Self-produced penultimate version of Chapter 6 of the volume *Autonomous Weapons Systems: Law, Ethics, Policy*, edited by Nehal Bhuta, Susanne Beck, Robin Geiβ, Hin-Yan Liu, Claus Kreβ. Cambridge University Press 2016. ISBN 978-1-107-15356-1, pp. 122-142.

6

# On Banning Autonomous Weapons Systems: From Deontological to Wide Consequentialist Reasons

Guglielmo Tamburrini

## Introduction

This chapter examines the ethical reasons supporting a moratorium and, more stringently, a pre-emptive ban on autonomous weapons systems (AWS). Discussions of AWS presuppose a relatively clear idea of what it is that makes those systems autonomous. In this technological context, the relevant type of autonomy is task autonomy, as opposed to the personal autonomy, which usually pervades ethical discourse. Accordingly, a weapons system is regarded here as autonomous if it is capable of carrying out the task of selecting and engaging military targets without any human intervention.

Since robotic and artificial intelligence technologies are crucially needed to achieve the required task autonomy in most battlefield scenarios, AWS are identified here with some sort of robotic systems. Thus, ethical issues about AWS are strictly related to technical and epistemological assessments of robotic technologies and systems, at least insofar as the operation of AWS must comply with discrimination and proportionality requirements of international humanitarian law (IHL). A variety of environmental and internal control factors are advanced here as major impediments that prevent both present and foreseeable robotic technologies from meeting IHL discrimination and proportionality demands. These impediments provide overwhelming support for an AWS moratorium – that is, for a suspension of AWS development, production and deployment

The author is most grateful to Jürgen Altmann, Noel Sharkey, Leen Spruit, Giuseppe Trautteur and the editors of this volume for their stimulating and helpful comments on a draft of this chapter. The research leading to these results has been partially funded by the Robotics Coordination Action for Europe program, which has received funding from the European Community Seventh Framework Programme (FP7/2007-2013) under Grant Agreement ICT-611247. The author is solely responsible for its content. It does not represent the opinion of the European Community, and the Community is not responsible for any use that might be made of the information contained therein.

at least until the technology turns out to be sufficiently mature with respect to IHL. Discrimination and proportionality requirements, which are usually motivated on deontological grounds by appealing to the fundamental rights of the potential victims,[1] also entail certain moral duties on the part of the battlefield actors. Hence, a moratorium on AWS is additionally supported by a reflection on the proper exercise of these duties – military commanders ought to refuse AWS deployment until the risk of violating IHL is sufficiently low.

Public statements about AWS have often failed to take into account the technical and epistemological assessments of state-of-art robotics, which provide support for an AWS moratorium. Notably, some experts of military affairs have failed to convey in their public statements the crucial distinction between the expected short-term outcomes of research programs on AWS and their more ambitious and distant goals. Ordinary citizens, therefore, are likely to misidentify these public statements as well-founded expert opinions and to develop, as a result, unwarranted beliefs about the technological advancements and unrealistic expectations about IHL-compliant AWS. Thus, in addition to the forms of consequential ignorance induced by the usual secrecy and reticence surrounding military technologies, the inadvertent or intentional failure to distinguish clearly between the long-term visionary goals of AWS research and its short-term outcomes hampers public debate about a moratorium and the related democratic deliberations on AWS.

Technical and epistemological assessments of AWS compliance with IHL play a central role in arguments for a moratorium, but they generally recede in the background of arguments for a pre-emptive ban on AWS. Notably, Peter Asaro advanced an argument for banning AWS,[2] which is independent of current and foreseeable failures to comply with IHL discrimination and proportionality requirements. Asaro's argument is distinctively based on the defence of human rights and dignity from a deontological standpoint.

Additional reasons for a pre-emptive ban on AWS are defended here from a consequentialist – rather than a deontological – standpoint in normative ethics. These reasons deserve special attention as they effectively countervail reasons for the future deployment of AWS that are equally advanced on consequentialist grounds in ethics. Consequentialist reasons for AWS future deployment are typically based on the expectation of IHL-compliant AWS that will bring down the numbers of dead combatants, innocent casualties and collateral damage, as a result of its targeting and engagement capabilities that surpass those of emotionally frail and cognitively more limited

[1] C. Heyns, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Execution, UN Doc. A/HRC/23/47, 9 April 2013.
[2] P. Asaro, 'On banning autonomous weapon systems: human rights, automation, and the de-humanization of lethal decision-making', *International Review of the Red Cross*, 94 (2012), 687.

human soldiers.[3] However, one should carefully note that these reasons flow from a fairly narrow appraisal of the expected consequences of AWS deployment. From a wider consequentialist perspective, the risk of a new arms race and global destabilization – up to and including nuclear destabilization – prevails over the allegedly good consequences of AWS deployment.[4] Accordingly, the consequentialist arguments in normative ethics are found to provide strong support for a pre-emptive ban on AWS, and they converge with deontological arguments that are based on the defence of human dignity and rights.

**On the definition of AWS**

Taking for granted the customary description of a weapons system as a set of devices and tools that are used for offensive or defensive fighting, the distinctive problem concerning AWS is circumscribing the class of weapons that deserve to be called autonomous. Philosophical definitions of what one may aptly call *personal* autonomy (or *p*-autonomy) are hardly useful in this circumstance, insofar as only conscious individuals, who are additionally assumed to be free and capable of acting on their genuine intentions, are *p*-autonomous. Therefore, no machine that one can make an educated guess about – that is, on the basis of current scientific and technological knowledge – satisfies the requirements for *p*-autonomy.[5]

More pertinent and informative for the purpose of defining and identifying AWS is the idea of *task* autonomy (*t*-autonomy), which is construed here as a three-place relationship between a system *S*, a task *t*, and another system *S'*. Roughly speaking, a system *S* is said to be autonomous at some task *t* from another system *S'* (*S* is *t*-autonomous from *S'*) if *S* accomplishes *t* regularly without any external assistance or intervention by *S'*. At the age of 10 months, babies usually fail to be autonomous at walking, insofar as they need parental support to do so, whereas most toddlers past the age of 15 months are autonomous from human beings or from any other supporting system at performing this task. Robots that one meets on factory floors are autonomous from human workers at a variety of assembling, painting and payload transportation tasks. Clearly, a system that is autonomous at a task *t* may fail to be autonomous at many other tasks: car factory robots and toddlers are usually unable to prepare good strawberry ice cream without significant external support. Moreover, a system that is autonomous at *t* from a system *S'* may fail to be autonomous at

[3] R.C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (CRC Press, 2009); R.C. Arkin, 'Lethal autonomous systems and the plight of the non-combatant', *AISB Quarterly*, 137 (2013), 1.
[4] J. Altmann, 'Arms control for armed uninhabited vehicles: an ethical issue', *Ethics and Information Technology*, 15 (2013), 137.
[5] The notion of *p*-autonomy, which plays pivotal roles in moral philosophy and law, will turn out to be useful in the ensuing discussion of moral responsibilities of human beings who are in charge of activating an autonomous weapon system (AWS).

the same task *t* from another system *S''*. For example, a driverless car is autonomous from humans at the task of driving but may depend on a GPS system to carry it out correctly.[6]

The US Department of Defense (DoD) proposed a definition of AWS that relies on *t*-autonomy, insofar as it involves a weapons system *S*, the complex task *t* of selecting and engaging military targets and human beings as the systems *S'* from which *S* must be autonomous. According to this definition, any weapons systems is autonomous 'that, once activated, can select and engage targets without further intervention by a human operator.'[7] The complex task of selecting and engaging targets can be divided into a variety of subtasks: sensory data must be acquired and processed in order to identify, track, select and prioritize targets before one can decide, on the basis of a given set of engagement rules, whether to apply force against them. Therefore, the design and implementation of all but the most rudimentary types of AWS must rely on artificial intelligence and robotic technologies for artificial perception and situational awareness, action planning and reactive behaviour. Thus, most AWS satisfying the DoD's definition are sensibly regarded as some sort of robotic system.

Assuming the DoD's definition in the ensuing discussion, let us now turn to consider some specific robotic systems that are autonomous according to this definition. A relatively simple case in point is the Samsung system SGR-A1 – a robotic stationary platform designed to replace or to assist South Korean sentinels in the surveillance of the demilitarized zone between North and South Korea.[8] The SGR-A1 can be operated in either unsupervised or supervised modes. In the unsupervised mode, the SGR-A1 identifies and tracks intruders in the demilitarized zone, eventually firing at them without any further intervention by human operators. In the supervised mode, firing actions are contingent on the judgment and the 'go' command of military officers. Thus, the SGR-A1 counts as an AWS according to the DoD's definition if it operates in the unsupervised mode, and it does not count as such, if it operates otherwise. In the latter case, it is better viewed as a combination of a decision-support system with a remote-controlled firing device.

The supervised SGR-A1 preserves almost every *t*-autonomy required of an AWS, insofar as it performs regularly, and without any human intervention, the perceptual and cognitive tasks of target identification and tracking in its intended operational environment. Accordingly, this robotic sentinel affords a vivid and straightforward illustration of the fact that a simple on/off operational mode switch can make the difference between an AWS and a non-autonomous weapon system. The risk of the SGR-A1 performing poorly at targeting and engagement tasks is reduced in either one of

[6] The notion of *t*-autonomy discussed here is closely related to the idea of independence for technological devices examined in Andrea Omicini and Giovanni Sartor's contribution to this volume.
[7] US Department of Defense, Directive 3000.09: Autonomy in Weapons Systems, 21 November 2012, 13–14 available at www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.
[8] A. Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons* (Ashgate, 2009).

its operational modes by a crucial environmental factor. The Korean demilitarized zone is severely constrained. Human access to areas monitored by the SGR-A1 is categorically prohibited; any human being detected there is classified as a target and perceptual models enabling one to discriminate between human targets and non-targets are available. The Korean robotic sentinel is unable to deal proficiently with more challenging perceptual discrimination and decision-making problems, such as those arising in more cluttered and highly dynamic warfare scenarios, where AWS are required to distinguish belligerents from non-belligerents and friends from foes. Accordingly, attributions of *t*-autonomies, which enable a weapon system to qualify as autonomous are inherently context dependent, insofar as suitable boundary and initial conditions must be in place for a system *S* to perform *t* correctly without any external intervention by human beings. This observation suggests that *t*-autonomy should be more accurately construed as a relationship between four elements: a system *S*, a task *t*, a system *S'* from which *S* does not depend to accomplish *t* and an environment where *t* must be performed.

Since an AWS can be correctly described as being autonomous in some environments and as non-autonomous in other environments, the problem arises as to whether and how one gets to know that a particular operating environment is included in the class of environments in which a weapons system counts as an AWS. This problem must be duly taken into account by AWS producers, insofar as they have to state the boundary conditions for the intended use of their products, and by military commanders too, insofar as they have to evaluate whether the operational scenario they are dealing with belongs to the class of environments in which the AWS can correctly perform the target selection and engagement task for which it was designed. Major scientific and technological hurdles have to be solved in order to put the prospective AWS producers and users in the right position to adequately address this problem. To illustrate, let us examine an analogous epistemic problem arising in the context of *t*-autonomies that is of interest to industrial and service robotics.

**How does one know that AWS will behave as it is intended?**

Environmental conditions contribute to how robotic behaviours are shaped in ways that one can hardly overrate. The tortuous paths that insect-like robots trace on a beach result from the application of a fairly uniform gait on uneven and unsteady walking surfaces. Variable illumination conditions may hinder the visual recognition of obstacles on the trajectory of both outdoor and indoor robots. The water just spilled on the kitchen floor, the Persian carpet recently placed in the living room and many other causal factors changing frictional coefficients may perturb the trajectory of mobile robots negotiating the floors of our homes. Accordingly, good models of robot

interactions with the environment must be available to predict, identify and reduce external sources of perturbations that the robot's control system cannot adequately deal with.

Adapting environments to the perceptual, cognitive and action capabilities of robotic systems is a heuristic strategy enabling one to rule out a wide variety of causal factors that jeopardize robotic compliance with task assignment. This strategy is extensively pursued in industrial robotics. For example, one may limit the dynamic behaviour and the sort of items allowed in the industrial robot workspace so that only item that the robot can properly recognize, manipulate or avoid contact with is permitted there. In particular, since human workers are a major source of change, which is both dynamic and difficult to predict, one must strictly regiment human-robotic interactions or fully segregate robots from the workspaces that are assigned to factory workers.

Human-robot segregation policies are no longer available when one moves away from assembly lines and other orderly industrial task environments towards the current frontiers of service and social robotics, where diverse human-robot interactions are often an integral part of the task requirements. To illustrate, consider the prospective use of autonomous mobile robots as assistants or caregivers to people in their homes, especially to elderly or disabled people. A carrier robot must be able to safely take a human being from, say, a bed to an armchair and back again, and a servant robot must be able to grasp cups and glasses and use them properly to serve beverages. In order to be granted permission to sell such robots, prospective manufacturers must supply proper evidence that the autonomous personal care and assistance robots they intend to commercialize are able to perform safely the required *t*-autonomies in normal operating conditions.

Accordingly, the International Organization for Standardization (ISO) 13482 (ISO 2014) demands that each autonomous care robot must be tested carefully for a 'sufficiently low' level of risk to users, and it points to the option of 'constraining the operational scenarios' as a suitable strategy for achieving this goal.[9] However, one cannot pursue this strategy to the point of transforming human dwellings into robotized factory floors or into the likes of the selfish giant's garden in Oscar Wilde's tale, where children are not admitted to play lest they might interfere with the proper working of some personal care and assistant robot. In the end, one can only hope to lower drastically the interaction risk by improving the robot control system, by limiting human-robot interaction to a minimum and by warning users about the chief environmental conditions that are likely to disrupt robotic behaviours.

Warfare scenarios resemble neither the orderly factory floors inhabited by industrial robots nor the relatively uneventful homes in which one usually lives. Each fighting side strives to

[9] Organization for International Standardization, *International Standard 13482: Robots and Robotic Devices: Safety Requirements for Personal Care Robots* (Organization for International Standardization, 2014), 33.

generate unexpected events that defy the opponent's predictions. And the interaction of partially known causal factors in warfare scenarios produces events that are unpredictable on the basis of past experience, knowledge of the battlefield situation and available models of warfare operations. It is in these unstructured and surprise-seeking conditions that AWS must operate. Here, one can neither resort to the ISO's recommendation of 'constraining operational scenarios', for one does not know and control all of the involved forces, nor confidently believe that the more important 'abnormal' situations perturbing the desired AWS behaviours have been properly taken care of.

**From AWS epistemology to moral reasons for a moratorium**

The epistemic predicament concerning unstructured warfare scenarios is enhanced by an appraisal of the state-of-art technologies for robotic perception. Consider the problem of recognizing *hors de combat* people—one has to be able to tell bystanders apart from foes and hostile opponents apart from surrendering, unconscious or otherwise inoffensive opponents. Identifying behaviours that conventionally or unconventionally carry surrender messages involve the viewpoint-independent classification of bodily postures and gestures in variable illumination conditions, in addition to an understanding of emotional expressions and real-time reasoning about deceptive intentions and actions in unstructured warfare scenarios. Thus, engineers who wish to endow an AWS with IHL-compliant competences face a distinctive challenge from the fact that adequate sets of rules for perceptual classification are difficult to isolate and state precisely.

Instead of attempting to furnish robots with an exhaustive set of rules for perceptual classification, one may opt instead to give them the capability of learning these rules from experience. The learning robot must identify classification rules on the basis of some finite set of training data, usually containing instances of correct and incorrect classifications. Once this learning phase is concluded, the reliability of the rule that has been learned can be assessed by theoretical or empirical methods.[10] In either case, however, the results that are obtained are contingent on a variety of background assumptions.[11] Thus, for example, the outcomes of empirical testing on learned rule reliability depend on the assumption that the training and testing data are significant representatives of the perceptual classification problems that the robot must solve. Similarly, probabilistic bounds on error frequency that one establishes within the more abstract mathematical framework of statistical learning theory are contingent on the assumption that the training data were

[10] T.M. Mitchell, *Machine Learning* (McGraw Hill, 1997).
[11] M. Santoro, D. Marino and G. Tamburrini, 'Robots interacting with humans: from epistemic risk to responsibility', *AI and Society*, 22 (2008), 301.

independently drawn from some fixed probability distribution.[12] Assumptions of both kinds are crucial in addressing the ethical and legal problems that concern learning robots,[13] but they are difficult to buttress in the case of human-robot interactions envisaged in service and social robotics and, *a fortiori*, in surprise-seeking, erratic and unstructured warfare scenarios.

Scientists in the field of robotics have frequently emphasized the formidable scientific and technological challenges that have to be met before one can realistically envisage IHL-compliant AWS. Thus, Noel Sharkey has stated that '[c]urrently and for the foreseeable future no autonomous robots or artificial intelligence systems have the necessary properties to enable discrimination between combatants and civilians or to make proportionality decisions'.[14] And Ronald Arkin has advanced an argument for a moratorium that hinges on the current and foreseeable limitations of robotics systems: '[T]he use and deployment of ethical autonomous robotic systems is not a short-term goal … There are profound technological challenges to be resolved, such as e ective in situ target discrimination and recognition of the status of those otherwise *hors de combat*.' For this reason, he claims, 'I support the call for a moratorium to ensure that such technology meets international standards before being considered for deployment'.[15]

Arkin's argument espouses a broad consequentialist framework for ethical theorizing about an AWS moratorium insofar as the moral permission to deploy them is contingent on discrimination and proportionality principles, in addition to the casualty reduction benefits that one may obtain from AWS that behave according to very conservative firing decisions that human soldiers cannot afford to apply in view of their legitimate self-preservation concerns. Within this consequentialist ethical framework, the epistemological reflections on state-of-art and foreseeable developments in robotics and artificial intelligence overwhelmingly support the moral obligation of suspending AWS development, production and deployment at least until compliance with IHL principles and other envisaged benefits has been convincingly demonstrated. One can adduce additional moral reasons for an AWS moratorium from a deontological standpoint in ethical theorizing.

---

[12] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edn (Springer, 2000).
[13] D. Marino and G. Tamburrini, 'Learning robots and human responsibility', *International Review of Information Ethics*, 6 (2006), 46.
[14] N. Sharkey, 'Saying 'no!' to lethal autonomous targeting', *Journal of Military Ethics*, 9 (2010), 369, 378.
[15] As a suitable benchmark to verify international standards, Arkin proposes the capability of a robot to behave as well as, or better than, our soldiers with respect to adherence to the existing international humanitarian law (IHL). Let us note in passing that the successful implementation of an ethical inference engine conforming, e.g., to the ethical governor architecture outlined in Arkin's scholarship, would be largely insufficient to enable an AWS to comply with IHL. Indeed, the moral arguments licensing any such conclusion must include among their premises perceptually corroborated statements concerning, e.g., the presence or absence of hostile combatants and non-belligerents, thereby presupposing an adequate solution to the perceptual classification problems mentioned above. R.C. Arkin, Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture, Technical Report GIT-GVU-07-11 (2007); R.C. Arkin, *Governing Lethal Behavior in Autonomous Robots.*

As Pablo Kalmanovitz emphasizes in his contribution to this volume, the deployment of an AWS is not an automated action but, rather, the deliberate decision of some military commander. In their deliberations, commanders are morally responsible for taking all reasonable steps to ensure that their orders comply with IHL proportionality and distinction requirements. Thus, in particular, commanders are morally responsible for activating an AWS only if their knowledge licences the judgment that the risk of running counter to IHL requirements is acceptably low. This moral requirement entails that commanders must be given in advance adequate information, based on extensive modelling and testing activities, about IHL-compliant activations of AWS and their boundary conditions. However, the epistemic uncertainties mentioned above vividly demonstrate that this information is presently unavailable. Therefore, commanders ought to refuse systematically to deploy AWS, given that they are not in a position to assert that the risk of running counter to IHL requirements is acceptably low. By the same token, the respect that is due to military commanders – *qua* agents who have to make real moral choices – demands that AWS should not be supplied as equipment for possible use in combat at least until the present epistemic uncertainties persist.

The latter conclusion about respect for the commander's moral agency is reinforced by considering the notion of dignity – in the sense of personal and rank dignity, which are discussed by Dieter Birnbacher in his contribution to this volume. Unlike human dignity, which is exclusively associated with human rights, Birnbacher points out that both personal and rank dignity are also associated with duties. Thus, in particular, the rank attributed to military commanders comes with the moral duty to assess the risk of violating IHL requirements and to set out their orders accordingly. Moreover, the commander's personal dignity is derivatively involved as well, insofar as the commanders' prerogatives and duties are an integral part of their culturally determined personal dignity.

To sum up, convincing arguments for an AWS moratorium have been advanced from both a deontological and a consequentialist standpoint in ethical theorizing. Arguments of both kinds involve as a crucial premise educated guesses about the current and foreseeable developments in robotics and artificial intelligence. Surprisingly enough, some public statements about AWS happen to neglect entirely these technological assessments and related epistemic predicaments, thus jeopardizing the correct development of democratic debates and giving rise to unjustified biases in decision-making processes about an AWS moratorium.

**Democratic decision making in the fog of AWS agnotology**

'A lot of people fear artificial intelligence', John Arquilla was quoted as claiming in a *New York Times* article of 28 November 2010, but 'I will stand my artificial intelligence against your human any day of the week and tell you that my A.I. will pay more attention to the rules of engagement and create fewer ethical lapses than a human force'.[16] This claim by the executive director of the Information Operations Center of the US Naval Postgraduate School echoes an earlier statement issued by Gordon Johnson of the Joint Forces Command at the Pentagon, who asserted that robotic soldiers 'don't get hungry, they're not afraid. They don't forget their orders. They don't care if the guy next to them has just been shot. Will they do a better job than humans? Yes'.[17]

There is an evident tension between these statements and the assessment of the current and foreseeable capabilities of robotic systems by scientists working in artificial intelligence and robotics.[18] One should carefully note that these statements, issued by experts in military affairs, are addressed to a wide audience of people who are generally unfamiliar with both weapon systems and the state of the art in artificial intelligence and robotic technologies. Tagged as well-founded expert opinions, these statements are likely to elicit or reinforce unrealistic beliefs and expectations in public opinion about robotic weapons in general and about AWS in particular. Since democratic decision making in technologically affluent societies is bound to rely on information that is supplied by experts, these unrealistic beliefs and expectations may unduly influence the formation of public opinion and democratic deliberations about AWS by ordinary citizens and their political representatives.

It is useful to distinguish the mechanism that gives rise to this form of consequential ignorance from other mechanisms and forms of ignorance production about AWS. Robert Proctor coined the word 'agnotology' to designate the study of the cultural production of ignorance and its effects on both individual and collective decision-making processes.[19] He identified various mechanisms of ignorance production, including both ignorance as an active construct and ignorance as selective choice. A blatant example of the former mechanism is the tobacco industry's policy of inducing doubts about the dangers of smoking. A more subtle example of the latter mechanism is any process of selecting research themes, which usually involves pruning alternative research themes.

---

[16] 'War Machines: Recruiting Robots for Combat', *New York Times* (28 November 2010), available at www.nytimes.com/2010/11/28/science/28robot.html?_r=0.
[17] 'New Model Army Soldier Rolls Closer to Battle', *New York Times* (16 February 2005), available at www.nytimes.com/2005/02/16/technology/new-model-army-soldierrolls-closer-to-battle.html.
[18] R.C. Arkin, 'Lethal autonomous systems'; N. Sharkey, 'Cassandra or the false prophet of doom: AI robots and war', *IEEE Intelligent Systems*, 23 (2008), 14; N. Sharkey, 'Grounds for discrimination: autonomous robot', *RUSI Defence Systems*, 11 (2008), 86; Sharkey, 'Saying 'no!' to lethal autonomous targeting'.
[19] R. Proctor, 'A missing term to describe the cultural production of ignorance (and its study)' in R. Proctor and L. Schiebinger (eds.), *Agnotology: The Making and Unmaking of Ignorance* (Stanford UP, 2008), 1.

Public debates and democratic deliberations concerning the prospective uses of AWS are hampered by ignorance as an active construct, insofar as secrecy and reticence often surround the development and deployment of military technologies. Moreover they are hampered by selective choice, to the extent that AWS research projects are preferred over research projects that more clearly prize the ethical, political and military advantages flowing from the meaningful human control on robotic weapons that Noel Sharkey examines in his contribution to this volume. And they are also hampered by what one may aptly call the *temporal framing* mechanism of ignorance production. This mechanism induces false beliefs about the time scales of envisaged technological advancements from a failure to convey clearly the distinction between the long-term (and often admittedly visionary) goals of ambitious technological research programs, on the one hand, and their expected short-term outcomes, on the other hand.[20]

The distinction between long-term and short-term goals is crucial to understand what actually goes on in many research programs in robotics. Long-term visions of research programs in home service robotics envisage robots that are endowed with the versatile competences of human butlers, tutors and assistants. These long-term visions are useful insofar as they play regulative and inspirational roles in inquiry. They should be clearly distinguished, however, from the short-term goals driving daily research activities on home robots. These technological short-term goals, unlike the underlying long-term visions, are expected to be feasible on the basis of state-of-the-art technologies and to feed the pipeline between technological inquiry and industry.[21] Thus, short-term goals of research on home robots are not concerned with the development of ideal butlers but, rather, with robots that selectively perform vacuum-cleaning jobs, tele-presence and medication-taking reminder services, assistance in emergency calls, and so on.

RoboCup affords another vivid illustration of the interactions between, and the respective roles of, the long-term and short-term goals of technological inquiry. RoboCup, which is familiar to the general public for its robotic soccer tournaments, cultivates the ambition of putting together a robotic soccer team that will beat the human world champion team. Fulfilling this long-term goal presupposes so many far-reaching advances in sensorimotor and cognitive skills of multi-agent robotic systems that one may sensibly doubt whether the research efforts of a few generations of

[20] Similarly asynchronous goal-pursuing processes are postulated in two-process models of scientific inquiry (Godfrey-Smith), which notably include Lakatos' scientific research programs and Laudan's research traditions. See P. Godfrey-Smith, *Theory and Reality: An Introduction to the Philosophy of Science* (University of Chicago Press, 2003); I. Lakatos, 'Falsification and the methodology of scientific research programmes' in J. Worrall and G. Currie (eds.), *Philosophical Papers*, Volume 1: *The Methodology of Scientific Research Programmes* (Cambridge UP, 1978), 8; L. Laudan, *Progress and Its Problems: Toward a Theory of Scientific Growth* (<publisher?>University of California Press, 1977). See also E. Datteri and G. Tamburrini, 'Robotic weapons and democratic decision-making' in E. Hilgendorf and J.-P. Guenther (eds.), *Robotik und Gesetzegebung* (Nomos Verlag, 2013), 211; G. Tamburrini, 'On the ethical framing of research programs in robotics', *AI and Society* (forthcoming).

[21] Tamburrini, 'On the ethical framing of research programs'.

committed scientists will suffice to bridge the gap between vision and reality. Beating the best human soccer team, the RoboCup manifesto acknowledges, 'will take decades of efforts, if not centuries. It is not feasible, with the current technologies, to accomplish this goal in any near term.' However, the RoboCup's elusive long-term goal has a significant role to play in the context of RoboCup research activities, insofar as it enables one to shape a fruitful research agenda by suggesting 'a series of well-directed subgoals', which are both feasible and technologically rewarding.[22] In particular, periodic RoboCup tournaments enable scientists to improve continually on the playing capabilities of robotic soccer teams and to identify, on the basis of the playing performances of the winning teams, the benchmarks that the robotic teams participating in the next tournaments will be confronted with.

Having an AWS achieve human-like capabilities in the way of IHL discrimination and proportionality requirements is comparable to the RoboCup's long-term, visionary goal of beating the human world champion soccer team with a team of robots. Both are formidable and possibly unattainable technological challenges. However, the RoboCup's long-term goal is prized for its role in shaping fruitful research agendas towards sub-goals that are feasible, technologically rewarding and morally permissible *per se*. In contrast, the long-term goal of AWS research may only result in the short term with a weapon system that fails to comply with IHL and that is not morally permitted on both deontological and consequentialist grounds.

Experts speaking to restricted circles of peers may abstain from signalling whether they are speaking from the long-term or short-term perspective of a research program. Indeed, shared background knowledge enables each member of the audience to identify which perspective the speaker is talking from. In public statements about their work, however, experts can no longer count on the shared background of tacit knowledge that shapes communication styles within communities of experts. Accordingly, experts wishing to provide correct and accessible public information must give adequate information about the expected temporal frame for the various goals of the research programs they talk about. Temporal framing mechanisms of ignorance production are activated by inadvertent or intentional failures to meet this challenge. And, clearly, the resulting agnotological effects are likely to be further amplified at the hands of both sensationalist media reporting and individual psychological responses.[23]

[22] See RoboCup, available at www.robocup.org/about-robocup/objective/. The full quotation is: 'Needless to say, the accomplishment of the ultimate goal will take decades of efforts, if not centuries. It is not feasible, with the current technologies, to accomplish this goal in any near term. However, this goal can easily create a series of well-directed subgoals. Such an approach is common in any ambitious, or overly ambitious, project.'

[23] For some pertinent psychological models explaining psychological responses overestimating or devaluating the expected outcomes of technological research, see F. Scalzone and G. Tamburrini, 'Human-robot interaction and psychoanalysis', *AI and Society*, 28 (2013), 297.

Moral responsibilities in democratic deliberation about AWS and other novel technologies are distributed in accordance with a sensible division of epistemic labour.[24] Citizens have the moral responsibility of reducing their consequential ignorance about scientific and technological matters, which threatens their moral values and aspirations. Experts carry the moral responsibility of supplying what is, to their best knowledge, correct and adequate scientific and technological information for democratic decision making. In many circumstances, however, citizens fail to collect the relevant background information, eventually choosing alternatives that conflict with their own interests and moral values. And experts exacerbate states of consequential ignorance by intentionally or inadvertently supplying inaccurate or incorrect information for social and political deliberation. It is worth contrasting, from this perspective of temporal framing ignorance production and its moral implications, the public statements by military experts quoted at the beginning of this section with the following statement about AWS attributed to Ronald Arkin in an *International Herald Tribune* article, dated from 26 November 2008: 'My research hypothesis is that intelligent robots can behave more ethically in the battlefield than humans currently can'.[25] Arkin is careful to emphasize that he is advancing a research hypothesis, thereby implying that his educated guess may be refuted like any other research hypothesis, remain unfulfilled for a long time to come or even be relinquished in the long run for persistent lack of substantial rewards. Elsewhere, he states that '[i]t is too early to tell whether this venture will be successful', emphasizing that there are 'daunting problems' of a technical nature that remain to be solved.[26] No trace of similar temporal and epistemic qualifications is found in the public statements by Arquilla and Johnson quoted at the beginning of this section.

In conclusion, temporal frame mechanisms of ignorance production give rise to screening-off effects and biases that unduly affect democratic debates and decision making about AWS. Communication ethics demands that all stakeholders in these debates carefully check their statements for these effects. Once temporal frames are correctly conveyed, one has to come to terms with the fact that developing perceptual, reasoning and action capabilities that may enable an AWS to surpass a human soldier in the way of IHL compliance is a formidable and possibly unattainable technological challenge – no less formidable than the RoboCup's regulative idea of beating the human world champion soccer team with a team of robots. Thus, a proper understanding of the involved temporal scales enables one to endorse a major premise of the arguments that, from both a deontological and consequentialist viewpoint in normative ethics, converge on the need for an AWS moratorium.

[24] P. Kitcher, *Science in a Democratic Society* (Prometheus Books, 2011).
[25] 'Robots May Be More "Humane" Soldier', *International Herald Tribune* (26 November 2008).
[26] Arkin, *Governing Lethal Behavior in Autonomous Robots*.

**Wide consequentialist reasons for banning AWS**

Let us finally turn to a consideration of the arguments for a ban on AWS. The epistemic predicaments that loom so large on arguments for a moratorium play only a minor role. Indeed, the upshot of the ethical arguments for banning AWS is to show that, no matter how well AWS will come to perform their targeting and engagement tasks, there are overriding moral reasons to forbid their use. Asaro's argument for banning AWS, which is supposed to apply to any conceivable AWS, no matter whether or how well it complies with IHL, is advanced from a distinctively deontological standpoint in normative ethics.[27] According to Asaro, human beings have the right of not being deprived of their life arbitrarily – that is, without the respect that other human beings owe to them as potential victims of lethal force. For killing decisions to count as non-arbitrary, Asaro argues, they must be taken on the basis of a responsible exercise of human judgment and compassion. Since AWS fail to meet these requirements, their use must be absolutely prohibited. In particular, Asaro remarks: 'The decision to kill a human can only be legitimate if it is non-arbitrary, and there is no way to guarantee that the use of force is not arbitrary without human control, supervision and responsibility. It is thus immoral to kill without the involvement of human reason, judgement and compassion, and it should be illegal.' And he goes on to claim that'[a]s a matter of the preservation of human morality, dignity, justice and law, we cannot accept an automated system making the decision to take a human life. And we should respect this by prohibiting autonomous weapon systems. When it comes to killing, each instance is deserving of human attention and consideration in light of the moral weight that is inherent in the active taking of a human life.'[28]

It was suggested that a ban on AWS can be supported exclusively if one endorses a similar deontological standpoint in normative ethics and shares with Asaro the view that potential victims of lethal force in warfare have certain inalienable rights: 'The moral support for a ban on the deployment of any autonomous robotic weapon depends entirely on whether it is decided that there

[27] Asaro, 'On banning autonomous weapon systems'.

[28] See Asaro, 'On banning autonomous weapon systems', 708. It is worth noting that Asaro's appeal to human dignity can be construed in terms of both a Kantian conception of human dignity and the recently revived conception of human dignity as rank, already mentioned above, and according to which a high-ranking status must be extended to every human being. See J. Waldron, *Dignity, Rank and Rights* (Oxford UP, 2013). This generalization of status affords the same kind of protection from degrading treatment offered by the familiar Kantian construals of dignity. According to Asaro, delegating to AWS life or death decisions results in inhumane and degrading treatment of potential victims. Christof Heyns examines Asaro's argument in the context of a call for an AWS moratorium (see Heyns, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Execution. Lieblich and Benvenisti, in their contribution to this volume, extend Asaro's motives for protecting potential victims of AWS attacks; and Birnbacher, in his contribution to this volume, critically examines Asaro's use of the notion of human dignity.

is a human right not to be the target of a robotic weapon'.[29] However, this claim overlooks the fact that significant arguments for a ban on AWS have been advanced neither on the basis of a deontological framework in normative ethics nor by appealing to fundamental human rights but, rather, by adopting a purely consequentialist standpoint. Any consequentialist argument for or against a ban on AWS is presently bound to focus on the expected, rather than the actual, consequences of their deployment since these weapon systems have not yet been deployed. Building on a distinction between narrow and wide consequentialist reasons, it is argued here that by sufficiently enlarging the temporal and spatial horizon of the expected consequences, the consequentialist reasons offered for a ban largely outweigh those offered for future AWS deployment.

To begin with, let us recall that the consequentialist reasons for the future deployment of AWS include reduced casualties not only in one's own and the opponents' camp but also among non-belligerents as a result of more accurate targeting and a more conservative decision to fire, which are free from human self-preservation concerns.[30] These reasons for the future deployment of AWS concern only expected battlefield performances and some of their outcomes. Consequences that one may expect on a more global scale are neglected, therefore revealing a narrow consequentialist perspective on AWS' future deployment. Instead, a broad consequentialist standpoint takes into account the expected effects on peace stability, on incentives to start wars by the newly introduced conventional armament, on the likelihood of escalation from conventional to nuclear warfare and on the disruption of extant nuclear deterrence factors. Sharkey takes a broad consequentialist standpoint when he points out that 'having robots to reduce the "body-bag count" could mean fewer disincentives to start wars', thereby suggesting that a reduction of one's own casualties in the short term cannot compensate for the higher numbers of casualties and destruction caused by increased numbers of conflicts that AWS may facilitate in the long term.[31] On more general grounds, Jürgen Altmann suggests that the list of issues that are usually addressed be extended if one wants to provide a balanced aggregate assessment of the expected costs and benefits flowing from future AWS deployment.

If new classes of conventional weapons are emerging, as is the case with armed uninhabited vehicles, they should be assessed with respect to questions such as do they make war more likely and do they raise other dangers. Envisioned short-term military advantages should be weighed

[29] J.P. Sullins, 'An ethical analysis of the case for robotic weapons arms control' in K. Podins, J. Stinissen and M. Maybaum (eds.), *Proceedings of the Fifth International Conference on Cyber Conflict* (NATO CCD COE Publications, 2013), 487, 497.
[30] Arkin, *Governing Lethal Behavior in Autonomous Robots*.
[31] Sharkey, 'Cassandra or the false prophet of doom', 16; see also F. Sauer and N. Schörnig, 'Killer drones: the silver bullet of democratic warfare?' *Security Dialogue*, 34 (2012), 363, 365.

against the probable long-term consequences for national, and, in particular, international, security.[32] AWS are potentially more threatening to global security than many other conventional weapons. In particular, swarms of aerial AWS that are capable of initiating coordinated attacks on great numbers of civilian infrastructures and military objectives raise serious concerns in connection with a new arms race and its expected impact on global destabilization. This observation shows the implausibility of the *ceteris paribus* assumption that the deployment of AWS on the battlefield will not have an ethically significant impact on causal factors and strategic reasons underlying the decision to start or escalate armed conflicts. However, this is exactly the implicit assumption from which the force of narrow consequentialist arguments for the future deployment of AWS entirely depends.

The threat of destabilization raised by swarms of AWS may be a sufficiently serious incentive for a conventional war. However, one should be careful to note that AWS, more than many other conventional arms, have the potential to deliver destructive attacks on strategic nuclear objectives. Swarms of AWS might be capable of delivering a powerful first strike against the opponent's nuclear arsenals, to the extent that they may thwart the opponent's second strike capability of responding with nuclear retaliation. In this scenario, traditional nuclear deterrence based on mutually assured destruction would no longer be appealing and first strike strategies would be prized instead.

Let us now try and assess from a wide consequentialist standpoint the aggregate of expected benefits and costs flowing from AWS deployment. By permitting the future deployment of AWS, one might expect reduced casualties among belligerents and non-belligerents in some battlefield scenarios. At the same time, however, one would significantly raise the risk of a new arms race and global destabilization, by providing incentives for the commencement of wars and by weakening traditional nuclear deterrence factors based on mutually assured destruction. As far as the latter kind of risk is concerned, one can hardly think of a more critical danger to humankind than the danger of setting off a nuclear conflict, and one can hardly think of a more desirable state to humankind than the state of nuclear peace preservation. Since the expected costs of an arms race and destabilization outweigh the sum of the expected benefits flowing from AWS future deployment, opting for a pre-emptive ban on AWS is tantamount to choosing the collective rule of behaviour that is expected to produce the most preferable set of consequences in a global geopolitical context.

[32] See also J. Altmann, 'Preventive arms control for uninhabited military vehicles' in R. Capurro and M. Nagenborg (eds.), *Ethics and Robotics* (IOS Press, 2009) 69, 80–1: 'Seen from a narrow standpoint of national military strength, these developments will provide better possibilities to fight wars and to prevail in them. However, if one looks at the international system with its interactions, the judgment will be different, in particular concerning armed robots/uninhabited systems. Destabilization and proliferation could make war more probable, including between great/nuclear powers.'.

In conclusion, there is a strong confluence on an international pre-emptive ban on AWS from both a deontological and a broad consequentialist standpoint in normative ethics. The introduction of a ban on AWS will raise the problem of enforcing international interdictions on the development, production and deployment of AWS. Runaway development of AWS under a ban will be facilitated by adaptations of cutting-edge robotic technologies that were originally intended for civilian applications. For example, European Union (EU) programs supporting research in robotics have excluded the funding of scientific and technological research for military applications. However, the outcomes of EU projects for, say, swarming robot technologies can be used straightforwardly to develop swarms of AWS, such as the swarms of autonomous robot boats already developed by the US Navy.[33] Artificial intelligence scientist Stuart Russell goes as far as suggesting that 'the technology already demonstrated for self-driving cars, together with the human-like tactical control learned by DeepMind's DQN system, could support urban search-and-destroy missions'.[34] Accordingly, in a comprehensive system of compliance for a pre-emptive ban on AWS measures,[35] one will have to include the careful assessment of advances in robotics that are made possible by non-military research programs so as to adequately benefit monitoring and early-warning procedures.

[33] For European Union-funded projects on swarming robots, see Horizon 2020, The Way of the Future: 'Swarming' Robots, 6 February 2014, available at http://ec.europa.eu/programmes/horizon2020/en/news/way-future-%E2%80%98swarming%E2%80%99-robots; for more information on navy swarm boats, see 'US Navy Could 'Swarm' Foes with Robot Boats, *CNN* (13 October 2014), available at http://edition.cnn.com/2014/10/06/tech/innovation/navy-swarm-boats/.

[34] S. Russell, 'Take a stand on AI weapons', *Nature* 521 (2015), 415.

[35] M. Gubrund and J. Altmann, 'Compliance measures for an autonomous weapons convention', *ICRAC Working Papers*, 2 (2013), available at http://icrac.net/resources/.