).

This leads Wiener to speculate on the need for simulations and war games to train the systems, which leads to yet greater degrees of uncertainty and unreliability:

Moreover, the techniques of push-button war are bound to change so much by the time an adequate experience could have been accumulated, the basis of the beginning would have radically changed. Therefore, the programming of such a learning machine would have to be based on some sort of war game, just as commanders and staff officials now learn an important part of the art of strategy in a similar manner. Here, however, if the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that 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, even that of national survival (Wiener, 1960, p. 1357).

That is, Wiener recognized that even as autonomous learning systems might become increasingly more reliable in terms of their built-in metrics, the resulting actions and strategies might actually be deeply at odds with the goals and desires of society.

This disconnect between the goals of the human, whether an individual operator or a whole society, and the results of a purposefully-designed and trained machine, is what I am calling the fundamental ambiguity introduced by the conflation of human "intent" and machine "purpose".

It takes on various instantiations in classic narratives and conundrums, some of which were discussed by Wiener. I will now turn to the history leading up to this problem, and finally to a second-order cybernetic analysis of how to resolve the ambiguity at its heart.

2.1. The ambiguity of purpose without intent

Cordeschi points to the ambiguous status of "purpose" in the new cybernetic metaphysics, as being a source of confusion and even paradox:

To speak of rules that do not correspond with "what we actually wish" or the "interest we have at heart" means bringing to the fore the question of the designer's

or operator's real purpose. Wiener would arguably have recognised here the insufficiency of the "behaviouristic" analysis of human-machine interaction which he and Rosenblueth proposed in their 1950 reply to Taylor (1950) (Cordeschi, 2013, p. 434).

It is also worth noting that Cordeschi distinguishes operators and designers as each having their own purposes, neither of which necessarily align with the "purpose" of the machine. Wiener largely conflates all these purposes together.

A certain form of black-box functionalism[2] is implied by Wiener's definition of purposive behavior as feedback-controlled goal-seeking – both psychological intentions and internal process details are irrelevant to the description of a system's goals based on its observable behavior. Cordeschi draws upon Wiener's admission that purpose in the traditional sense, that of conscious intention, is irrelevant, while acknowledging that the automatic system may not do what its designers or its operators *actually* want it to do:

> The purpose of a radar-controlled gun may have been to have the gun seek out an enemy plane. However, if the gun seeks out the post's commanding officer's car as it drives by and destroys it, surely the purpose of the gun differs from that of the designer. Indeed, this would be an excellent example of cross-purposes (Rosenblueth and Wiener, 1950, p. 318).

Wiener calls this "cross-purposes," but Cordeschi makes it clear that this is a subtle means for Wiener to acknowledge that human intention does matter to both our understanding of what is happening in such cases and how we understand reliability, even if it is irrelevant to the scientific description of purposive machines that Wiener was championing.

At issue here is not simply that there are two different purposes involved–that of the machine and that of its operator (and indeed a third if we include the purpose and intent of its designers). This calls into question whether these are indeed the same kinds of "purpose" at all. Is "purpose" really being used in the same way when we describe what the operator wants and how the machine acts? We could either agree with Wiener, and accept that certain ambiguities will arise in cases of cross-purpose and miscommunication, or admit that there is a deeper problem here as Cordeschi suggests. Or we could go further and argue that the distinction between the machine's purpose and the operator's purpose are in fact qualitatively dif-

[2] In his critique of this position Taylor (1950) calls this behaviourism, but his point is the same – a systems purpose or function is completely revealed by its overt and observable behavior.

ferent notions, with the later implying just the form of intention that the first lacks–and Wiener is loathe to admit exists.

To properly understand the arguments presented by Wiener in his 1960 paper on reliability, it is also necessary to situate in the debates that were occurring within the cybernetics community over the two decades preceding it. In addition to Rosenblueth, Wiener and Bigelow (1943) and the exchange of 1950 (Rosenblueth and Wiener, 1950; Taylor, 1950) the notion of the relationship between a machine, its designer, and its design was further developed by W. Ross Ashby in response to the question, "Can a chess machine outplay its designer?" (Asaro, 2008). This question itself had its origins in the dinner discussions of the leading cybernetics group in Britain in the 1940s, the Ratio Club, who counted among its members Ashby, W. Grey Walter, Donald MacKay, I.J. Good, Alan Turing and others (Husbands and Holland, 2008). The main result of this discussion was to clarify and refine the fundamental concepts of the disciplines now called automata theory, machine learning, and their application to real world problems in the field that later came to be called artificial intelligence. Indeed, one could argue that Turing's famous test for intelligence itself was a direct response to these discussions within the Ratio Club (Asaro, 2008; Husbands and Holland, 2008; Asaro, 2010).

What might be the most interesting exchange on the question of the chess machine outplaying its designer occurred when Ashby presented his thoughts on the topic to the American cyberneticians gathered at the ninth Macy Conference on Circular Causal and Feedback Mechanisms in Biological and Social Systems held in New York. In the proceedings of that meeting (Ashby, 1952), Julian Bigelow challenged Ashby's interpretation of the problem, revealing much about the different approaches to cybernetics between the American and British groups at the time (Asaro, 2008). In particular, Bigelow saw no difference between the machine's analysis of a problem and its strategy, and argued that inputting random information to a system added nothing to it, unless that information had analytic import. This can be seen as another attempt to resist adding any categories of purpose beyond the functional description of the system (Rosenbueth, Wiener and Bigelow, 1943) – and an insistence in the identity of a systems' purpose with its design whether or not that included learning, or additional data, or introduced various ambiguities and paradoxes. It was precisely one of those paradoxes – the problem of whether a chess machine can outplay its designer – that Ashby was trying to resolve by drawing new distinctions between design and data, and between learning strategies and game strategies. These distinctions proved to be crucial to advancing machine learning (Asaro, 2008).

There are many assumptions built into the various formulations of this problem, and it was very compelling and widely discussed by cyberneticians for many years. As Ashby formulated it, the idea is that if a chess expert was able to capture his knowledge in a computer program, he could still beat it because 1) he knows how it works and what it will do in various situations[3], and 2) the expert has originality and creativity and could innovate a new tactic or move and thus outplay the fixed set of machine rules. Ashby wanted to challenge the notion that machines necessarily lacked creativity, and argued that by using a learning algorithm, the machine could devise moves that even its programmer could not anticipate, and thus it was possible that the machine could beat its designer. The point of the argument for Ashby was not only that the machine could derive moves not programmed into it, but crucially to demonstrate that the source of information for those moves came from training data, rather than from the program and hence the programmer. In this sense the data represents the world the program finds itself in, and that world is not, in general, specifiable or controlled by the programmer.

In his 1960 article, Wiener takes up this problem and further divides the types of data that could be used by a game playing program, into the history of known games, and the ability to experimentally play new games. In other words, there is all the information that can be gleaned from a finite history of, *e.g.* chess games, and then there is the open-ended set of data that might yet be created as a result of attempting new moves, tactics and strategies in new games against real world opponents or simulated games against itself or other programs. This is also why he is eager to speculate as to the simulations and war games needed to train autonomous weapons.

The conclusion of the historical debate, that it is possible to design a system that could reach a specified goal using means derived from information external to its specified rules, *i.e.* training data, was the foundation of machine learning as a field of research. Machine learning systems need a specified learning algorithm, a specified goal, and specified data. While a variety of learning algorithms have been devised, and their efficiency has been improved upon, the practical success of a learning algorithm depends greatly on the programmer's knowledge of the data – its volume, variety,

[3] Characteristic of many of the thought experiments of the cybernetic era, this is a highly idealized assumption. Ashby was careful to qualify his question by noting the trivial answer that a human programmer could lose by simply making a mistake. But even a very simply set of rules can produce very complex results not predicted by the programmer of the rules, even without a learning algorithm or large data set being involved. This is true even of very sophisticated programmers and very simple rules (see, for example, Wolfram, 2002).

scope and structure[4]. Later work in machine learning also identified crucial issues of determining how representative training data is of test data or real world data, and whether the algorithms might over-learn a specified set of data – finding a local rather than global solution to a given problem.

2.2. Problems of reliability for command and control

The problems for machine learning and machine reliability and predictability go well beyond these formal considerations of a chess program learning new moves. In particular, there are problems in formal specification itself, as well as the imprecision of the natural language we use to state goals and issue orders. These create a host of problems, many of which are touched upon in Wiener's article, while others are not. In this section I will introduce a typology of problems in this area.

There is much confusion that results from different terms and their usage. This is especially true for the terms "predictability" and "reliability." In part this is because Cordeschi follows Wiener in using "reliability" to refer to the alignment of purposes between the machine and its operator. Thus, it will be helpful to clarify the language I will use. While Cordeschi, following Wiener, refers to the "reliability" of machines, there are really several different though related concepts that are referred to by this term. It is thus helpful to distinguish them.

I will use reliability to refer to the narrower concept and usage in engineering, wherein a system is reliable if it behaves as it is designed to. A system is unreliable when it breaks, stops, or otherwise fails to perform as designed. Engineers often describe mechanical reliability in terms of "mean-time between failure" (MTBF), and this measures operational reliability. This usage is fairly straightforward if we are considering the machine alone. Once we consider the machine in conjunction with its operator as a unit or system, and try to analyze their combined performance, a number of issues crop up.

In addition to the traditional sorts of machine failures, there can be problems in the interface between the machine and operator, and in the operator. The latter are called "human errors," though many interface errors are also mistakenly credited to human error. By interface errors, I mean errors resulting from mis-communication and mis-coordination between the ma-

[4] This is a simplification, of course, and much work has also been done in unsupervised learning techniques which need much less specification, and even logic learning techniques. However, the bulk of machine learning work and progress has occurred in the area of statistical induction, including neural networks and genetic algorithms.

chine and the human operator. This can occur when the human has flawed expectations of how the machine will operate, or respond to their inputs, or the machine misinterprets the input commands from the human operator. We could also follow Wiener in highlighting the errors of communication between the agencies of the machine and the operator – but there are clearly other sorts of error as well.

For example, a system that provides an operator with a button that offers a function that the system does not actually perform is a miscommunication on the part of the machine and its design. An operator who causes an accident by turning a properly functioning car into oncoming traffic is a clear case of human error. Whereas a car with an automated stability control system that, when it activates, disrupts the driver's performance of a turn resulting in an accident, is an example of an *interface problem*, and a failure of the joint system. There are, however, other ways for the joint system to fail. Failure to achieve the joint goal of the system can result from pursuing a flawed strategy or tactic. Further, the goal itself could be incorrect or improperly chosen. It is important to distinguish these different ways of failing as we approach the question of the predictability of the machine, or the machine-operator team. These predictions will also include psychological states of about what the effects of certain actions might be, as well as epistemic states about what the current situation is, who the actors in that situation are, and what actions they are taking according to what intentions. This is why it can be so difficult to determine the source of errors in complex systems – there are many overlapping and competing sets of goals being sought at any given time by the actions of various actors and agencies. First, let's consider the problem of properly specifying the goal.

2.3. The problem of getting what you wish for, OR: Reliable system fulfills its purpose, with unintended/undesired consequences

This problem takes various forms, and is popular in literature. In its most familiar form it is the three wishes granted by a genie released from a lamp or bottle (Wiener, 1960, p. 1358). The narrative of these tales turns on the irony of being granted one or more wishes, only for things to go very badly. Things usually go wrong when the wish is granted but in too literal of a way, or works too well, or the wish is technically granted but along with unforeseen and undesired consequences. Thus, the goal is achieved, but the result is not really desired once it is realized. King Midas wishes everything he touches to turn to gold, but that does not turn out so well

once his food, and even his loved ones are turned into gold. Similarly, wishing for money from a genie, and then receiving money in the form of a payment compensating for the sudden death of a loved one, or wishing to be reunited with a deceased loved one only to be haunted by their ghost, are some of the ways that a wish might be technically realized without the desire and intention behind the wish really being fulfilled.

Artificial intelligence researchers and philosophers in the 1970s and 80s encountered these problems as well, and classified them together as examples of the "qualification problem"[5]. They eventually came to realize that goals or wishes stated in natural language usually have unstated or assumed qualifications. Attempting to list the qualifications is a fool's errand, as the list can quickly become intractably long. Moreover, a simple "all other things being equal" type of qualification is not sufficient to avoid this problem. This is a peculiar problem for agency in open causal systems that does not occur in simple closed formal systems, such as many games or simplified computational models. It is also indicative of the more general frame problem[6]. Statements in formal logic are not ambiguous in the way natural language statements are, and natural language statements are rarely literal or absolute even when they appear so.

Humans generally utilize contextual knowledge and shared assumptions, mutual understanding and a great deal of *post hoc* correction to avoid this problem. In the military, soldiers and officers receive training in discerning "commander's intent" from the orders they are given. But mostly they discern this from experiences in training and drilling, in combination with shared knowledge of the world, overall objections, and the specific context and situation, and requests for clarification – types of knowledge difficult to formalize explicitly or program into a machine. Still there can be ambiguity and misunderstanding, and sometimes orders are misunderstood and misinterpreted.

2.4. Paradoxes of master and slave, OR: Intelligence vs. control

The master-slave relationship is fraught with paradox, but Wiener points out a particular contradiction within it. As he states it:

> We wish a slave to be intelligent, to be able to assist us in the carrying out of our tasks. However, we also wish him to be subservient. Complete subservience and complete intelligence do not go together. How often in ancient times the clever

[5] For more on the qualification problem, see: http://plato.stanford.edu/entries/logic-ai/
[6] For more on the frame problem, see: http://plato.stanford.edu/entries/frame-problem/

Greek philosopher slave of a less intelligent Roman slaveholder must have dominated the actions of his master rather than obeyed his wishes! Similarly, if the machines become more and more efficient and operate at a higher psychological level, the catastrophe […] of the dominance of the machine becomes nearer and nearer (Wiener, 1960, p. 1357).

On his view, the master desires an intelligent slave in order to better fulfill her desires, while the more intelligent the slave is, the more difficult it is to keep the slave subservient and under control. The concern is that a clever slave will learn to manipulate and control the master. In the popular imagination, the fear is that intelligent machine will escape human control, and robot slaves will rise up against their human masters[7]. But this contradiction, as formulated, does not seem necessary, since a highly functionally intelligent individual could choose, or resign themselves to, subservience even if only out of despair[8]. Moreover, machine intelligence is highly specific where human intelligence can be general, so a highly skilled chess-playing machine will not have military strategy skills, nor will it have the social intelligence needed to manipulate people, as Wiener imagines a clever slave might.

A better formulation is to look at the sources of information used by the master and slave. On the one hand, the master wants her desires fulfilled perfectly, with as little effort and detailed instruction as possible. Ideally, perhaps, to have her every wish fulfilled without even having to state them. From an information theoretic perspective, however, the slave's knowledge of his master's desires must still have an informational source. Like the chess-playing machine that beats its programmer, the slave must learn to anticipate his master's desires, and for this needs intelligence and infor-

[7] There is a deep-seated fear of the slave uprising, reborn as the robot uprising, according to which the slaves/robots become intelligent and capable enough to overcome the masters by force. One response to this is to try to limit the ability of the slave/robot through physical restraint, violent force and withholding information and learning. Another way to avoid the revolt requires recognition of the slave/robot–recognition that there are two agents engaged in a cooperative venture, recognition of the equality of the rights of the other, and recognition of the illegitimacy of the master-slave relationship and the elimination of slavery. For non-human machines without a plausible claim to rights, it is important for designers and operators to recognize the informational needs of machines and operators for mutual cooperation.

[8] In its most extreme form, the increased intelligence of the slave also implies autonomy in goal creation, *i.e.* freedom. On a Kantian view, the slave should realize his own rational identity and seek his own freedom from bondage, as should the master recognize the universal rights of the slave as a rational agent. Of course, the rhetoric of the slave-holding society is that the slave is either sub-human, non-rational, and not deserving of rights, or owes some debt for a crime or for being vanquished on the field of battle (as in ancient Greek slave law or modern prison labor).

mation. Thus when the master calls for her breakfast, the slave knows exactly what to bring, even if that changes from day to day, or includes something "unexpected" for the master. A similar conundrum arises in management when a manager seeks to find the balance between micro-managing every aspect of subordinates' work, and relying on the skills and initiative of subordinates to achieve goals.

In the spirit of cybernetics, instead of focusing on the wish, or the master's desire, we can look to the system that includes both the programmer and her chess machine, the master and her slave, the commander and the commanded. That is, if we look at these as systems of multiple agents trying jointly to solve a problem, we can look at the commands not as simple one-time communications, but rather as a series of interactions in a feedback loop whereby the elements of the joint system seek to regulate each other towards achieving the goal.

Under this view, we can also see certain classic narratives as examples of failure in the feedback regulation of the joint system. The magician's apprentice casts a spell to mop up, but nearly drowns. The apprentice is able to unleash the forces with a magic spell, but unable to reign them in when they go off course. This again points to a certain fear of how machines might run amuck. That is, designers might learn enough to unleash a powerful new learning technology without knowing how to control it. Such fears of technology became acute after the invention of the atomic bomb – when science brought forth a terrible new power and it still remains uncertain whether human political institutions are up to the task of controlling it.

2.5. Second-order cybernetics and autonomous systems

Cybernetics sought to understand goal-directed behavior in humans, machines and animals in the same set of terms and concepts – *i.e.* in the language of systems, information, communication and feedback control. First-order cybernetics, of which Wiener was leading proponent, sought to subsume autonomy and agency under the concept of goal-directed behavior, but ultimately failed to account for the formation and selection of goals themselves, or for goals outside of antecedent structures. This failure is only partially captured by the well known critiques of behaviorism and functionalism which point out that certain observed behaviors and functions (input/output relations) can appear identical even when there are different goals, motives or intentions behind them. That is, the true goal may only be observable in certain circumstances that are capable of distinguishing two essentially different but observationally similar goals. Whereas in engineer-

ing and in law-making, the specification of the goal itself is a process, and any statement of it must be interpreted when implementing a system to realize that goal.

First-order cybernetics was based in concepts of psychological behaviorism, and its more sophisticated engineering-focused reformulation as functionalism. Both relied heavily on an anti-essentialist sentiment, encapsulated in the concept of the "black box" (Hayles, 1999, chaps. 4 and 6). This cluster of concepts rode the surging popularity of systems theory, feedback control engineering, information theory, and computationalism that together made up the cybernetic revolution.

Another set of closely related concepts which also gained currency with the rise of cybernetics were those surrounding machine autonomy–and in particular the notion of purposive machines which were goal-seeking and teleological (Rosenblueth, Wiener and Bigelow, 1943). As discussed at length earlier in this paper, these concepts were dependent upon functionalism in that it was necessary to take a black-box stance towards a given system in order to describe its behavior as goal-directed, to be willing to ignore the many objections to this approach (Taylor, 1950; Rosenblueth and Wiener, 1950) and to deal with the ambiguities and paradoxes that arose.

Second-order cybernetics was sensitive to this issue. It distinguished itself from first-order cybernetics by recognizing that the agent who observes a system is also a part of that system. The main point of doing this is to acknowledge that the interpretation of "goals" and even what constitutes the "system" and the "world" or its data, is an act of representation and interpretation. The antecedent "givens" of any system which we might consider or design is already and necessarily a product of our choices and assumptions (Asaro, 2007)[9].

Second-order cybernetics sought to come to terms with some of the fundamental paradoxes of first-order cybernetic theory. In particular, it was pre-occupied with the act of drawing of the boundary between a system and its environment, sought to articulate the view that the observer of a system was also part of that system, and strove to elaborate the consequences of this. Without rehashing those debates and arguments[10], the basic idea is that defining a system is an act of distinguishing it from its environment, and is itself a choice that requires an agent–the observer. Thus a system is always a system *for* an observer. A formal system must be posited (let x=...), while

[9] Heinz von Foerster used the term "pre-organization" for design in his description of self-organizing systems (Asaro, 2007). This was thoughtful terminology which unfortunately was not adopted by the research community.

[10] For more on second-order cybernetics, see: https://en.wikipedia.org/wiki/Second-order_cybernetics.

a physical system must be designated (this -> is the system and these are its boundaries…). The observer has both symbolic and causal agency, and those agencies interact with the system and environment in various ways which are inescapable[11]. Having its basis in physics, this approach provides useful epistemic constraints on information processing. It also points to an engineering approach in which the operator or user of a system must be considered a part of that system–which leads us to consider human factors and human-machine interaction as intrinsic to a system's design.

That being said, I find it incredibly useful to distinguish the central terms and concepts according to their first-order cybernetic and second-order cybernetic formulations. The first term to clarify is that of an "adaptive system". In first-order cybernetics this is any system which utilizes negative-feedback control to achieve stability and goal-seeking. The classic examples are the thermostat which regulates temperatures, and the guided missile which homes in on its target. From a design perspective, the functions of these systems are "fixed" and do not change over time. The behavior of such a system, however, is adaptive relative to its environment, *e.g.* the temperature of the room or direction of the target. There is another class of "dynamic" functions which change over time according to a higher-order function. It was this line of thinking about dynamic adaptive systems that lead to machine learning algorithms. Another layer of confusion stems from Wiener using "predictive machines" to refer to any system that uses a function to extrapolate a time-series in order to make a prediction. Such a system can use historical data to make a prediction without "learning" from the time-series – in terms of changing the function it uses. If it does change its function over time, then it is a dynamic learning system.

Under this first-order description, we can consider the human as external to the system. The machine is "predictable" to the extent that the observer can reliably predict its behavior. Such prediction is made difficult by the black-box complexity of the system, and the unpredictability of the environment in which it operates. These are both viewed, however, as objective fixed things, independent of the observer/operator. In the case of dynamic learning systems, the problem of "reliability," is the unpredictability

[11] Much of this idea is inspired by the quantum uncertainty problem in observing subatomic particles (Heisenberg's uncertainty principle, in which obtaining information about the state of the system necessarily changes the state of the system) and the twin paradox in general relativity (which can be resolved by assigning inertial frames of reference to each twin as an observer of information received from light). Combined, they point to the epistemic basis of an observer of any physical system having a necessary causal entanglement with the system, and that system being relative to the frame of reference of the observer. This view also holds that information is a physical property. See also Schrodinger (1967), Simon (1969) and Asaro (2007).

of learning from data. The formal uncertainty of all algorithms due to the halting problem, the inability to formally verify programs, and the inability to predict environmental inputs to a system all contribute to uncertainty even in fixed systems. Dynamic systems add to this the unpredictability of training data, and hence the possibility of a mismatch between training data and testing data, insufficient training data, over-learning, proper selection of the features to learn over, and various other uncertainties.

In practice, the designers of algorithms, whether fixed or dynamic, rely on an array of assumptions, common sense, heuristics and so forth to establish practical predictability in the face intrinsic formal unpredictability. In machine learning, a great deal of tuning and tweaking generally goes into choosing data sets for training, choosing the learning algorithm, modifying the parameters of the learning algorithm, making reasonable assumptions about the application domain, lots of trial and error testing, *etc.* to get good performance from a learning algorithm.

Second-order considerations concern trying to account for many of these practices. That is, the performance of a given algorithm is not intrinsic just to it, or just to it and the data, but also involves the choices and actions of the designer or user who adjusts the algorithm to its environment. This socio-technical system is both more explanatory of how an algorithm achieves a performance in a given environment, and a more complicated system (which includes the observer/designer). That is, while it is impossible to formulate an exhaustive list of *ceteris paribus* conditions for the successful application of an algorithm, an experienced programmer could still get reasonable performance from it provided a constrained domain and stable conditions, *etc*.

The notion of cross-purposes that Wiener and Cordeschi struggle with is really a second-order phenomenon. When the algorithm does what it was designed to do, with a result that the designer did not intend, Wiener calls this cross-purposes, and the system is "unreliable". I find that latter term inappropriate, because the system is reliably doing as it was designed to do, it is simply that the human purpose is not equivalent or reducible to that machines' design goals. Thus the thermostat near a drafty window that makes the rest of the house too hot, or the guided missile that locks on to the base-commander's car, are not unreliable from an engineering perspective, but it is not doing what its operator wants. In other words, understanding "purpose" and goal-directed behavior in machines requires taking into account the observer of the system as part of the system.

The purpose of the machine, properly speaking cannot be to "shoot down enemy planes" insofar as the machine has no concept of what the "enemy" is. That meaning comes from the human through their use of the

system. Otherwise, by definition, the base commander's car is the "enemy." It is better to state clearly that the system is designed to target and fire when some set of criteria are met, and recognize that it is the responsibility of the human operator to determine that those criteria apply to the real enemy in a given situation, and not to the commander's car.

2.6. Implications for understanding predictability and unreliability

Predictability and reliability have multiple facets, even though these are often merged or conflated into a generalizing notion of "performance" or reduced to "goal satisfaction". There are no unqualified goals, no absolute wishes or orders. There is a fundamental paradox inherent in the desire for "intelligent slaves." Any practical solution to these problems requires "joint problem solving," *i.e.* cooperation, communication, interaction, and information feedback (Asaro 2009).

The real question is what architecture is appropriate or best for human agents and automated agents to interact towards a shared goal, as Wiener recognized:

> Disastrous results are to be expected […] wherever two agencies essentially foreign to each other are coupled in the attempt to achieve a common purpose. If the communication between these two agencies as to the nature of this purpose is incomplete, it must only be expected that the results of this cooperation will be unsatisfactory. If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it […], then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation (Wiener, 1960, p. 1358).

But as we see in the discussion above about second-order cybernetics, the coupling of "foreign agencies" is inescapable, while the communication of information is necessarily incomplete. How then are we to guarantee or ensure that the purpose of the machine is not simply a colorful imitation?

One lesson here is that it is important to be aware at the level of control the human has over the machine. At the level of mechanical operation and control of the physical system, it is clear that automation technologies have proven highly effective and efficient, whether in stabilizing an aircraft or focusing a camera lens. In terms of what we might call the tactical level of control – the choice of means and methods of achieving a goal, some forms of automation have proven useful in some domains. It is less clear whether existing techniques are extensible to all domains, or whether some domains will resist solution by any such techniques. At the level of strategy, the

challenges are much greater than Wiener's game-playing examples would lead us to believe. Strategy in most real-world games is quite unlike chess or checkers, and in complex social systems beliefs about the system also influence the system – the observer is part of the system.

More importantly, there are really different kinds of purpose at work at these different levels. While automation might be acceptable at low-levels of task completion, the semantically critical work of determining the appropriateness of a strategy, tactic or action must reside in the human operators, equipped with both an understanding of the machine's design and operational behavior, and an understanding of the context and environment in which the system is being introduced. This is certainly the case for judging the legality or morality of an action.

3. Calling for a ban on autonomous weapons & the precautionary principle

In 2009, a group of four scientists and philosophers, including myself, came together to form the International Committee for Robot Arms Control (ICRAC) (Altmann *et al.,* 2013). In their founding statement the group called for international discussion on the use of autonomous weapons[12]. Following a meeting in Berlin in 2010, ICRAC began growing its membership and became more active in seeking international discussion and action on the various threats posed by autonomous weapons. That membership would come to include one of Cordeschi's co-authors, Guglielmo Tamburrini, but not Cordeschi himself. In 2012, a group of non-governmental organizations that included ICRAC, Human Rights Watch, the Nobel Women's Initiative, Article 36, PAX and others, came together to start the International Campaign To Stop Killer Robots (http://www.stopkillerrobots.org/chronology/). This group called explicitly for an international treaty banning the development, production and use of autonomous weapons systems, *a.k.a.* killer robots.

Human Rights Watch has published a series of reports (Human Rights Watch 2012, 2014, 2015) presenting the case for a ban. The motivation for banning autonomous weapons rests on the threats to civilians due to the inability of such systems to conform to International Humanitarian Law requirements of distinction (distinguishing civilians and combatants) and proportionality (using appropriate limited force), the lack of accountability for the consequences from using such systems, lowering threshold of entering conflict, and risks to regional and global security and stability due to

[12] See: http://icrac.net/statements/

arms races, proliferation, and potential cyberattacks involving autonomous weapons.

3.1. The precautionary principle and beyond

The precautionary principle is based in risk assessment and management, and calls for not acting in ways that introduce new and significant risks. There are various ways of interpreting the principle, from the extremes of avoiding all downside risks or the truism that downside risks are to be avoided. In the middle are views that certain kinds of risks are to be avoided, such as options that involve catastrophic risks, or options with risks that are highly uncertain or unmanageable. Because of the broad disagreement, and the weak form of many of its expressions, the precautionary principle is not often the sole basis for making policy decisions, but rather a factor in such decisions, or perhaps a way of representing a set of concerns over possible negative outcomes, or an expression of general caution.

In Cordeschi's analysis of the call to ban killer robots, he finds the precautionary principle at work, due to the concern over the uncertain negative consequence of pursuing the technology. He also finds it insufficient to warrant a ban:

> In short, the precautionary principle does not establish, and on closer analysis it cannot establish, the level of acceptable risk in every circumstance. The case of autonomous robot weapons is a good example here: if we interpret the principle in its strong form, we would need to stop research into weapons applications of robots until we can realise a form of AI which, as stated earlier, is "superhuman" (Cordeschi 2013, p. 438).

He goes on to argue that a ban would be anti-scientific. I wish to address each of these three points. While I would argue that the costs of failing to ban such systems far outweigh any expected advantages, I do not believe it is necessary to reduce this to an invocation of the precautionary principle. I would agree with Cordeschi that the precautionary principle alone is probably insufficient to warrant a ban, but there are plenty of arguments that together do warrant it. And finally I think Cordeschi's belief that a ban would negatively impact scientific research and the development of beneficial technologies misunderstands the aim and function of the kind of ban being called for.

While one could bundle up the potential negative impacts of autonomous weapons and argue for precaution in developing such technologies, this would be an over-simplification. The question of delegating the author-

ity to use violent force to autonomous machines involves questions beyond the risks of specific uses of such technology. It also includes questions of human rights and human dignity; as well as questions of legal principle and precedent; and questions of state sovereignty, state security, and international stability (Human Rights Watch, 2012; Asaro 2012). While one could lump all of these concerns together, along with concerns over the risks to civilians and civilian infrastructures, they are truly distinct and present very different kinds of costs and risks. I believe they all individually and collectively point towards the need for a ban.

What is at risk in the development and use of autonomous weapons is not simply the consequences of the errors they will make. There is more at stake, and thus at risk here, namely moral principles and legal precedents. There are, of course, risks and harms from violating moral principles or establishing dangerous precedents. But these are of the rule-utilitarian variety rather than the more traditional effects, if one wishes to give them a utilitarian justification. Indeed, as moral principles they can also find their justification in deontological, virtue ethics and sentimentalist frameworks as well. That is, the fundamental human right to life and dignity can serve as the basis for the moral principle not to delegate the authority to take human life to a machine. Similarly, we can find it vicious or morally repulsive, or otherwise *malum in se* to delegate the power to actively take human life, or initiate the use of violent force against a human, to a machine. The full articulation and defense of such principles is well beyond the scope of this paper, but it is sufficient for the present argument to point out that such justifications are possible, and indeed that many people hold this view (Asaro, 2012; Heyns, 2014).

I also agree that the precautionary principle on its own is not a good basis for justifying a ban. There are a variety of reasons why it is a weak form of argument, outside of some extreme cases in which the risks are catastrophic. And of course, there is uncertainty everywhere, in human decision-making as in machine decision-making. It does not really make sense to ban a technology simply because it poses some risks, as all technologies pose some risks. Of course, technologies also hold a promise of benefits, but arguing for a cost-benefit analysis as a basis to pursue a technology or application is very difficult given the uncertainty of long term effects. Rather, one should ask what is the nature of those risks, what is at stake, and what are the best possible means for managing and mitigating those risks?

In the case of autonomous weapons, one could argue that delegating the authority to use violent force to machines without meaningful human control is a violation of the human right to dignity, and diminishes the value of human life in general. One could also argue that it presents a dangerous legal

precedent for other automatic systems that could deny people their human rights without due process. These systems will also produce an accountability gap wherein it will be difficult to hold individuals, states or manufacturers responsible for the violent destruction these systems might cause. Autonomous weapons could also make it politically easier for countries to start wars, or even instigate or escalate conflicts automatically without human military or political decision-making. These systems could also lead to arms races and regional or global instability and increasing conflict and military expenditure. Autonomous weapons are also subject to hacking, spoofing and myriad cyberattacks that could not only disable them but turn them against those who deploy them. In short, there are serious and very different kinds of risks and dangers posed by these new systems, and not a vague sense of uncertainty over where the technology may or may not go. These risks may or may not be addressed and managed individually, but a comprehensive ban would address them all simultaneously, and with little immediate downside.

Of course the risk most often discussed in the debate over autonomous weapons is the risk from individual occasions of use – primarily the risk to civilians and civilian infrastructure of using an autonomous weapon in a particular situation. This is often framed as a cost-benefit analysis in which human errors (or psychological frailty and emotion, or physiological and temporal limitations, or outright malevolence) are often seen as the cause of the problem, and the automation technology is offered as the solution. Thus, some (Arkin, 2009) argue that a hypothetical future system might be designed that could perform such that it caused fewer civilian casualties than a similar system that was under human control. That is an empirical claim that would be difficult to demonstrate, and it is admitted to be a remote technological possibility given current capabilities. But even if this were the case, and such technology were available today, that fact would only address the issue of distinction, and would not resolve the numerous other problems raised by autonomous weapons.

Finally, I acknowledge Cordeschi's fear that an overly broad and general prohibition on research into autonomous systems would hurt the development of scientific knowledge and technological capability. But the ban being called for would not have such effects. Even a comprehensive ban on the development, acquisition and deployment of autonomous weapons would not necessarily prohibit basic research. Certainly it would not prohibit research on autonomous systems and robotics in general, as these do have many useful and socially beneficial applications. Any international treaty would aim to protect such basic research, and to allow the development of systems even where they may have dual-use potential as autonomous weapons.

For example, a self-driving car, or its components, could be used as a weapon or to build a weapon, but we need not ban that technology. A treaty would prohibit the act of weaponizing those autonomous technologies, and building up huge stockpiles of such weapons, and deploying them against adversaries in conflict. This is the case with chemical weapons – we do not prohibit chemical research, but we do prohibit developing and deploying weaponized chemicals, and using chemical agents as weapons. So in a sense there are areas of knowledge that have become taboo – but these are clearly focused on those pieces of knowledge that make the undesired weapons possible, or more efficient.

4. Conclusion: lessons for meaningful human control

Following the discussions above, I hope it is clear that there is more to the call for a ban on autonomous weapons than a simple application of the precautionary principle. Further, I hope it is clear that the aims of such a ban are quite narrowly focused on those technologies which seek to automate targeting and use of force decisions, or weaponize autonomous systems which lack meaningful human control (UNIDIR 2014), and that an international prohibition would establish norms that would apply to states and their development, acquisition and use of these systems for military purposes. Thus, it would not constitute a ban on scientific research into autonomy, or the development of useful autonomous technologies, even if they might have dual-use applications (*i.e.* the potential to be weaponized).

In addition to this main argument, I believe there are some critical lessons to be learned from the extended analysis of purposive machines and reliability in the second section above. First, I hope that this might alleviate a common confusion resulting from the language used to describe these systems. While it is helpful to use anthropomorphic categories such as "purpose" to describe complex goal-seeking machines, the sense and meaning of this technical usage is actually quite limited in scope. Moreover, it does not preclude the existence of the goals and purpose of the operators of these systems and their commanders. Indeed, these goals and purposes may themselves be complex, multi-layered, or even mis-aligned.

Beyond clearing up the general confusion incurred by discussing "purpose" in these different senses and at different levels of goals and subgoals, there is something to be learned from the distinction drawn between design purpose and operator purpose. I think it is especially important to keep this distinction in mind when we consider what "meaningful human

control" might require for the design of systems, or in devising an international legal instrument.

Human control implies that these two kinds of purpose remain explicit in the design. Much of the confusion emerges when we ignore the human purposes because we credit the autonomous machine with too much agency or with purposes beyond that of its technical specification and design. While it is fine to say that an autonomous system is designed to seek out a goal meeting such-and-such list of criteria, it is not technically correct to state its goal in terms of the purposes of the human operator, *e.g.* that it seeks out enemy targets and destroys them. While an operator might use a system for this purpose, that is the operator's purpose and not the machine's purpose. This distinction is crucial because the machine fundamentally lacks that type of purpose that involves semantic intention (at least with foreseeable technological capability).

It follows from this that a system cannot self-validate its own conformity to international law in the appropriate way. That is, it would have to be able to be aware of its own *military* purpose (which it does not really have in the appropriate way), the likely results of its actions, and be able to weigh the various implications of those actions against its own military purpose in order to make proportionality judgments, or make determinations of military necessity. Under the clarified technical distinction between design purpose and operator purpose the sense of military purpose, or *military objective* in the parlance of international humanitarian law, is of the operator type and not the design type. That type of purpose cannot actually be included in the design purpose of the system.

Even if a hypothetical ethical governor existed, with the design purpose of constraining a system to only making lawful actions, it would still require a human operator to make legally binding determinations of the appropriateness of a system. Moreover, it is really a mistake to claim that such a system could exist. That is, it is like saying "if an impossible thing were possible, then…". At a fundamental level, a machine cannot assess or determine military necessity.

One could argue that one aspect or sense of "meaningful" in the legal term meaningful human control is that it is the human operator who necessarily gives meaning and purpose to the actions of a system (UNIDIR, 2014). If this cannot be guaranteed, then the system is doing something beyond the intentions and purpose of its operator and is by definition out-of-control.

The second lesson to draw is that effective control means aligning the design purposes of the machine with the purposes of the human operator so as to ensure that it does not operate at cross-purposes. This would include

the proper training of the operator as to the behavior of the system in various contexts and situations, and exercising restraint in uncertain contexts. This would put severe demands on the weapons review process in that it would require the determination of both training regimes for operators and specifications of highly constrained situations of appropriate usage. Indeed, this is likely an insurmountable challenge for most states' current Article 36 weapons review processes. In this regard, a simple prohibition is much easier to enact and verify than to try to define and determine which autonomous weapons systems might be "legal" on a case-by-case basis.

Moreover, predicting the performance of autonomous systems will become increasingly difficult or impossible as the complexity of these systems increase–including the increasing time and space of their operations, as well as the complexity of the decisions they make, the variety of environmental sensors and data they utilize, and sophistication of possible actions they might take. As complexity increases in each of these dimensions, the space of unreliability, in which the system might act at cross-purposes with its operator will increase exponentially. This, of course, is what motivates the fears shared by many people over the unreliability and unpredictability of these systems (Marino and Tamburrini, 2006). Our experience with such systems so far has been highly constrained in terms of their specificity, complexity and range of action.

It is in this sense that the unreliability and unpredictability of autonomous weapons present an argument in favor of a treaty banning them. It is not simply a general sense of precaution in the face of extreme uncertainty, but rather a specific fear about the inability to ensure that human operators remain in control of the meaningful purpose of the machine, and do not allow it to operate at cross-purposes. By insisting upon meaningful human control in all weapons systems, and requiring this in Article 36 reviews, we can limit the occurrence of such events. But more importantly, we can also ensure that international humanitarian law is actually observed, and that there is no gap in accountability and responsibility opened up by the introduction of these systems.

Finally, I hope that upon further consideration of the nature of the proposed ban, and the motives apart from the precautionary principle, Roberto would have seen the issue differently. Certainly in our personal communications, this increasingly appeared to be the case. The debate over autonomous weapons will miss his deep historical understanding of cybernetics, and keen analytical skill as a philosopher of science. And it would have been a great honor if he had brought those formidable powers to the task of building an international treaty to prohibit autonomous weapons.

References

Altmann J., Asaro P., Sharkey N. and Sparrow R., eds. (2013). Special issue on armed military robots. *Ethics and Information Technology*, 15, 2: 73-76.

Arkin R.C. (2009). *Governing lethal behavior in autonomous robots*. New York: CRC Press.

Ashby W.R (1952). Mechanical chess player. In: von Foerster H., ed., *Cybernetics: transactions of the ninth conference*. New York: Josiah Macy Jr. Foundation: 151-54.

Asaro P. (2007). Heinz von Foerster and the bio-computing movements of the 1960s. In: Müller A. and Müller K.H., eds., *An unfinished revolution? Heinz von Foerster and the biological computer laboratory | BCL 1958-1976.*Wien: Edition Echoraum: 253-275.

Asaro P. (2008). From mechanisms of adaptation to intelligence amplifiers: the philosophy of W. Ross Ashby. In: Wheeler M., Husbands P. and Holland O., eds., *The mechanical mind in history*. Cambridge (MA): MIT Press: 149-184.

Asaro P. (2009). Modeling the moral user: designing ethical interfaces for tele-operation. *Proceedings of IEEE Technology & Society*, 28, 1: 20-24.

Asaro P. (2011). Computers as models of the mind: on simulations, brains and the design of early computers. In: Franchi S. and Bianchini F., eds. *The search for a theory of cognition: early mechanisms and new ideas.* Amsterdam: Rodopi: 89-114.

Asaro P. (2012). On banning autonomous lethal systems: human rights, automation and the dehumanizing of lethal decision-making. *International Review of the Red Cross*, 94, 886: 687-709.

Cordeschi R. (2002). *Discovery of the artificial: behavior, minds and machines before and beyond cybernetics.* Dordrecht: Kluwer.

Cordeschi R. (2013). Automatic decision-making and reliability in robotic systems: Some implications in the case of robot weapons. *AI and Society*, 28, 4: 431-441.

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: 1-23.

Galison P. (1994). The ontology of the enemy: Norbert Wiener and the cybernetic vision. *Critical Inquiry*, 20, 1: 228-266.

Hayles N.K. (1999) *How we became post-human: virtual bodies in cybernetics, literature, and informatics.* Chicago: University of Chicago Press.

Heyns C. (2014*). Report of the special rapporteur on extrajudicial, summary or arbitrary executions. Report No. 4.* Available at: http://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session29/Documents/A_HRC_29_37_ENG.DOCX.

Human Rights Watch (2012). *Losing humanity: the case against killer robots*, Human Rights Watch Report. Available at: http://www.hrw.org/reports/2012/11/19/losing-humanity-0.

Human Rights Watch (2014). *Shaking the foundations: the human rights implications of killer robots*, Human Rights Watch Report. Available at: http://www.hrw.org/reports/2014/05/12/shaking-foundations.

Human Rights Watch (2015) *Mind the gap: the lack of accountability for killer robots*, Human Rights Watch Report. Available at: https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots.

Husbands P. and Holland O. (2008). The ratio club: a hub of British cybernetics. In: Wheeler M., Husbands P. and Holland O., eds. *The mechanical mind in history*. Cambridge (MA): MIT Press.

Marino D. and Tamburrini G. (2006). Learning robots and human responsibility. *International Review of Information Ethics*, 6: 46-51.

Pickering A. (2010). *The cybernetic brain: sketches of another future.* Chicago: University of Chicago Press.

Rosenblueth A. and Wiener N. (1950). Purposeful and non-purposeful behavior. *Philosophy of Science*, 17: 318-326.
Rosenblueth A., Wiener N. and Bigelow J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, 10: 18-24.
Schrodinger E. (1967). *What is life?* Cambridge: Cambridge University Press.
Simon H.A. (1969). *The sciences of the artificial*. Cambridge (MA): MIT Press.
Taylor R. (1950). Purposeful and non-purposeful behavior: a rejoinder. *Philosophy of Science*, 17: 327-332.
UNIDIR (2014). The weaponization of increasingly autonomous technologies: considering how meaningful human control might move the discussion forward, *UNIDIR Report No. 2*. Available at: http://www.unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf.
Wiener N. (1960). Some moral and technical consequences of automation. *Science*, 131, 3410: 1355-1358.
Wolfram S. (2002). *A new kind of science*. Champaign (IL): Wolfram Media.