Published as Human-Machine Autonomies (2016) In N. Bhuta, S. Beck, R. Geis, H.-Y. Liu, & C. Kreis (Eds.), *Autonomous Weapons Systems* (pp. 75-102). Cambridge, UK: Cambridge University Press.

Human-Machine Autonomies

Lucy Suchman and Jutta Weber

> We are responsible for the world of which we are a part, not because it is an arbitrary construction of our choosing but because reality is sedimented out of particular practices that we have a role in shaping and through which we are shaped.
>
> Karen Barad, *Meeting the Universe Halfway*[1]

> [R]esearch and development in automation are advancing from a state of automatic systems requiring human control toward a state of autonomous systems able to make decisions and react without human interaction. DoD will continue to carefully consider the implications of these advancements.
>
> US Department of Defense, *Unmanned Systems Integrated Roadmap*[2]

This chapter takes up the question of how we might think about the increasing automation of military systems not as an inevitable 'advancement' of which we are the interested observers but, rather, as an effect of particular world-making practices in which we need urgently to intervene. We begin from the premise that the foundation of the legality of killing in situations of war is the possibility of discrimination between combatants and non-combatants. At a time when this defining form of situational awareness seems increasingly problematic,[3] military investments in the automation of weapon systems are growing. The trajectory of these investments, moreover, is towards the development and deployment of lethal autonomous weapons – that is, weapon systems in which the identification of targets and the initiation of fire is automated in ways that preclude deliberative human intervention. Challenges to these developments underscore the immorality and illegality of delegating

[1] K. Barad, *Meeting the Universe Halfway* (Durham: Duke University Press, 2007), 390.

[2] US Department of Defense (DoD), *Unmanned Systems Integrated Roadmap: FY2013–2038* (DoD, 2013), 15.

[3] Christiane Wilke observes that the figures of civilian and combatant are not only gendered and aged (women and children being the canonical instances of the first category) but also raced. Both, moreover, are increasingly problematic, as 'the rise of the figure of the "unlawful combatant" … is accompanied by a corresponding rise of the figure of the illegitimate, non-innocent, suspicious civilian'. C. Wilke, 'Civilians, Combatants and Histories of International Law', 28 July 2014, available at http://criticallegalthinking.com/2014/07/28/civilians-combatants-histories-international-law/. See also D. Gregory, *Keeping Up with the Drones*, 20 November 2014, available at http://geographicalimaginations.com/2014/11/20/keeping-up-with-the-drones/.

responsibility for the use of force against human targets to machines, and the requirements of international humanitarian law that there be (human) accountability for acts of killing. In these debates, the articulation of differences between humans and machines is key.

The aim of this chapter is to strengthen arguments against the increasing automation of weapon systems, by expanding the frame or unit of analysis that informs these debates. We begin by tracing the genealogy of concepts of autonomy within the philosophical traditions that animate artificial intelligence, with a focus on the history of early cybernetics and contemporary approaches to machine learning in behaviour-based robotics. We argue that while cybernetics and behaviour-based robotics challenge the premises of individual agency, cognition, communication and action that comprise the Enlightenment tradition, they also reiterate aspects of that tradition in the design of putatively intelligent, autonomous machines. This argument is made more concrete through a reading of the US Department of Defense's (DoD) *Unmanned Systems Integrated Roadmap: FY2013–2038*, particularly with respect to plans for future autonomous weapon systems (AWS). With that reading in mind, we turn to resources for refiguring agency and autonomy provided by recent scholarship in science and technology studies informed by feminist theory. This work suggests a shift in conceptions of agency and autonomy, from attributes inherent in entities to effects of discourses and material practices that either conjoin humans and machines, or that delineate differences between them. This shift leads in turn to a reconceptualization of autonomy and responsibility as always enacted within, rather than as being separable from, particular human-machine configurations. We close by considering the implications of these reconceptualizations for questions of responsibility in relation to automated/autonomous weapon systems. Taking as a model feminist projects of deconstructing categorical distinctions while also recognizing those distinctions' cultural-historical effects, we argue for simultaneous attention to the inseparability of human-machine agencies in contemporary war fighting and to the necessity of delineating human agency and responsibility within political, legal and ethical/moral regimes of accountability.

In proposing a reconceptualization of autonomy in the context of this chapter, we wish to be clear that our discussion is in no way meant to diminish the importance, or the possibility, of taking an operational approach to defining what have been categorized as lethal autonomous weapons. Mark Gubrud proposes that we begin with the definition offered by the US DoD, which states that an AWS is 'a weapon system that, once activated, can select and engage targets without further intervention by a human operator.[4] This includes human-supervised AWS that are designed to allow human operators to override the operation of the weapon system but that can select and engage targets without further human input after activation'.[5] Taking up the key phrase 'select and engage', Gubrud observes that 'selection' or targeting is complicated by the fact that 'the status of an object as the target of a weapon is an attribute of the weapon system or persons controlling and commanding it, not of the object itself … an object harmed without having been selected is called "collateral damage," be it a house, a garden, or a person'.[6] Target selection, Gubrud argues, is where the crucial questions and indeterminacies lie, and the operator, 'the final human in the so-called "kill chain" or "loop"'[7], should be the final decision point. Gubrud concludes that insofar as any weapon system involves the delegation of responsibility for target selection and engagement from operator to machine (whatever the precursors to that delegation in terms of intelligence reports, target lists and the like), that system is in violation of the principle of human control.

It is as a way of addressing these questions that those campaigning for a ban on lethal autonomous weapons have insisted on the need to preserve 'meaningful human control' over target selection and engagement.[8] The word 'meaningful' here is meant to anticipate and reject the proposition that any form of oversight over automated target identification constitutes 'human control'. Noel Sharkey offers a list of progressively greater levels of human control:

1. human engages with and selects target and initiates any attack
2. program suggests alternative targets and human chooses which to attack
3. program selects target and human must approve before attack
4. program selects target and human has restricted time to veto
5. program selects target and initiates attack without human involvement.[9]

On Sharkey's analysis, while Levels 1 and possibly 2 provide for what he identifies as 'the minimum necessary conditions for the notion of meaningful control',[10] the rest do not.[11]

[4] M. Gubrud, Autonomy without Mystery: Where Do You Draw the Line?, 9 May 2014, available at http://gubrud.net/?p=272.
[5] DoDDirective 3000.09: Autonomy in Weapon Systems, 21 November 2012, available at www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.
[6] Gubrud, Autonomy without Mystery.
[7] *Ibid*.
[8] 'Article 36: Key Areas for Debate on Autonomous Weapon Systems. Memorandum for Delegates at the Convention on Certain Conventional Weapons', Paper presented at the Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 13-16 May 2014, available at www.article36.org/wp-content/uploads/2014/05/A36-CCW-May-2014.pdf.
[9] See N. Sharkey, 'Staying in the loop: human supervisory control of weapons', ch. 2 in this volume.

In this chapter, we develop the argument, implicit in these discussions, that the adjudication of questions of autonomy and responsibility requires as its unit of analysis specific configurations of humans and machines. As we elaborate below, contemporary social theory has effectively challenged the premise that autonomy can be adequately understood as being an intrinsic capacity of an entity, whether human or machine, shifting the focus instead to the capacities for action that arise out of particular socio-technical systems. The concept of 'configuration' further orients us to relevant assumptions regarding humans, machines and the relations between them and to the practical consequences of particular human-machine assemblages.[12] Thus, to understand the agencies or capacities of either people or technologies requires an analysis of the dynamics of the socio-technical relations through which they are conjoined. Different configurations effect different distributions of agency between persons and technologies, making different capacities for action possible. In thinking about life-critical technical systems, it is the question of what conditions of possibility a particular configuration affords for human responsibility and accountability that is key.[13]

Autonomy: from Enlightenment reason to cybernetics

As background to this argument, we turn next to a brief review of shifting conceptualizations of autonomy as they have developed within the fields of cybernetics, artificial intelligence and robotics since the mid-twentieth century. Within the context of the modern episteme, one function of the concept of autonomy has been to posit an essential difference between humans and machines. Introduced by Enlightenment thinkers, autonomy was grounded in the idea of the individual self-determination[14] of the liberal subject. In Immanuel Kant's conception, the

[10] N. Sharkey, 'Towards a Principle for the Human Supervisory Control of Robot Weapons', *Politica and Società*, 2 (2014), 305-24.

[11] Sharkey in this volume cites the US DoD Science Board Task Force's review of many DoD-funded studies regarding 'levels of autonomy', which concluded that such designations are not particularly helpful in as much as 'they focus too much attention on the computer rather than on the collaboration between the computer and its operator/supervisor to achieve the desired capabilities and effects'. DoD, Directive 3000.09, 48. We return to the question of agency in human-machine configurations below.

[12] See L. Suchman, *Human-Machine Reconfigurations: Plans and Situated Actions* (New York: Cambridge University Press, 2007); L. Suchman, 'Configuration', in C. Lury and N. Wakeford (eds.), *Inventive Methods* (London: Routledge, 2012), 48.

[13] For further discussion regarding responsibility and liability for autonomous weapon systems (AWS), see G.S. Corn, 'Autonomous weapon systems: managing the inevitability of "taking the man out of the loop"', ch. 10 in this volume; N. Jain, 'Autonomous weapons systems: New frameworks for individual responsibility', ch. 13 in this volume; H.-Y. Liu, 'Refining responsibility: Differentiating two types of responsibility issues raised by autonomous weapons systems', ch. 14 in this volume.

[14] Self-determination or self-government are the English terms for the German concept 'Selbstbestimmung'. The term self-government already includes a cybernetic notion as governor, which is the translation of the Greek word cybernetes.

compliance of the human subject with moral law is the basis for human dignity.[15] And though autonomy of communities was a well-known concept in ancient Greece, autonomy now signified for the first time the idea of the right of self-determination of individual subjects. While nineteenth-century natural sciences debated the mechanistic, versus the vitalistic, nature of the living – life's deterministic or dynamic nature – the then dominant discourse of the humanities promoted the idea of the singularity of the human, and made it a widely accepted idea in liberal political discourse.

The reconfiguration of this concept of autonomy took its start in the 1920s and 1930s with the 'new sciences' of system theory and cybernetics. The biologist Ludwig von Bertalanffy, in his general systems theory,[16] conceptualized all living organisms as systems based on homeostatic balance. In this new logic, all organisms were regarded as being able to maintain steady states as well as their structure and identity in interaction with their environment and to regenerate and reproduce themselves.[17] This system of logic was ascribed not only to single organisms but also to collectives, whether they were biological, technical, economic or social.[18] This idea enables, in turn, the translation of organic and non-organic entities – of the material and non-material – into objects of communication and control.

The important transformation was to background defining features and intrinsic properties of organisms (including humans), which had been the main focus of concern prior to this, and to focus instead on goal-oriented behaviour. This combination of a powerful systems analogy and the concept of self-regulation, as well as the shift from essence to behaviour, increasingly blurred the boundary between humans and machines.[19] Systems theory along with cybernetics shifted the paradigm of science from energy towards information and from intrinsic properties of entities towards their behaviour. The

[15] '*Autonomie des Willens ist die Beschaffenheit des Willens, dadurch derselbe ihm selbst (unabhängig von aller Beschaffenheit der Gegenstände des Wollens) ein Gesetz ist. Das Prinzip der Autonomie ist also: nicht anders zu wählen als so, dass die Maximen seiner Wahl in demselben Wollen zugleich als allgemeines Gesetz mit begriffen seien*.' ('Autonomy of the will is the property of the will through which it is a law unto itself (independently of all properties of the objects of volition). Kant, *Grundlegung zur Methaphysik der Sitten*, (Frankfurt a.M.: suhrkamp, [1785] 1977), 74. The principle of autonomy is thus: 'Not to choose otherwise than so that the maxims of one's choice are at the same time comprehended with it in the same volition as universal law'. I. Kant, *Groundwork for the Metaphysics of Morals* (New Haven: Yale University Press, [1785] 2002), 58.

[16] L. von Bertalanffy, 'Der Organismus als Physikalisches System Betrachtet', *Die Naturwissenschaften*, 33 (1940), 521; see also H. Penzlin, 'Die Theoretische und Institutionelle Situation in der Biologie an der Wende vom 19. zum 20. Jh.' in I. Jahn, R. Löther and K. Senglaub (eds.), *Geschichte der Biologie: Theorien, Methoden, Institutionen, Kurzbiographien*, 3rd edn (Heidelberg, Berlin: Spektrum, 2000), 431.

[17] K. Gloy, *Das Verständnis der Natur.* volume I: *Die Geschichte des Wissenschaftlichen Denkens* (München: Beck, 1995).

[18] G. Leps, 'Ökologie und Ökosystemforschung' in I. Jahn, R. Löther and K. Senglaub (eds.), *Geschichte der Biologie: Theorien, Methoden, Institutionen, Kurzbiographien*, 3rd edn (Heidelberg, Berlin: Spektrum, 2000), 601.

[19] N. K. Hayles, 'Computing the Human' in J. Weber and C. Bath (eds.), *Turbulente Körper, Soziale Maschinen: Feministische Studien zur Wissenschaftskultur* (Opladen: Leske and Budrich, 2003), 99.

cyberneticians' interest in the behaviour of a system was driven by their involvement in military research, which occurred during the Second World War as Norbert Wiener worked on an anti-aircraft predictor. The calculation of aircraft trajectories was made possible only by neglecting the intrinsic features of the pilot and his machine and conceptualizing them as one entity – as a system – while concentrating on their behaviour.[20] Though Wiener did not succeed in building the predictor during the Second World War, cybernetics nonetheless successfully articulated the belief 'that machines and organisms were behaviourally and in information terms "the same"'.[21]

Cybernetics can be interpreted as an early technoscience, which aimed at constructing (anti-)systems with teleological behaviour. Cybernetics blackboxed not only machines but also any entities, including non-human and human organisms. In his book *The Human Use of Human Beings*,[22] Wiener claims that instead of materiality it is the organization or form of an entity that guarantees its identity in its ongoing transformation processes.[23] In principle, he sees no difference between the transport of matter or messages.[24] And it is not only specific materiality that is regarded as being irrelevant. Wiener as well as Claude Shannon introduced a new and purely formal concept of information, which sidestepped the context and meaning of information to ensure its computability. Both interpret information to be 'a principle of statistical quantification whose universal scope is equalled only by its indifference towards the specific nature of signals (physical, biological, technical or human)'.[25] While Shannon prioritized linearity in the famous sender- receiver model,[26] the concept of circular causation was central to Wiener's idea of communication.[27] Organisms and machines feed back certain information from specific parts of the system into the whole, which now has become an

[20] P. Galison, 'The Ontology of the Enemy: Norbert Wiener und the Cybernetic Vision', *Critical Inquiry*, 1 (1994), 228; P. Edwards, *The Closed World: Computers and the Politics of Discourse in Cold War America* (Cambridge, MA: MIT Press, 1996).

[21] G. Bowker, 'How to Be Universal: Some Cybernetic Strategies, 1943–70', *Social Studies of Science*, 23 (1993), 107.

[22] N. Wiener, *The Human Use of Human Beings: Cybernetics and Society* (Boston: Riverside Press, 1950); see also J. Weber, 'Blackboxing Organisms, Exploiting the Unpredictable: Control Paradigms in Human-Machine Translation' in M. Carrier and A. Nordmann (eds.), *Science in the Context of Application* (Springer, 2011), 409.

[23] In the history of philosophy – from Aristotle to contemporary approaches in philosophy of mind – we find a polarization of substance and form, matter and information. See T. Adorno, *Negative Dialektik* (Frankfurt a.M.: suhrkamp, [1966] 1982); J. Weber, *Umkämpfte Bedeutungen: Naturkonzepte im Zeitalter der Technoscience* (New York: Campus, 2003). These approaches take for granted that matter is passive and the form is imprinted on matter – it gets 'informed'. This approach has been extensively criticised by Marxists, phenomenologists, feminist philosophers, discourse theoreticians and post-colonial theory scholars.

[24] Wiener, *The Human Use of Human Beings*; see also Weber, *Umkämpfte Bedeutungen*.

[25] C. Lafontaine, 'The Cybernetix Matrix of French Theory', *Theory, Culture and Society*, 24 (2007), 27, 31.

[26] C. Shannon and W. Weaver, *The Mathematical Theory of Communication* (Urbana: University of Illinois Press, 1949).

[27] Lafontaine, 'The Cybernetix Matrix of French Theory'.

information network; this information is then supposed to help to regulate and thereby enhance the performance of the whole network.

The strong focus on information and feedback, and on the interaction of systems, is an identifying development of cybernetics,[28] which gave up on analysing intrinsic features of organisms, materiality or nature in favour of the frame of a functionalist logic. However, in the seminal paper 'Behavior, Purpose and Teleology', Norbert Wiener, Arturo Rosenblueth and Julian Bigelow claim that active purposeful behaviour is primarily based on negative feedback, 'signals from the goal [that] are necessary at some time to direct the behavior'.[29] Against the idea of functional relationships, Rosenblueth, Wiener and Bigelow claim a dependant, inter-objective relation between a system and its goal, which is not intentional but also non-random. They interpret purpose as 'the awareness of voluntary activity'.[30] In this way, cybernetics endows every single actor – whether human or machine – with a certain autonomy and 'elbow-room', by conceptualizing a systems' behaviour as at least partially teleological and adaptive:

> Perhaps the most fundamental contribution of cybernetics is its explanation of purposiveness, or goal-directed behaviour, an essential characteristic of mind and life, in terms of control and information. Negative feedback control loops which try to achieve and maintain goal states were seen as basic models for the autonomy characteristic of organisms: their behaviour, while purposeful, is not strictly determined by either environmental influences or internal dynamical processes. *They are in some sense 'independent actors' with a 'free will'*.[31]

On this basis, one could argue that the idea of autonomous systems begins with the cybernetician's claim of purposeful and goal-oriented behaviour as an attribute of any system. But what does this mean? And why do Francis Heylighen and Cliff Joslyn use quotation marks in their reference to the free will of the machine?

While the Enlightenment concept of autonomy is grounded in the idea of a free and self-aware subject, one which can self-determinedly and consciously choose its maxims,[32] the

[28] K. Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (University of Chicago Press, 1999).

[29] A. Rosenblueth, N. Wiener and J. Bigelow, 'Behavior, Purpose and Teleology', *Philosophy of Science*, 10 (1943), 18, 19.

[30] *Ibid*., 18.

[31] F. Heylighen and C. Joslyn, 'Cybernetics and Second-Order Cybernetics' in R. Meyers (ed.), *Encyclopedia of Physical Science and Technology*, 3rd edn (New York: Academic Press, 2001), 3 (emphasis added).

[32] Which need to be generalizable to be ethical according to Kant. See Kant, *Groundwork for the Metaphysics of Morals*.

cyberneticians explain purposeful behaviour not in rational-cognitivist terms but, rather, as a pragmatic physiological mechanism that can be automated: 'A torpedo with a target-seeking mechanism is an example. The term servo-mechanisms has been coined precisely to designate machines with intrinsic purposeful behavior'.[33] In this respect, cybernetics does not rely on assumptions of representation, symbol processing or pre-programmed plans to execute behaviour but, rather, on a pragmatic idea of a system's performance in interaction with its goal. The rhetoric of purpose and self-determination primarily rests on the fact that the system 'self-decides' again and again how to adjust its behaviour to achieve its goal. While the course of the machine towards the goal is not pre-programmed – in the case of the target-seeking torpedo, for example – the goal is pre-given. At the same time, the machine is at least partially flexible in seeking how to achieve its goal. In the logic of the cyberneticians, voluntary action (the philosopher's 'free will') and dynamic goal-oriented behaviour are more or less synonymous.

The dynamic relation between the servo-mechanism and the target – between the system and its goal – becomes possible through a tight coupling of system and environment. System and environment are regarded as separate, but closely interacting, entities. Wiener and his colleagues were interested in feedback – purposeful 'non-extrapolative, non-predictive' behaviour,[34] which could only be realized on the basis of the intimate interaction between different objects in a dynamic system-environment relation. In order to integrate the non-predictive into their calculations, they understood that the control of dynamic systems cannot be static or (too) centralized. Cybernetics is not so much about the exact calculation of behaviour but, rather, about its probabilistic estimate,[35] which is also the reason why the cyberneticians were interested in probability and game theory. And though concepts such as purpose, behaviour and teleology were stigmatized in late nineteenth- and early twentieth-century biology as vitalistic and non-scientific, cybernetics now managed to reformulate them as grounding concepts of a new, flexible technoscience of communication and control.[36]

The systems analogy, as well as the understanding of systems as goal-directed and purposeful, is a central pre-condition for the idea of the 'autonomy' of so-called smart and intelligent (war) machines. As developed by Ludwig von Bertalanffy, and further elaborated

[33] Rosenblueth, Wiener and Bigelow, 'Behavior, Purpose and Teleology', 19.

[34] *Ibid*.

[35] For the differences in the epistemological approaches of Wiener and von Neumann, see J. Lenhard, 'Computer Simulation: The Cooperation between Experimenting and Modeling', *Philosophy of Science*, 74 (2007), 176.

[36] M. Osietzki, 'Das "Unbestimmte" des Lebendigen als Ressource Wissenschaftlich-Technischer Innovationen: Menschen und Maschinen in den Epistemologischen Debatten der Jahrhundertwende' in J. Weber and C. Bath (eds.), *Turbulente Körper, soziale Maschinen: Feministische Studien zur Wissenschaftskultur* (Opladen: Leske & Budrich, 2003), 137.

by the cyberneticians, the systems analogy made it possible to shift the analysis of the information sciences from the intrinsic properties of entities towards their behaviour. The concept of behaviour was redefined as purposeful, moreover, insofar as any system's performance was directed through its interactions with its stipulated goal. The meaning of autonomy thereby shifted from the philosophical idea of the capacity of a self-aware and self-determined subject conforming to a (generalizable) moral law towards the technoscientific idea of autonomy as the operations of a pragmatic, physiological servo-mechanism.

Symbol-processing artificial intelligence

For manifold reasons, cybernetics did not dominate the field of artificial intelligence in the long run.[37] Already in the late 1960s, the symbol-processing approach of artificial intelligence, which was oriented towards mathematics and logic, won over the more biologically oriented approaches of cybernetics and early connectionism. Traditional, symbolic artificial intelligence is dominated by the paradigm of information processing in which intelligence, the brain, and the calculation of symbols are equated. Intelligence is seen less as a property than as a capability to think – which is understood as the processing of symbols and, correspondingly, as the computing of algorithms. This research paradigm also abstracts from the physical and concentrates on the representation of knowledge – that is, the adequate modelling of the world via symbols and logical inference as the decisive features of intelligence. Intelligence and the human brain are regarded as fundamentally computational in structure and function. Input is given, then it is processed, and finally output is generated. This procedure of input processing output was translated into the sense-think-act cycle of humans (and machines). The system receives input from the outside world via sensors (sense), interprets the sensory data via symbol processing and develops a plan (think). As output, the system executes an action according to the plan (act). Accordingly, symbolic artificial intelligence repeats traditional, rational-cognitive conceptions of human intelligence in terms of planning. It does not promote the idea of autonomy of technical systems in the sense of the randomly based, self-learning behaviour of so-called new artificial intelligence.[38] The symbolic approach worked very well in strongly rule-based environments such as chess playing or factory manufacturing but ran into severe problems when applied to mobile robots in dynamic, real-world environments.

[37] J.P. Dupuy, *The Mechanization of Mind* (Princeton University Press, 2000); Weber, 'Blackboxing Organisms'.

[38] We return to new artificial intelligence in the following section.

Definitions of what a robot comprises share the common requirement that a machine can engage in a sequence of ‘sense, think and act’ or perception, reasoning and action. The question of what counts as sensing or perception is key here, however. Does ‘sense, think and act’ refer to an assembly line robot that performs an action ‘at a certain location in a coordinate system representing real space’[39] or through machine ‘vision’ in a highly controlled environment where the consequences of failure are acceptable? Or does it invoke sensing and perception as dynamic, and contingent, capacities in open-ended fields of (inter)action with potentially lethal consequences? This leads as well to the question of what any instruction or plan presupposes about the capabilities required to carry it out. In the 1970s and 1980s, US researchers working in the field of artificial intelligence adopted the premise that plans, understood as a precondition for rational action, could be implemented as a device for structuring cognition and action in computational machines.

In *Plans and Situated Actions*,[40] the first author challenged this approach, proposing that rather than thinking of plans as cognitive control structures that precede and determine actions, they are better understood as cultural devices produced and used within specific sites of human activity. One entailment of this proposition is that planning is itself a form of situated activity that results in projections that bear consequential, but irremediably indeterminate, relation to the actions that they anticipate. Most importantly (and problematically) for the project of designing autonomous machines, plans and any other form of prescriptive specification presuppose competencies and *in situ* forms of interaction that they can never fully specify. The corollary of this is that the efficacy of plans relies upon the ability of those who ‘execute’ or ‘implement’ them to find the relation of the conditions and actions specified to some actual, particular occasion. And how to do that is not, and cannot be, fully specified. Prescriptive specifications such as plans, instructions and the like, in other words, presuppose an open horizon of capabilities that irremediably exceed their representational grasp.

Behaviour-based robotics

In the mid-1980s, a new, behaviour-based artificial intelligence and robotics developed that reinvented many insights of traditional cybernetics as it tried to avoid representations of the world and stressed the importance of (real-world) experience, negative feedback, situatedness,

[39] S.M. Riza, *Killing without Heart: Limits on Robotic Warfare in an Age of Persistent Conflict* (Dilles, VA: Potomac Books, 2013), 14.

[40] Suchman, *Human-Machine Reconfigurations*. Suchman, ‘Configuration’.

autonomy of the system and a tight coupling of system and environment.[41] Behaviour-based, or situated, robotics is inspired by first-order cybernetics but also by the theory of dynamic systems. The interest in materiality and embodiment that this approach promoted is now regarded by many as a necessary condition for real intelligence.[42] Roboticist Rodney Brooks adopted an idea of situated action as part of his campaign against representationalism in artificial intelligence and within a broader argument for an evolutionarily inspired model of intelligence.[43] For Brooks, 'situated' means that creatures reflect in their design an adaptation to particular environments. At the same time, the forms of adaptation to date are primarily focused on navigation, and the environment is delineated principally in terms of physical topographies. Brooks' situatedness is one that is largely emptied of sociality, and the creature's 'interactions' with the environment comprise variations of conditioned response, however tightly coupled the mechanisms or emergent the effects.

Nevertheless, cybernetics as well as behaviour-based artificial intelligence aims at overcoming the static and mechanistic paradigm of the 'traditional' sciences[44] in order to encompass dynamic and complex behaviours of organic and technical systems. In new artificial intelligence/behaviour-based robotics, the idea of autonomous systems gains momentum. Definitions of autonomy range – depending on the context and task of the system – from autonomy of energy supply or mobility to autonomy through adaptivity, embodied intelligence and learning behaviour (realized as computable, technical processes). While projects in autonomous energy supply or mobility aim to go beyond automation, they are mostly regarded as middle-range steps towards the achievement of full autonomy in the sense of the capability to operate 'in the real world without any form of external control'.[45] More ambitious approaches in robotics aim at adaptive learning behaviour intended to make the machine independent from human supervision and intervention.[46]

[41] R. Brooks, 'A Robust Layered Control System for a Mobile Robot', *IEEE Journal of Robotics and Automation* (1986) 14; L. Steels, 'Towards a Theory of Emergent Functionality' in *Proceedings of the First International Conference on Simulation of Adaptive Behavior* (Cambridge, MA: MIT Press, 1990), 451.

[42] K. Dautenhahn and T. Christaller, 'Remembering, Rehearsal and Empathy: Towards a Social and Embodied Cognitive Psychology for Artifacts', available at ftp://ftp.gmd.de/GMD/ai-research/Publications/1996/Dautenhahn.96.RRE.pdf; R. Pfeifer and C. Scheier, *Understanding Intelligence* (Cambridge, MA: MIT Press, 1999); S. Nolfi and D. Floreano, *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines. Intelligent Robots and Autonomous Agents* (Cambridge, MA: MIT Press, 2000).

[43] R. Brooks, *Cambrian Intelligence: The Early History of the New Artificial Intelligence* (Cambridge MA: MIT Press, 1999); R. Brooks, *Flesh and Machines: How Robots Will Change Us* (New York: Pantheon, 2002).

[44] A. Pickering, 'Cybernetics and the Mangle: Ashby, Beer and Pask', *Social Studies of Science*, 32 (2002), 413.

[45] G. Bekey, *Autonomous Robots: From Biological Inspiration to Implementation and Control* (Cambridge, MA: MIT Press, 2005).

[46] J. Beer, A. Fisk, and W. Rogers, 'Towards a Psychological Framework for Level of Robot Autonomy in Human-Robot Interaction' (Technical Report HFA-TR-1204. Atlanta, GA: Georgia Institute of Technology,

New robotics takes on the cybernetic idea of goal-oriented, 'purposeful' behaviour and tight system-environment coupling but reaches beyond it. Adaptive and biologically inspired robotics wants to include random behaviour as well.[47] There is a new interest in unpredictability and the unknown, as an integral factor of control and the systematization and exploitation of processes of trial and error.[48] While traditional artificial intelligence worked with pre-given rules for the robot's sensing and acting behaviours, behaviour-based robotics claims to build robots that can handle unpredictable situations in real-world environments. Therefore, biologically inspired concepts such as adaptation, imitation and experience-based learning[49] are the centre of attention.

Robots are posited to learn either through imitation (supervised learning) or through autonomous self-exploration. In the latter case, they should deduce the implicit general rules of a specific experience and adapt them in future situations. Learning is conceptualized as permanently acquiring new behaviours through autonomous self-exploration and through interaction with the environment via trial and error. The improved performance of the system is built on structural changes of the system (a kind of permanent self-reorganization). The basis of the autonomous learning process is unsupervised learning algorithms (as in value-based or reinforcement learning), which are supposed to enable agents to develop new categories and thereby adapt to the environment, though the new 'relevant configurations have to be selected using a value system'.[50] So while, on the one hand, the physical – and not the computational – structure of the agent and the 'tight coupling of embodiment, self-organization and learning'[51] are regarded as highly relevant to machine learning, on the other hand, the performance of the machine depends upon a pre-given value system in which the behavioural goals of the agent are inscribed. The value systems then 'evaluate consequences of behaviour'.[52]

Behaviour-based robotic agents seem to have more autonomy – understood as self-guided behaviour – in comparison to traditional agents of symbolic artificial intelligence and even of cybernetics. The agenda of continuous self-exploration on the basis of self-learning

School of Psychology, 2012); T. Fong et al., 'A Survey of Socially Interactive Robots', *Robotics and Autonomous Systems,* 42 (2003), 143-166.

[47] E.g., R. Pfeifer, 'On the Role of Embodiment in the Emergence of Cognition and Emotion', January 2000, available at www.ifi.unizh.ch/groups/ailab/publications/2000.html; Nolfi and Floreano, *Evolutionary Robotics*.

[48] Weber, 'Blackboxing Organisms'.

[49] A. Billard, S. Calinon and R. Dillmann, *Learning from Human Demonstration. Handbook of Robotics* (Cambridge, MA: MIT Press, 2013). Fong et al., 'A Survey of Socially Interactive Robots'; Sigaud and Peters, 'From Motor Learning to Interaction Learning in Robots', *Studies in Computational Intelligence*, 264 (2010), 1.

[50] Pfeifer and Scheier, *Understanding Intelligence*, 500.

[51] R. Pfeifer, M. Lungarella and F. Iida, 'Self-organization, embodiment, and biologically inspired robotics', *Science*, 318 (2007), 1088, 1090.

[52] Pfeifer and Scheier, *Understanding Intelligence*, 498–9.

algorithms makes so-called emergent[53] or unpredictable behaviour at least partially possible. The behaviour is not pre-programmed, but rather the outcome of a kind of systematized tinkering and situated experimenting of the system with its environment. However, this exploration is guided by pre-given value systems to make an 'assessment' of the experiences of the system possible. New experiences must be 'evaluated' through these pre-given values and categories. It seems that the robot behaviour shifts to another level: it is no longer totally pre-programmed but, instead, more flexible. The behaviour is mediated by random effects of the system's architecture[54] or learning algorithms, which can result in interesting, so-called emergent effects of the robot's behaviour, which are then exploited via post-processing.[55] The systems are regarded as being autonomous because of their sometimes unforeseen and even more rare (random) problem-solving behaviour. To address these machines as partners,[56] however, means to ignore the extent to which today's behaviour-based robots also rely on traditional symbolic artificial intelligence approaches, including huge amounts of pre-given systems structures (that is, the system architecture) and variables (such as the value system), as well as pre-programmed, determining software programs.

In their enthusiasm for the new, but nevertheless quite limited, capacities of behaviour-based agents, some roboticists' claims for real autonomous systems are greatly exaggerated. The claims are also grounded in a profound semantic shift in the meaning of autonomy, which is primarily defined as the capability to explore random real-world environments, by which sometimes unforeseen and useful behaviour might emerge. If the behaviour of these robots appears much more flexible than that of traditional robots, it is only because of the extremely static, non-dynamic behaviour of agents built in the tradition of symbolic artificial intelligence. The so-called autonomy of the behaviour-based agent is ontologically quite different from the original understanding of autonomy as self-determination, the ability to choose one's own (ethical) maxims of acting, or at least to comply with the dominant moral law.

As a result of the not exactly calculable behaviour of the agents of new artificial intelligence, and the invisibility of the underlying variables, categories and value systems in the robots' architecture, many people are intrigued by the more dynamic ontology of the bio-

[53] Emergence is understood in this context as the development of something qualitatively new on a higher and more complex level – a process that cannot be explained on a causal basis as a linear evolution or growth of complexity. See Hayles, *How We Became Posthuman*, p. 225.

[54] Brooks, 'A Robust Layered Control System'; for a critique, see, Weber, 'Blackboxing Organisms'.

[55] Specifically, reading the log files of the robot and trying to deduce how a new behaviour pattern of the robot was generated.

[56] Brooks, 'A Robust Layered Control System'; T. Christaller et al., *Robotik. Perspektiven für Menschliches Handeln in der Zukünftigen Gesellschaft* (Springer, 2001).

cybernetic sciences. In science communication, roboticists even enforce these impressions by dubious promises of agents that will soon be intelligent, develop human-like capabilities and (possibly if not inevitably) overtake humans in their moral and creative capabilities.[57] The ever-increasing competition between human and machinic autonomy seems to have reached its point of culmination in the contemporary discussion of the right of 'autonomous' weapons to decide the life and death of human beings. In the debate on AWS, it becomes even more obvious how autonomy is configured as self-sufficient, adaptive and self-determined performance, on the one hand, and pre-programmed, fully automated execution under perfect human control, on the other. These two imaginaries are profoundly intermingled, with questionable rhetorical and practical effects.

Figuring the future of machine autonomy in military robotics

To see how the traces of these histories are co-mingled in contemporary rhetorics of AWS, we turn to a reading of the US DoD's *Unmanned Systems Integrated Roadmap: FY2013–2038* (*USRM*). In a grammatical construction that posits a future knowable in the present (along with a characteristic elision of the difference between description and aspiration), the *USRM* informs us that '[t]he future of autonomous systems is characterized as a movement beyond autonomous mission *execution* to autonomous mission *performance*'.[58] 'Execution' and 'performance' are differentiated within the text by the former's reliance on a pre-programmed plan, while the latter involves the realization of goals that may change dynamically over a mission's course. Implicitly positing the existence of a future system capable of engaging in autonomous performance, pre-programming in this imaginary 'goes beyond system operation into laws and strategies that allow the system to self-decide how to operate itself'.[59] With that said, the document's authors are quick to restore the human to the loop. On the one hand, goals are directed by humans, while, on the other hand, 'automation is only as good as the software writer and developer because the control algorithms are created and tested by teams of humans':

[57] Think, e.g., of the prediction of roboticists that a robot soccer team will defeat the human world champion soccer team in 2050 or that there will be reliable software for moral decision making for lethal weapon systems in the foreseeable future. For the soccer example, see H. Kitano and M. Asada, 'The RoboCup Humanoid Challenge As the Millennium Challenge for Advanced Robotics', *Advanced Robotics*, 13 (2000), 723. For a moral decision-making software, see R. Arkin, P. Ulam and A. Wagner, 'Moral Decisionmaking in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception', *Proceedings of the IEEE*, 100 (2012), 3. For a critique of the latter claim, see P. Asaro, 'How Just Could a Robot War Be?' in P. Brey, A. Briggle and K. Waelbers (eds.), *Current Issues in Computing And Philosophy* (Amsterdam: IOS Press, 2000), 50.

[58] US DoD, *Unmanned Systems Integrated Roadmap*, 66 (emphasis in the original).

[59] *Ibid*., 66.

> In these algorithms, the 'patterns of life' are critical to automation and must be observed and captured properly to ensure accuracy and correctness of a decision-making process within the software. Ensuring accuracy and correctness requires a continual process in which the observe – orient – decide – act (OODA) loops in the software are continually updated via manual analysis, training, and operator understanding of algorithm inputs and outputs. The human brain can function in dynamic environments and adapt to changes as well as predict what will happen next. In simplistic terms, the algorithms must act as the human brain does.[60]

This passage is problematic, on our analysis, on several grounds. First, it presupposes that relevant circumstances can be rendered algorithmically, and still adequately, as 'patterns of life,' a form of profiling that has been effectively critiqued in assessments of the use of related techniques in campaigns of targeted killing.[61] Second, the reference to 'a decision-making process within the software' elides the difference between algorithmic and judgmental 'decision', again presuming the possibility of the latter's translation into the former. Finally, while insisting on the continued necessity of human oversight in the form of 'updating', the passage concludes by invoking a brain-based figure of human cognition and reasserting the possibility of its algorithmic replication.

Having set out the requirements for machine intelligence, the *USRM* goes on to provide a three-part account of the future of research and development in autonomous systems, beginning with:

> **4.6.1 Today's State (2013–2017)**
> In general, research and development in automation is advancing from a state of automatic systems requiring human control toward a state of autonomous systems able to make decisions and react without human interaction.[62]

While framing this section of the report as the current state of the art, the opening statement

[60] *Ibid*., 67.

[61] International Human Rights and Conflict Resolution Clinic (Palo Alto, CA: Stanford Law School) and Global Justic Clinic (New York: NYU School of Law) *Living Under Drones: Death, injury and trauma to civilians from US drone practices in Pakistan*, September, 2012, available at http://chrgj.org/wp-content/uploads/2012/10/Living-Under-Drones.pdf; C. C. Heyns, 'Targeting by Drones: Protecting the Right to Life', Paper presented at the European University Institute and Global Governance Programme on Targeted Killing, Unmanned Aerial Vehicles and EU Policy, European University Institute in Florence, 22 February 2013.

[62] US DoD, *Unmanned Systems Integrated Roadmap*, 68.

again conflates the descriptive with the promissory. The ‘in general’ implies not only a trend or tendency but also a kind of inevitability.[63] The document goes on to acknowledge that at present ‘systems that are autonomous require highly structured and predictable environments’[64] but with the implication that this is just a temporary phase, rather than a characterization of the results of the past fifty years or more of research and development in machine intelligence and robotics. The discussion of ‘today’s state’ and the ‘near term’ of the next four years includes a figure[65] that sets out ‘[t]he Army’s Vision for 5 Problem Domains’ in research and development in robotics, in which the domains are anthropomorphized as ‘Think-Look-Move-Talk-Work’. Figure 24 of this text warrants a closer reading.

[63] For a critique, see N. Sharkey, ‘The Evitability of Autonomous Robot Warfare’, *International Review of the Red Cross*, 94 (2012), 787–99; N. Sharkey and L. Suchman, ‘Wishful Mnemonics and Autonomous Killing Machines’, *AISBQ Quarterly, Newsletter of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour*, 136 (2013), 14. In his chapter in this volume, Dan Saxon posits that increasing speed and concomitant arguments of military necessity and advantage will further undermine the Directive’s already too vaguely specified standard of ‘appropriate levels of human judgment over the use of force’. D. Saxon, ‘A Human Touch: Autonomous Weapons, DOD Directive 3000.09 and the Interpretation of “Appropriate Levels of Human Judgment over the Use of Force”’, ch. 9 in this volume.

[64] US DoD, *Unmanned Systems Integrated Roadmap*, 68.

[65] *Ibid*., 70, Figure 24.

Table 4.1: Army's Vision for Five Problem Domains (Think – Look – Move – Talk – Work)

	Barriers to Achieving our Vision — >	Simplistic and Shallow World Model	Mobility-Focused Perception	Tele-operated or (at best) Scripted Planning	No Shared Understanding of Missions and Roles	Missing or Shallow Learning Capabilities
		World Model is either at only a metric level precluding reasoning, or at only a cognitive level without physical grounding	Objects in the world are perceived primarily only as mobility regions not as discrete objects of semantic and cognitive importance	Bots are almost always tele-operated or at best only perform sample scripted behaviors – and scripting all needed behaviors is not tractable	Bots are opaque and distributed, and cannot explain what they are doing – primarily because they don't know	Bots must be explicitly programmed to do tasks, so it is intractable to product the needed scope of behavior. Any learning capability is shallow and lacks generalization
"Think"	**Adaptive Tactical Reasoning**					
	Understand tasks, missions (METT-TC)	World model needs to represent concepts such as missions, tasks, and generally METT- TC [Mission, Enemy, Terrain, Troops - Time, Civlian Consideration].		Robots need to generate behaviors pertinent to achieving the mission, adapt to changing situation.	Robots need to be able to follow instructions given at a semantic or cognitive level, not just "goto (x,y)."	
	Follow semantic instructions					
	Generate behaviors to achieve mission, adapting to changing situation					
	Understand teammates and what they need to know					
"Look"	**Focused Situational Awareness**					
	Maintain SA relevant to current task/mission	World model needs to represent, maintain, monitor, and correct all info needed for SA.	Robot needs to contribute to the general SA of the unit, noting salient observations.		Robot needs to report on salient observations as needed to other elements of its unit.	Robot should learn by comparing its observations and actions to those of its human counterparts.
	Contribute to general SA of unit					
	Look for salient unforeseen events					
	Observe and report on salient activity					
"Move"	**Safe, Secure and Adaptive Movement**					
	Move cognitively in reaction to safest route in the world (as people do) with GPS or other metric crutches	World model needs to store and operate upon all entities needed to relate movement to tactical constraints.	Robot must perceive all entities in its environment relevant to safe, secure, and adaptive movement.	Robots must move in a tactically correct manner and react to changes in mission or circumstances.		Robot needs to learn from its movement experience whether from mobility challenges or tactical behavior.
	Move in tactically and continually relevant manner					
	Adjust to mobility challenges such as terrain, weather, barriers					
"Talk"	**Efficient Interactive Communication**					
	Receive and acknowledge semantic instructions	World model needs to have shared mental models as a basis for human–robot interaction.	Robot needs to send and information relevant based on a shared perception (common ground).		Robot needs to receive and acknowledge cognitive-level instructions and similarly explain its own behavior.	Robot needs to be able to learn through cognitive- level interaction with human teammates.
	Explain own behavior					
	Report information relevant to mission					
	Seek guidance as needed					
"Work"	**Interaction With Physical Wold**					
	Inspect and manipulate objects	World model needs to represent wide variety of objects to be manipulated.	Robot needs to perceive well enough to interact effectively with objects in a 3D world.	Robot needs to figure out how and when to manipulate or transport objects as needed.		Robot needs to learn from interaction with the physical world, e.g. when door is locked.
	Transport objects as needed					
	Open doors, windows, hoods, trunks, etc.					
	Use tools as needed					

Reproduced from the *Robotic Collaborative Technology Alliance (RCTA) FY2012 Annual Program Plan*, Figure 24 sets out the US Army's vision in the familiar form of a matrix, a representational device designed to ensure systematic and comprehensive consideration of orthogonal relations between two sets of categories, while at the same time asserting the systematicity of the analysis that it represents. The columns set out the five 'barriers to achieving our vision'. These name familiar problem areas that have vexed the project of artificial intelligence since its inception. For example, the premise that the army's vision requires its autonomous devices to have a world model adopts a conventional symbolic-processing approach to artificial intelligence, based in the encoding of a representation of the 'world' in which the device is to act, as a precondition for its effective and appropriate operation. However, while the characterization of existing models as 'simplistic and shallow' suggests that the challenge is to develop models that are more complex and deep, the wider premise that autonomous agency relies upon, and can be achieved through, the encoding of a model of the world as an *a priori* for action has, as we have discussed earlier, come under widespread critique, both within the field of artificial intelligence and among its philosophical and cultural critics.

While there is no question that human actors are continually engaged in rendering the world intelligible, it does not follow that this is done through a process of mapping between some cognitive model 'inside' the head of the individual and a world 'out there'. Rather, 'the world' is a very general gloss for an open horizon of potentially relevant circumstances. How a circumstance is articulated as such and made relevant, moreover, is not given in advance, but, rather, the recognition and/or articulation of something as a relevant circumstance is part of the ongoing, generative practices through which actions are rendered sensible and accountable. This helps to account, in turn, for the remaining problems, or rather characteristics, of the state of the art: machine 'perception' narrowly construed as obstacle avoidance; reliance on remote operation or pre-scripted behaviours and the irremediable incompleteness of the latter; inability to comprehend the situation of one's action and the lack of anything beyond the most technical sense of 'learning' from experience in ways that can inform future actions. The column headings of Figure 24 indicate, in sum, troubles in the conception of machine autonomy at work in the *USRM*, insofar as it presupposes the possibility of specifying relevant conditions of combat and appropriate responses within the range of capacities for sensing that are built into the system.

So what of the other axis of the matrix, the 'five problem domains' labelled 'Think-Look-Move-Talk-Work'? Treated as separable 'domains,' each of these prescribes a corresponding requirement for the ‘world model’, reiterating the premise that the model provides the basis for effective action. The remainder of the cells are filled with general characterizations of those capabilities that the robot ‘must’ or ‘needs to’ have – for example, ‘understand teammates and what they need to know’ or ‘contribute to general s[ituational] a[wareness] of the unit’ or ‘seek guidance as needed.’ However, these are precisely those abilities that the *x*-axis of the matrix has identified as being resistant to all of the efforts to achieve them to date. One way to read this figure, then, is as a demonstration of the limits of an approach to autonomy based on modelling and planning and of the decomposition of human action into multiple, separate domains. The persuasive intent of the matrix is not, however, to call the project of model-based robotics into question but, rather, to urge that efforts be redoubled.

While this figure makes clear the significant and substantial unsolved problems that face attempts to create autonomous, intelligent robots, there is no indication of the timeframe for their solution. Nonetheless, section 4.6.2 of the *USRM* assures us that ‘[t]he middle-term future state in the 2017–2022 time frame will consist largely of a further maturation of near-term capabilities … and move the capability further along the scale from automation to autonomous behavior,’ while section 4.6.3 on the ‘Long-Term Future State (beyond 2022)’ asserts again that ‘[t]he long-term state for unmanned systems will bring further maturation of the middle-term capabilities. It will also bring higher levels of automation’.[66] Both of these sections then go on to sketch out the imagined or desired next configurations of automated/autonomous systems for each of the armed services, but without having addressed the fundamental problems that continue to resist technological solution.

Refiguring autonomous agency

Our starting observation, set out in this chapter’s opening section, is that the project of machine intelligence is built upon, and reiterates, traditional notions of agency as an inherent attribute and autonomy as a property of individual actors. This conception of agency has been profoundly challenged, however, within contemporary science and technology studies. While focused on relations of subjects (scientists, technologists) and objects (natural kinds, artefacts) within the techno-sciences, these studies are a rich resource for a broader reconceptualization of autonomous agency.

In the field of science studies, Andrew Pickering develops the metaphor of the

[66] US DoD, *Unmanned Systems Integrated Roadmap*, 71.

‘mangle’ to argue that what he names ‘material agency’ is always temporally emergent in practice[67] rather than fixed in either subjects or objects. Karin Knorr-Cetina adopts a trope of ‘epistemic cultures’ to think about laboratories as mutually shaping arrangements of scientists, instruments, objects and practices aimed at the production of observably stabilized instantiations of ‘reality effects’.[68] The notion of ‘reconfiguration’ is central to her analysis, as the process through which subject/object relations are reworked.[69] Charles Goodwin’s analyses of what he names ‘professional vision’[70] demonstrate in detail how the acquisition of professional competency comprises processes through which practitioners learn to ‘see’ the objects of their profession, at the same time that those objects are reflexively constituted through the same practices by which they become intelligible.[71] Taken together, these analyses support an understanding of agencies as always relational and give us, in turn, a different way of conceptualizing the problem of attributions of knowledge and agency to machines. The problem is less that we attribute agency to computational artefacts than that our language for talking about agency, whether for persons or artefacts, presupposes a field of discrete, self-standing entities. Latour takes us closer to the domain of the weapon system, with his reflections on the gun:

> You are different with a gun in your hand; the gun is different with you holding it. You are another subject because you hold the gun; the gun is another object because it has entered into a relationship with you. The gun is no longer the gun-in-the-armory or the gun-in-the-drawer or the gun-in-the-pocket, but the gun-in-your-hand … If we study the gun and the citizen [together] … we realize that neither subject nor object … is fixed. When the [two] are articulated … they become ‘someone/something’ else.[72]

These inquiries re-specify agency from a capacity intrinsic to singular actors (human or artefactual) to an effect of subject/object relations that are distributed and always contingently enacted. In the words of feminist theorist Karen Barad, ‘agency is not an attribute but the

[67] Pickering, ‘Cybernetics and the Mangle’.

[68] K. Knorr-Cetina, *Epistemic Cultures: How the Sciences Make Knowledge* (Cambridge, MA: Harvard University Press, 1999), 26–33.

[69] Suchman, *Human-Machine Reconfigurations*.

[70] C. Goodwin, ‘Professional Vision’, *American Anthropologist*, 96 (1994), 606; C. Goodwin, ‘Seeing in Depth’, *Social Studies of Science*, 25 (1995), 237.

[71] R. Prentice, ‘The Anatomy of a Surgical Simulation: The Mutual Articulation of Bodies in and through the Machine’, *Social Studies of Science*, 35 (2005), 837; M. Myers, ‘Molecular Embodiments and the Body-Work of Modeling in Protein Crystallography’, *Social Studies of Science*, 38 (2008), 163.

[72] B. Latour, *Pandora’s Hope: Essays on the Reality of Science Studies* (Cambridge, MA: Harvard University Press, 1999), 179–80.

ongoing reconfigurings of the world'.[73] Methodologically, this view of the nature of socio-material agencies has two broad implications. First, it demands attention to the question of frames, of the boundary work through which a given entity is delineated as such. Beginning with the premise that discrete units of analysis are not given but made, we need to ask how any object of analysis – human or machine or a combination of the two – is called out as being separate from the more extended networks of which it is part. This work of cutting the network is a foundational move in the creation of socio-technical assemblages as objects of analysis or intervention.[74] In the case of the robot, or autonomous machine more generally (as in the case of the individual human as well), this work takes the form of modes of representation that systematically foreground certain sites, bodies and agencies while placing others offstage. Our task as analysts is then to expand the frame to a wider field of view that acknowledges the effects created through a particular framing, while also explicating the hidden labours and unruly contingencies that inevitably exceed its bounds.

Implications for the debate over AWS

Applied to the case of weapon systems, these methodological shifts have profound political and moral consequences. With respect to automation and autonomy, an understanding of agency not as an attribute of either humans or machines but, rather, as an effect of particular human-machine configurations opens the possibility of explicating the systematic erasures of connection and contingency through which discourses of autonomous agency operate. And it opens as well the question of how to configure socio-technical assemblages in such a way that humans can interact responsibly in and through them.[75] At the same time, we face a certain tension in thinking about responsibility and the human in these terms. Cybernetics and new artificial intelligence abandon the idea of intrinsic properties of humans or non-humans and stress the interaction of systems and their environments. And as we have discussed, contemporary science and technology studies – particularly those that have been informed by feminist theory – effectively dissolve the problematic idea of human autonomy in favour of attention to the human/non-human relations through which what we call human agency is produced as one effect. However, in engaging discourses of autonomous weaponry, it seems crucial to articulate the particular agencies and responsibilities of the human war fighter and

[73] Barad, *Meeting the Universe Halfway*, 141.

[74] M. Strathern, 'Cutting the Network', *Journal of the Royal Anthropological Institute*, 2 (1996), 517.

[75] This task is made more difficult by the lack of transparency that characterizes initiatives in AWS development. See S. Knuckey, 'Autonomous weapons systems and transparency: Towards an international dialogue', ch. 8 in this volume.

their resistance to translation into executable code. This is not so much a contradiction to be resolved, we would argue, but a trouble with which we need to stay. As well as recognizing the epistemic situatedness of our concept of autonomy, we need to explore the ways in which our agencies are entangled with, and dependent upon, the technological world today and to analyse our particular agencies within the assemblages that we configure and that configure us.[76]

In the opening pages of *Killing without Heart*,[77] Air Force Colonel M. Shane Riza reflects on the shifting agencies and responsibilities of the 'human in/on the loop' and the 'string of events [that] we technological warriors facetiously call the "consecutive miracles" that comprise the effective functioning of technologically advanced weapon systems'.[78] He points to the ways in which pilots and engineers in the field are called upon to mitigate shortcomings of weapons development contracts in their failure to fully address the contingencies of use. At the same time, he emphasizes that the issue for him is not the dependency of the fighter on the technology: 'I am comfortable in the knowledge that my mastery, such as it was, of the technology at my fingertips successfully took me into battle and brought me back'.[79] Rather, what he is concerned about is the decision (not) to kill, which he defines as the red line between automation and robotic autonomy. Riza makes a strong distinction between automated weaponry and 'autonomous killers'[80] and proposes that meaningful discussion of developments in weapon systems requires that we 'come to grips with the clear distinction between automation and autonomy and navigate the all-too-unclear realm of the latter's spectrum'.[81] He includes landmines and machine guns among automated weaponry, while autonomy is exemplified by 'a small tracked robot carrying a shotgun or assault rifle with the ability to select and fire on targets of its own choosing'.[82] This categorization is challenged, however, by the analogy between land mines and 'killer robots' made by campaigners such as Article 36's Matthew Bolton,[83] who observes that the campaign to ban land mines was based precisely on their 'autonomy', albeit a self-firing triggered not by 'decision' but, rather, by the simple trip wire of a proximate body. And Riza agrees with this further on, when he notes that:

[76] See P. Kalmanovitz, 'Judgment, liability, and the risk of riskless warfare', ch. 7 in this volume.
[77] Riza, *Killing without Heart*.
[78] *Ibid*., 4.
[79] *Ibid*., 6.
[80] *Ibid*., 12.
[81] *Ibid*., 13.
[82] *Ibid*., 12.
[83] Article 36: Ban Autonomous Armed Robots, 5 March 2012, available at www.article36.org/statements/ban-autonomous-armed-robots/.

> [a]ntipersonnel mines of the kind that kill hundreds of innocent people every year are indiscriminate by their very nature, and indiscriminate killing had been against the law of war in written form for a hundred years before the [Ottawa Accord of 1997] – and against the norms of behavior for a millenium before that. We should have known better than to field them.[84]

It is the question of discrimination, specifically between combatants and non-combatants, that becomes crucial, and recourse to the human, whether 'in' or 'on the loop', is complicated by the nexus of intensifying speed and increasing automation that characterizes modern weapon systems. Citing the 'friendly fire' incidents of the 2003 invasion of Iraq, when Patriot missiles shot down two allied aircraft killing their crews, Riza concludes that '[t]he decision to fire in these instances were made by humans, but their decisions were radically influenced – perhaps to the point of abdication – by basic artificial intelligence'.[85] This incident troubles the clarity of the line between automation and autonomy, along with the questions of agency and responsibility that the human in/on the loop is imagined to resolve.[86]

The interrelated dangers of increasing automation in weapon systems and the shift towards weapon autonomy pose two critical challenges. On the one hand, we need to understand the ways in which automation establishes its own circular logics of necessity, as the shortened time frames that result become, in turn, the justification for further automation. Following Riza, we can understand that the 'loop' in which humans and machines are conjoined in contemporary weapon systems, whether the humans are figured as 'in' or 'on' that loop, diminishes the possibility of judgments not to kill. In this logic of no time for communication or consideration, machine autonomy becomes the necessary extension to automation. At the same time that we identify the connecting logics of automation and autonomy, however, we need as well to articulate their differences. More specifically, if our concern is to interrupt the vicious cycle of automation in war fighting, and the political and economic investment in a future of autonomous weapons that it justifies, one strategy is to make the discontinuity between automation and autonomy more evident.[87] To do that, we need a critical examination of the assumptions that underwrite conceptions of autonomy, whether human or machine, in

[84] Riza, *Killing without Heart*, 29.
[85] *Ibid*., 20.
[86] See also the chapter by Christof Heyns in this volume. C. Heyns, 'Autonomous weapons systems: Living a dignified life and dying a dignified death', ch. 1 in this volume.
[87] This is the basis for the 'Campaign to Stop Killer Robots', a coalition of non-governmental organizations dedicated to the development of an arms control ban on lethal autonomous weapons. See www.stopkillerrobots.org.

the fields of artificial intelligence and robotics. We also need to develop a concept of autonomy that comprises fully the socio-political dimensions of human–machine interaction. Applied to weapon systems, this means that the question is less about automation versus autonomy than it is about what new forms of agency are enabled by contemporary configurations of war fighting and with what political, ethical, moral and legal consequences.