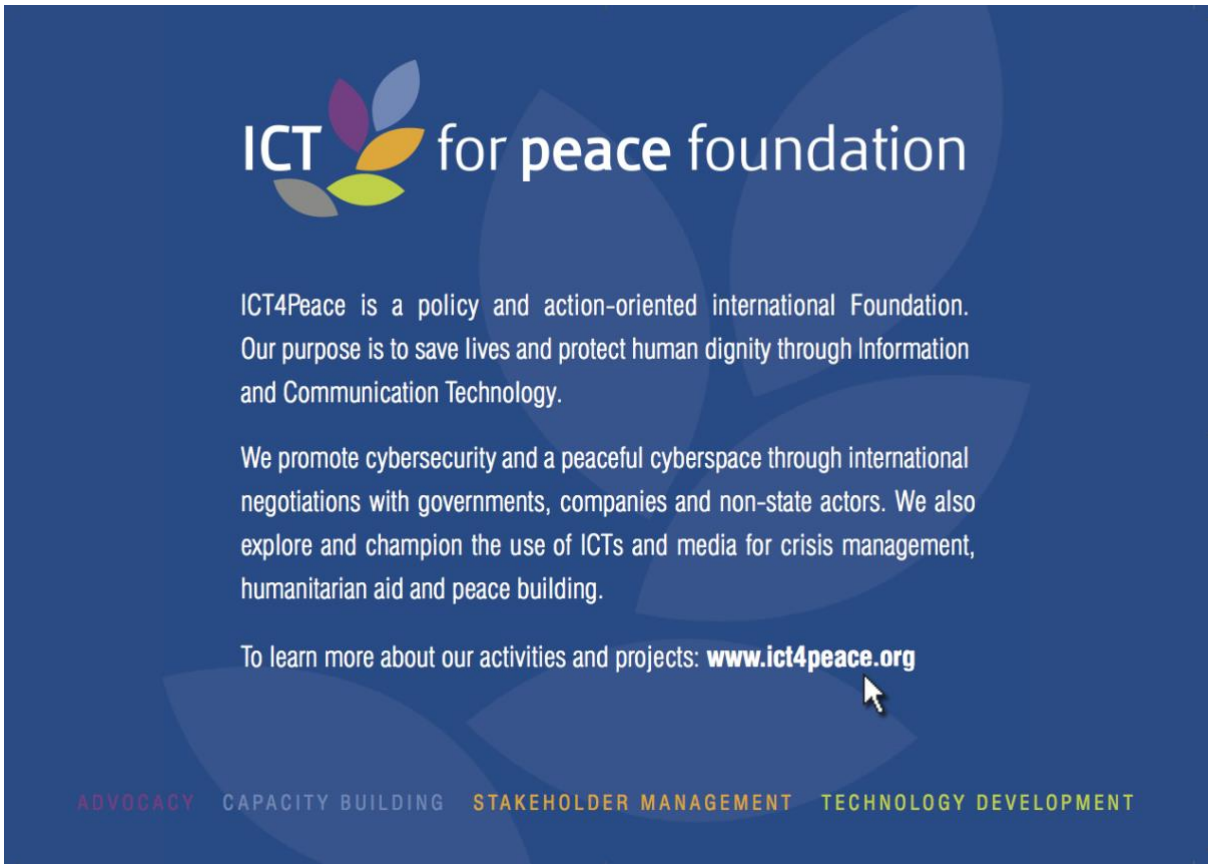
Artificial Intelligence: Autonomous Technology (AT), Lethal Autonomous Weapons Systems (LAWS) and Peace Time Threats

By Regina Surber, Scientific Advisor, ICT4Peace Foundation and the Zurich Hub for Ethics and Technology (ZHET)



Zurich, 21 February 2018

Table of Contents

List of Abbreviations

AAAI	Association for the Advancement of Artificial Intelligence
ACM	Association for Computing Machinery
AI	Artificial Intelligence
AT	Autonomous Technology
CBM	Confidence Building Measures
CCW	Convention on Certain Conventional Weapons
DNA	Deoxyribonucleic Acid
EPSRC	Engineering and Physical Science Research Council
EURON	European Robotics Research Network
FLI	Future of Life Institute
GAN	Generative Adversarial Network
GGE	Group of Governmental Experts
HRW	Human Rights Watch
IDSIA	Instituto Dalle Molle di Studi sull'Instelligenza Artificiale / Dalle Molle Institute for Artificial Intelligence Research
IEEE	Institute of Electrical and Electronics Engineers
IHL	International Humanitarian Law
IHRL	International Human Rights Law
LAWS	Lethal Autonomous Weapons Systems
UN	United Nations

1 Introduction

The main purpose of this paper is to inform the international community on the risks of Autonomous Technology (AT) for global society. AT can be said to be the essence of Lethal *Autonomous* Weapons Systems (LAWS), which have triggered a legal and policy debate within the international arms control framework of the United Nations Convention on Certain Conventional Weapons (UN CCW) that is now entering its fifth year. Since LAWS highly challenge existing International Humanitarian Law (IHL) due to their capacity of replacing a human operator on a weapons platform, the CCW's tasks of, i.a., ensuring that the concepts of legal accountability and human responsibility do not become void, and assessing whether LAWS are legal under IHL, are of utmost importance.

However, LAWS are not the only manifestation of the security risks of AT. This paper will demonstrate further ways of the actual and potential weaponization of AT that are currently not yet fully addressed by the UN organizations. Moreover, AT not only poses risks to global society if *weaponized*, but can pose tremendous systemic risks to global society and humanity also when *not weaponized*. This potentially dangerous transformative power of AT, which is beyond the scope of the CCW's mandate, will be the thematic core of this paper. Based on a risk assessment of *not-weaponized* AT, the paper will present thought-provoking impulses that can shape an international interdisciplinary debate on the risks of AT specifically and of emerging technologies more generally.

In addition, this paper highlights risks underlying the application of terms originally referring to *human* traits to technological artefacts, such as 'intelligence', 'autonomy', 'decision-making capacity' or 'agent'. It will argue that this unquestioned so-called 'anthropomorphism' leads to a premature overvaluation of technology and a simultaneous potential devaluation of human beings, and will present ideas for linguistic substitutes.

At the same time, the paper will illustrate that the 'classical' understanding of 'autonomy' as human '*personal* autonomy' has, in fact, donated its meaning to the current technological use of the term. However, this fact risks to be obfuscated by the broadening pool of diverse definitions and understandings of 'autonomy' for technological artefacts. Thereby, the paper will unearth the current paradigm shift in human technological creation and self-understanding that underlies the ongoing debate on AT and LAWS: The fact that humans are creating technological artefacts that may lose their instrumental character because we gradually give away control and responsibility for the outcomes of their usage. Locating the core challenge of AT, AI and any emerging technology in this still subtle but pervasive change in the understanding of the human-technology relationship, this paper will also provide conclusions and recommendations that are of a more general and long-term character.

The paper will be structured as follows: Chapters 2 and 3 will describe the current understandings, uses as well as the risks of those uses of the terms 'Artificial Intelligence' (AI) and 'Autonomous Technology' (AT). Chapter 4 will introduce the term 'Lethal Autonomous Weapons System' (LAWS), which will lead over to Chapter 5 on the international discussions within the UN CCW and this UN debate's limitations. Chapter 6 will present further ways of

weaponizing AT that are ignored by the UN CCW, yet need immediate attention. Chapter 7 shows threats of AT for global society during peace-time. Chapter 8 presents three arguments for shaping an international debate on AT, AI and LAWS. Chapter 9 concludes and presents a list of recommendations. Eleven lists of principles for ethical/ responsible research on AI, AT and Robotics can be found in the annex.

2 Artificial Intelligence (AI)

AI are two letters that represent the financially most lucrative scientific field that currently exists.[1] Moreover, they represent something that is often regarded as the fuel of the fourth industrial revolution, which is taking place at an unprecedented pace compared to any other in human history.[2] However, the question of what AI really *is* most often receives a rather vague and elusive answer. The reason for this lack of clarity may by two-fold.

First, the term 'Artificial Intelligence' includes the term 'intelligence.' 'Intelligence' originally has been used as a characteristic of humans. However, there neither exists a general understanding of this natural trait, nor a standard definition, despite a long history of research and debate.[3]

Precisely due to the growing research on AI, there exist strong incentives to define what the term 'intelligence' shall mean. This need is especially acute when artificial systems are considered that are significantly different to humans. This is the reason why researchers at the Swiss AI Lab IDSIA (Instituto Dalle Molle di Studi sull'Intelligenza Artificiale) created a single definition based on a collection of 70 definitions of 'Intelligence' by dictionaries, psychologists and AI researchers. They state that *'intelligence measures an agent's ability to achieve goals in a wide range of environments.*' This general ability includes the ability to understand, to learn

[1] The Economist, 2017, Coding Competition, The Battle in AI, The Economist Online, December 7, 2017, available at: https://www.economist.com/news/leaders/21732111-artificial-intelligence-looks-tailor-made-incumbent-tech-giants-worry-battle?frsc=dg%7Cehttp://www.economist.com/news/leaders/21732111-artificial-intelligence-looks-tailor-made-incumbent-tech-giants-worry-battle?frsc=dg%7Ce (accessed on December 11, 2017).

[2] See e.g. Wan, James, 2018, Artificial Intelligence is the fourth industrial revolution, Lexology.com, January 18, 2018, available at: https://www.lexology.com/library/detail.aspx?g=fccf419c-6339-48b0-94f9-2313dd6f5186 (accessed on January 31, 2018); Kelnar, David, 2016, The fourth industrial revolution: a primer on Artificial Intelligence (AI), Medium.com, December 2, 2016, available at: https://medium.com/mmc-writes/the-fourth-industrial-revolution-a-primer-on-artificial-intelligence-ai-ff5e7fffcae1 (accessed on January 31, 2018); Wright, Ian, 2017, Artificial Intelligence and Industry 4.0 – Taking the Plunge, Engineering.com, October 19, 2017, available at: https://www.engineering.com/AdvancedManufacturing/ArticleID/15871/Artificial-Intelligence-and-Industry-40--Taking-the-Plunge.aspx (accessed on January 31, 2018). Some experts also say that we are currently in the middle of a digital revolution, see e.g. Helbing, Dirk, 2017, A Presentation on Responsible Innovation and Ethics in Engineering, November 11, 2017, available at: https://www.youtube.com/watch?v=Jyv3QpRp9LA (accessed on February 7, 2018). Research by McKinsey suggests that AI could potentially transform global society '[…] ten times faster and 300 times the scale, or roughly 3000 times the impact,' Dobbs, R., Manyika, J. and Woetzel, J., 2015, The four global forces breaking all trends, McKinsey&Company, available at: https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-four-global-forces-breaking-all-the-trends (accessed on February 3, 2018).

[3] Helbing, Dirk, 2018, Personal Interview, February 9, 2018. For a list of 70 definitions of 'Intelligence' see Legg, Shane, and Hutter, Marcus, 2007, A Collection of Definitions of Intelligence, Frontiers in Artificial Intelligence and Applications Vol 157, 17-24.

and to adapt, since those are the features that enable an agent to solve a problem in a wide range of environments.[4]

It must be highlighted that the driving force behind the above-mentioned definition was to create a definitional reference point useful for *both human* as well as *technological artefacts*.[5] This ignores the fact that the term 'intelligence' was originally used to refer to a natural *human* capacity; and without a clear understanding of this human trait, we could possibly risk an overvaluation of technology and a devaluation of human beings.[6]

And second, a reason for confusion about the meaning of AI may lie in the fact that the term AI is used to refer to two distinct but interrelated understandings. The distinction of these two possible understandings of AI will be highlighted here by two definitions of AI. However, we do not claim for these definitions to gain universal validity, as they would merely increase the existing pool of possible choices of such definitions. Yet, they should provide the reader with a first sense of caution when dealing with the application of originally 'human terms' such as 'intelligence' or 'autonomy' to technological artefacts. At first glance, it might seem accurate and comprehensive to apply originally human terms to technological artefacts, since the latter are increasingly capable to perform 'actions' that resemble those of humans. However, the elaborations in this paper will show that this could prove to be risky.

On the one hand, AI refers to a scientific field, whose modern history started with the development of stored-program electronic computers,[7] but whose intellectual roots can already be found in Greek mythology.[8] As a scientific field, AI can be regarded as the attempt to answer the question of how the human brain gives rise to thoughts and feelings. AI as a research field began with the idea that '[…] every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.'[9] Therefore, AI refers to '[…] the *study* of the computations that make it possible to perceive, reason, and act';[10] it is the '[…] *effort* to make computers think […];'[11] and it is the '[…] *art* of creating machines that perform functions that require intelligence when performed by people.'[12]

[4] Legg and Hutter, 2007, 8.
[5] Legg, Shane, and Hutter, Marcus, 2006, A formal measure of artificial intelligence, Proc. 15th Annual Machine Learning Conference of Belgium and The Netherlands, 73-80, 73.
[6] Helbing, 2018.
[7] A computer that stores program instructions in electronic memory.
[8] See e.g. on the bronze man Talos from Crete, who can be regarded as incorporating the idea of an intelligent robot: Appollodorus, The Library, Book 1, Chapter 9, Section 26, Frazer, Sir James George (trnsl.), 1921, Cambridge, MA: Harvard University Press; London: William Heinemann Ltd; Apollonius Rhodius, Argonautica, Book 4, Section 1638 et seq., Seaton, R.C. (trnsl.), 1912, London: William Heinemann; Cohen, J., 1966, Human Robots in Myth and Science, London: Allen and Unwin.
[9] McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E., 1955, A proposal for the Dartmouth summer research project on artificial intelligence, 1.
[10] Winston, Addison-Wesley, 1992, Artificial Intelligence 3rd ed., Boston, MA: Longman Publishing Co, emphasis added.
[11] Haugeland, John, 1985, Symbolic Computation: Artificial Intelligence: The Very Idea, Cambridge, MA: The MIT Press, emphasis added.
[12] Kurzweil, Raymond, 1990, The Age of Intelligent Machines, Chapter 1: The Roots of Artificial Intelligence, 2, emphasis added.

Bearing in mind the above-mentioned risk of devaluating humans in creating a definition of (artificial) intelligence without a human reference, AI shall here be understood as

(1) a scientific undertaking that is aiming to create software or machines that exhibit traits that resemble human reasoning, problem-solving, perception, learning, planning, and/ or knowledge.

Core parts of research on AI include: 'knowledge engineering,' which aims at creating software and machines that have abundant information relating to the world; 'machine learning', which is the modern probabilistic approach to AI and that studies algorithms that 'learn' to predict from data; 'reinforcement learning', a sub-discipline of machine learning and currently the most promising approach for general intelligence that studies algorithms that learn to act in an unknown environment through trial and error; 'deep learning'[13], a very fast-moving and successful approach to machine learning based on neural networks, which has enabled recent breakthroughs in computer vision and speech recognition;[14] 'machine perception', which deals with the capability of using sensory inputs to deduce aspects of the world, 'computer vision', the capability of analyzing visual inputs; and 'robotics', which deals with robots and the computer systems for their control.[15]

On the other hand, AI is also referred to the 'knowledge' or 'capacity' *embedded* in software or hardware architecture that are the result of the research on AI (1). Such capacities of software or hardware, e.g. the capacity to 'recognize' faces or voices or to 'drive' without a human behind a steering wheel, can be understood as artificially created intelligence – or AI. In this sense, AI can be regarded as a resource or a commodity, because it can be traded. Tech giants around the world are competing about the brilliance of their respective algorithms.[16] Therefore, AI can be regarded both as a formless potential foundation of wealth as well as a resource for political leverage.[17]

In this sense, AI can also be understood as

(2) the formless capacity embedded in software and hardware architecture which enables the latter to exhibit traits that resemble human reasoning, problem-solving, perception, learning, planning, and/ or knowledge.

[13] For a more detailed description of deep learning, see p. 5.
[14] Leike, J., AI Safety Syllabus, 80.000hours.org, available at: https://80000hours.org/ai-safety-syllabus/ (accessed on February 3, 2018).
[15] See e.g., Techopedia.com, Artificial Intelligence, available at: https://www.techopedia.com/definition/190/artificial-intelligence-ai (accessed on January 31, 2018).
[16] The Economist, 2017, Battle of the brains, Google leads in the race to dominate artificial intelligence, December 7, 2017, available at: https://www.economist.com/news/business/21732125-tech-giants-are-investing-billions-transformative-technology-google-leads-race (accessed on January 31, 2018).
[17] See e.g. CNBC, 2017, Putin: Leader in artificial intelligence will rule the world, September 4, 2017, available at: https://www.cnbc.com/2017/09/04/putin-leader-in-artificial-intelligence-will-rule-world.html (accessed on February 7, 2018); Metz, Cade, 2017, Google is already late to China's AI revolution, February 2, 2017, Wired.com, available at: https://www.wired.com/2017/06/ai-revolution-bigger-google-facebook-microsoft/ (accessed on February 7, 2018), Armbruster, Alexander, 2017, Künstliche Intelligenz: Google-Manager Eric Schmidt warnt vor China, Frankfurter Allgemeine Zeitung Online, November 2, 2017, available at: http://www.faz.net/aktuell/wirtschaft/kuenstliche-intelligenz/kuenstliche-intelligenz-google-manager-eric-schmidt-warnt-vor-china-15273843.html (accessed on February 7, 2018).

Current AI in the second sense of the term (2) is known as 'narrow' or 'weak' AI, in that it is designed to perform a narrow task, such as *only* driving a car or *only* recognizing faces. The long-term goal of many researchers, however, is to create so-called 'general' or 'strong' AI, sometimes also called 'artificial human-level intelligence'.[18] General AI is the formless capacity embedded in general purpose systems that are comparable to that of the human mind.[19] If general AI was achieved, this might also lead to 'artificial superintelligence', which can be defined as '[…] any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.'[20]

3 Autonomous Technology (AT)

AT is a result of research in the fields of AI and robotics, but also draws on other disciplines such as mathematics, psychology and biology.[21] Currently, there exists no clear understanding and no universally valid definition of the term 'autonomous' or AT in the context of AI and robotics. However, there do exist different attempts.

Sometimes a purely operational understanding of 'autonomy' is used. In this sense, the term 'autonomous' may refer to any outcome by a machine or software that is created without human intervention. This could include, e.g., a toaster's ejection of a bread slice when it is warm. In this form, autonomy would be equivalent to automation[22] and would not be limited to digital technology but could be used in analog technology or mechanics as well.[23] Hence, this understanding does not locate AT exclusively within the research field of modern AI.

Some experts use a narrower understanding and limit the use of the attribute 'autonomous' to more complex technological processes. They argue that AT extends beyond conventional automation and can solve application problems by using materially different algorithms and software system architectures.[24] This perspective is narrower and clearly locates the emergence of AT within the research of modern AI.

In this sense, the key benefit of AT is the ability of an autonomous system to '[…] explore the possibilities for action and decide 'what to do next' with little or no human involvement, and to do so in *unstructured situations* which may possess *significant uncertainty*. This process is, in practice, *indeterminate* in that we cannot foresee all possible relevant information […]. 'What to do next' may include […] a step in problem-solving, a change in attention, the creation

[18] Müller, Vincent C., and Bostrom, Nick, 2016, Future Progress in Artificial Intelligence: A Survey of Expert Opinion, In: Müller, Vincent C., (ed.), Fundamental Issues of Artificial Intelligence, Synthese Library; Berlin: Springer, 553-571, 553.
[19] Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S., and Sowa, J. F., 2012, Mapping the landscape of human-level artificial general intelligence, AI Magazine, 33(1), 25-42.
[20] Bostrom, N., 2014, Superintelligence: Paths, dangers, strategies, Oxford: Oxford University Press, Ch. 2.
[21] Atkinson, David J., 2015, Emerging Cyber-Security Issues of Autonomy and the Psychopathology of Intelligent Machines, Foundation of Autonomy and Its (Cyber) Threats: From Individuals to Interdependence: Papers from the 2015 Spring Symposium, 6-13, 7.
[22] Christen, Markus, Burri, Thomas, Chapa, Joseph, Salvi, Raphael, Santoni de Sio, Filippo, and Sullins, John, 2017, An Evaluation Scheme for the Ethical Use of Autonomous Robotic Systems in Security Applications, Digital Society Initiative (DSI) of the University of Zurich, DSI White Paper Series, White Paper No. 1, 36.
[23] Helbing, 2018.
[24] Land mines are an often-cited example of an automated weapon, see e.g. Ibid., 46.

or pursuit of a goal, and many other activities […].'[25] In other words, a system is 'autonomous' if it can change its behavior during operation in response to events that are *unanticipated*,[26] e.g. a self-driving car's reaction to a traffic jam, a therapist chatbot's[27] answer to a person lamenting about her disastrous day, or a missile defense system that intercepts an incoming hostile missile, like Israel's Iron Dome.

The theoretical AI approach that is at the core of AT in its narrow understanding, and that enables technological systems to perform the above-mentioned actions without a human operator, is deep learning. Deep learning software tries to imitate the activity of layers of neurons in the human brain. Through improvements in mathematical formulas and the continuously increasing computing power of computers, it is possible to model a huge number of layers of virtual neurons. Through an inflow of a vast amount of data, the software can recognize patterns in this data and 'learn' from it.[28] This is key for 'autonomous' systems' reaction to unanticipated changes: due to new data inflow, the software can recognize new patterns and adapt to a changing 'environment'. Thereby, an autonomous system can modify its actions in order to follow its goal or agenda.

It is crucial to highlight that deep learning mechanisms are so complex that a human being cannot comprehend why a technological process based on deep learning creates the outcome it does.[29] Hence, outputs of autonomous systems may not only come as a surprise due to their core capacity of choosing a course of action undetermined by a human operator, but also due to the impossibility of locating the technological 'trigger' for a certain output.

At this stage one could address the fear that software or machines could independently create something that may resemble free will. It is a fact that autonomous systems may perform actions that are both unanticipated and ex post untraceable. However, the initial programming of the software with the potential for future 'autonomous behavior' is the engineer's and programmer's decision, and not an unavoidable fact. It is up to humans to discuss and set standards that ensure the development of beneficial and safe technology.

Since there exists no agreement whether automated systems (e.g. a toaster) should already be regarded as autonomous (no human operator controls the ejection of the warm bread), some experts see it useful to think of 'autonomy as a continuum'[30] or of 'degrees of

[25] Atkinson, 2015, 7, italics added. For further elaborations a limited use of the term 'autonomy' to its more complex forms, see Russell, Stuart J. and Norvig, Peter, 2014, Artificial intelligence: a modern approach, Third Edition, Pearson Education: Harlow; Van der Vyver, J.-J. et al., 2004, Towards genuine machine autonomy, in: Robotics and Autonomous Systems, Vol. 46, No. 3, 151-157.

[26] Watson, David P., and Scheidt, David H., 2005, Autonomous Systems, Johns Hopkins APL Technical Digest 26(4), 368-376, 368.

[27] See e.g. the 24/7 Woebot that chats in order to improve someone's mood, available at: https://woebot.io/ (accessed on February 14, 2018).

[28] Burkhalter, Patrick, 2018, Personal Interview, February 14, 2018.

[29] Ibid. See also Knight, Will, 2017a, The Dark Secret at the Heart of AI, MIT Technology Review, April 11, 2017, available at: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (accessed on February 16, 2018).

[30] Asaro, Peter, 2009, How just could a robot war be?, Proceedings of the 2008 Conference on Current Issues in Computing and Philosophy, 50-64, 51; Nicholas Marsh, Defining the Scope of Autonomy: Issues for the Campaign to Stop Killer Robots 2 (2014), available at: http://file.prio.no/Publication_files/Prio/Marsh%20(2014)%20-%20Defining%20the%20Scope%20of%20Autonomy,%20PRIO%20Policy%20Brief%202-2014.pdf (accessed on February 14, 2018); Michael Biontino, Summary of Technical Issues: CCW Expert Meeting on Lethal Autonomous Weapons Systems 1

autonomy'.[31] They would characterize automated processes or semi-autonomous processes as 'autonomous', however to a lower degree than 'fully autonomous' systems.[32] This takes into account a blurring of definitional borders and reflects the lack of a clear definition of 'autonomy' in AI and Robotics, but does not fill this gap.

There is also no agreement regarding whether or not a system could be classified as 'autonomous' if only certain aspects of its capacities function without human intervention. Some experts argue that, e.g., a system that can function independently from external energy sources (autarkic), or one that can adapt its programming behavior based on previous data acquired ('learning'), could already be regarded as 'autonomous'.[33]

Some experts also claim that the attribute 'autonomous' is used for a technological artefact when it becomes (nearly) impossible for a human being to intervene in a technological process. In this sense, 'autonomy' is not a term that covers a set of clearly defined characteristics (e.g. an artificial agent's capacity to 'learn', to be autarkic, to function independently from human control), but one that describes the *result of a technological process for which the human cannot or does not want to bear responsibility*.[34]

This view is influenced by the highly important, and thus not negligible fact, that the term 'autonomy' has a rich philosophical history and refers to an unquantifiable attribute intrinsic to human personhood. There are two distinct but interrelated understandings of 'autonomy' as a human attribute.

'Personal autonomy', on the one hand, refers to self-governance or the capacity to decide for oneself and follow a course of action in one's life, independent of moral content.[35] This necessarily leads to personal responsibility for the course of action taken.

On the other hand, 'moral autonomy,' usually traced back to Immanuel Kant, can be understood as the capacity of an individual human to deliberate, understand and give oneself the moral law. For Kant, it is by virtue of our autonomy that we are moral beings that can take on moral responsibility. At the same time, we are moral to the extent that we are autonomous.[36]

This second classical understanding of *moral* autonomy, connected with the fact that the term 'autonomy' is used when referring to software and machines, may have prematurely supported the idea of, and fueled discussions about 'autonomous' robots that may also behave

(2014), available at http://www.unog.ch/80256EDD006B8954/%28httpAssets%29/6035B96DE2BE0C59C1257CDA00553F03/$file/Germany_LAWS_Technical_Summary_2014.pdf (accessed on February 14, 2018).

[31] Christen et al., 2017, 10.

[32] Schörrig, Niklas, 2017, Automatisierung in der Militär- und Waffentechnik, 27. ETH-Arbeitstagung zur Sicherheitspolitik, Autonome Waffensysteme und ihre Folgen für die Sicherheitspolitik, February 3, 2017.

[33] Christen et al., 10.

[34] Helbing, 2018.

[35] Dryden, Jane, Internet Encyclopedia of Philosophy, Autonomy, available at: https://www.iep.utm.edu/autonomy/ (accessed on February 1, 2018).

[36] Kant, Immanuel, 1998 (1785), Groundwork for the Metaphysics of Morals, Cambridge: Cambridge University Press.

morally and *ethically*.[37] Both a precise technological understanding as well as careful linguistic usage may minimize or eliminate the risk of a (potentially unconscious) terminological confusion.[38]

However, it is barely possible to completely strip off a term from its 'classical' meaning. And the fact that 'autonomy', when used to characterize technological processes, does so when the latter create outcomes for which humans have a hard time taking control – in other words, when they actually *do* give away the capacity to decide for an action that leads to a technological process' outcome – there clearly exists an overlap of the 'classical' understanding of *personal* autonomy and the technological use of the term.

Due to this common contextual denominator of *personal* autonomy and 'autonomy' for technological artefacts, one could argue that the international debate about a definition of 'autonomy' for artefacts clearly distinguished from *personal* autonomy is misguided. The reason is that the technological use of the term 'autonomy' precisely uses this term in order to highlight a notion of 'self-governance' of an artefact. And whether or not this 'self-governance' is in fact technologically possible, one must *not* ignore that research endeavors to create 'autonomous' systems bear an immense risk of going hand in hand with losing human control over outputs (deep learning) and relinquishing human responsibility for outcomes. And this risk is independent of the term itself. In other words, a distinct definition of 'autonomy' for artefacts, measurable and potentially existing to degrees, obfuscates the fact that humans are creating technological instruments that may lose their instrumental character because we gradually give away responsibility for the outcomes of their usage.

Consequently, agreeing that 'autonomy' for artefacts is a term willingly borrowed from its 'classical' usage of personal self-governance, and intrinsically linked to responsibility, would shed a different light on the creation of autonomous artifacts and thus lead to a different question: Why are we aiming at limiting the space for responsible human action instead of increasing it? It is highly important not to lose oneself in technological definitions of 'autonomy'. 'Autonomy' for artefacts is a term that could function as an excuse for relinquished human responsibility for 'ugly' and potentially immoral outcomes, i.a., the killing of human beings in the case of LAWS.

4 Lethal Autonomous Weapons Systems (LAWS)

AT can supplant human beings in decision-making processes in certain areas. This can have an enormous potential for good (e.g. autonomously driving cars for visually impaired people,

[37] Arkin, Ronald, 2009, Ethical Robots in Warfare, IEEE Technology and Society Magazine 28(1), 30-33; Arkin, Ronald, 2010, The case for ethical autonomy in unmanned systems, Journal of Military Ethics 9(4), 332-341; Arkin, Ronald, 2017, A roboticist's perspective on lethal autonomous weapon systems, UNODA Occasional Papers No. 30, New York: United Nations, 35-37; Anderson, M., Anderson, S., and Armen, C., 2004, Towards Machine Ethics, AAAI-04 Workshop on Agent Organizations: Theory and Practice, San Jose, CA; Anderson, M., Anderson, S., and Berenz, V., 2016, Ensuring Ethical Behavior from Autonomous Systems, Proc. AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments, available at: http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12555 (accessed on February 4, 2018); Moor, J., 2006, The Nature, Importance, and Difficulty of Machine Ethics, IEEE Intelligent Systems, July/August, 18-21; McLaren, B., 2005, Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning, 2005 AAAI Fall Symposium on Machine Ethics, AAAI Technical Report FS-05-06.

[38] See 6.2.

surgical robots[39]). However, in addition to promising applications of AT, autonomous software can be (and arguably already is) integrated into robots that can select and engage a (military) target (e.g. infrastructure and potentially also combatants) without a human override.[40] Often-called Lethal Autonomous Weapons Systems (LAWS), as yet, there exists no agreed definition of LAWS. One reason for this lack of definition is that there exists, as highlighted above, no general understanding of the term 'autonomy' in AI and robotics.

The general idea is that a LAWS, once activated, would, with the help of sensors and computationally intense algorithms, identify, search, select, and attack targets without further human intervention. Whether the human being can still overpower or veto an autonomous weapon's 'decision' in order for it to be called a LAWS, is also debated.[41] However, military operational necessity precisely seems to require weapons systems that can function once human communication links break down.[42] Furthermore, state-of-the-art research on AI is currently creating software which can 'learn' entirely on its own[43] and even 'learn' to 'learn' on its own.[44] Hence, (precursor) technologies for creating fully 'human-out-of-the-loop'[45] weapons systems already exist.

From a military perspective, LAWS have many advantages over classical automated or remotely controlled systems: LAWS would not depend on communication links; they could operate at increased range for extended periods; fewer humans would be needed to support military

[39] See e.g., Strickland, Eliza, 2017, In Flesh-Cutting Task, Autonomous Robot Surgeon Beats Human Surgeons, IEEE Spectrum, October 13, 2017, available at: https://spectrum.ieee.org/the-human-os/biomedical/devices/in-fleshcutting-task-autonomous-robot-surgeon-beats-human-surgeons (accessed on February 1, 2018).

[40] The aim of this paper is not to provide examples or a list of existing weapons with autonomous capabilities. A continuously updated list can be found through e.g. Roff, Heather, and Moyes, Richard, 2016, Dataset: Survey of Autonomous Weapons Systems, Global Security Initiative: Autonomy, Robotics & Collective Systems, Arizona State University, available at: https://globalsecurity.asu.edu/robotics-autonomy (accessed on February 15, 2018).

[41] The US Department of Defense defines a weapons system as autonomous if it '[…] can select and engage targets without further intervention by a human operator.' Department of Defense, Directive 3000.09, November 21, 2012, 13-14; The UN Special Rapporteur on extrajudicial, summary or arbitrary executions adds the element of choice: 'The important element is that the robot has an autonomous "choice" regarding selection of a target and the use of lethal force.' Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christoph Heyns, UN doc. A/HRC/23/47, § 38; Human Rights Watch (HRW) distinguishes level of autonomy in weapons systems and contrasts the terms 'human-out-of-the-loop' and 'human-on-the-loop'. A 'human-out-of-the-loop' weapon is '[…] capable of selecting targets and delivering force without any human input or interaction […].' In other words, a 'human-out-of-the-loop' weapon's decision cannot be vetoed by a human being. On the other hand side, a 'human-on-the-loop' weapon can '[…] select targets and deliver force under the oversight of a human operator who can override the robots' actions […]'. According to HRW, both types can be considered 'fully autonomous weapons' when supervision is so limited that the weapon can be considered 'out-of- the-loop.' Docherty, B., 2012, Losing Humanity: The Case Against Killer Robots, Human Rights Watch, November 2012, available at: https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots (accessed on February 1, 2018); The ICRC defines autonomous weapons systems as '[…] (a)ny weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention.' ICRC, 2016, Convention on Certain Conventional Weapons, Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), April 11 – 15, 2016, Geneva, Switzerland, 1.

[42] Adams, T., 2002, Future Warfare and the Decline of Human Decision making, Parameters, U.S. Army War College Quarterly, Winter 2001-02, 57-71.

[43] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A, Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. and Hassabis, D., 2017, Mastering the game of Go without human knowledge, Nature vol. 550, 354-359.

[44] See e.g., Finn, Chelsea, 2017, Learning to Learn, Berkeley Artificial Intelligence Research, available at: http://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/ (accessed February 2, 2018).

[45] Docherty, B., 2012.

operations; their higher processing speeds would suit the increasing pace of combat;[46] by replacing human soldiers, they will spare lives; and with the absence of emotions such as self-interest, fear or vengeance, their 'objective' 'decision-making' could lead to overall outcomes that are less harmful.[47]

However, the use of LAWS may also generate substantial threats. Generally, LAWS may change how humans exercise control over the use of force and its consequences. Further, humans may no longer be able to predict who or what is made the target of an attack, or even explain why a particular target was chosen by a LAWS. This fact raises serious legal, ethical, humanitarian and security concerns.[48] From a humanitarian and ethical point of view, LAWS could be regarded as diminishing the value of human life as a machine and not a human being 'decides' to kill.[49] Also, the physical and emotional distance between the programmer or engineer of a LAWS and the targeted person may generate an indifference or even a 'Gameboy Mentality' on the side of the former.[50] From a security perspective, LAWS could be dangerous because they may also be imperfect and malfunction.[51] Moreover, the greater the technology advances, the more the level of autonomy of a LAWS increases. This, further, leads to an increased unpredictability of outcomes of LAWS and may enable the interaction of multiple LAWS as e.g. self-organizing swarms.[52]

The focus of scholarly inquiry of the legality of LAWS was mainly on IHL,[53] which presents significant challenges for both the development and the use of LAWS, since the latter would face problems to meet IHL's requirements of distinction,[54] proportionality[55] and precaution.[56] [57] Moreover, the nature of autonomy in a weapons system means that the lines of

[46] Thurnher, J., 2014, Examining Autonomous Weapons Systems from a Law of Armed Conflict Perspective, in: Nasu, H., and McLaughlin, R. (eds.), New Technologies and the Law of Armed Conflict, TMS Asser Press, 213-218.
[47] ICRC, 2011, International Humanitarian Law and the Challenges of Contemporary Armed Conflicts, Official Working Document of the 31st International Conference of the Red Cross and the Red Crescent, November 28 – December 1, 2011.
[48] Geneva Academy, 2017, Autonomous Weapons Systems: Legality under International Humanitarian Law and Human Rights, https://www.geneva-academy.ch/news/detail/48-autonomous-weapon-systems-legality-under-international-humanitarian-law-and-human-rights (accessed on February 2, 2018).
[49] UN Doc. A/HRC/23/47, § 109.
[50] Sassòli, Marco, 2014, Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified, International Law Studies Vol. 90, 308-340, 317.
[51] ICRC, 2014, Expert Meeting on 'Autonomous weapons systems: technical, military, legal and humanitarian aspects', March 26 – 28, 2014, Report of November 1, 2014, available at: https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014# (accessed on February 1, 2018).
[52] ICRC, 2016, 3.
[53] The reason for this legal focus on LAWS based almost exclusively on IHL is the fact that the UN CCW is underpinned by IHL, see also 3.1.5. This fact appears in an even odder light when considering that the first international thematic reference on autonomy in weapons systems was expressed by UN Special Rapporteur on Rapporteur on extrajudicial, summary or arbitrary executions, Christoph Heyns, in UN doc. A/HRC/23/47, § 38, for the Office of the High Commissioner for Human Rights.
[54] Art. 48, 49 51 (2) and 52 (2) Protocol Additional to the Geneva Conventions of August 12, 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), June 8, 1977.
[55] Art. 51 (5) (b) and Art. 57 Protocol I.
[56] Art. 57 (1) Protocol I.
[57] Brehm, Maya, 2017, Defending the boundary: Constraints and requirements on the use of autonomous weapons systems under international humanitarian and human rights law, Geneva Academy of International Humanitarian Law and Human Rights, 22; see also Bolton, M. 'From Minefields to Minespace: An Archeology of the Changing Architecture of Autonomous Killing in US Army Field Manuals on Landmines, Booby Traps and IEDs', 46 Political Geography (2015) 41–53.

responsibility for an attack by a LAWS may not always be clear. Therefore, LAWS also challenge the legal concept of accountability.[58]

Recently, LAWS have also been discussed in the light of International Human Rights Law (IHRL), whose benchmark for the legal use of force is higher than under IHL.[59] However, the emphasis on IHRL falls behind the strong focus on IHL within the UN CCW forum.

5 The debate at the United Nations Convention on Certain Conventional Weapons (UN CCW)

LAWS were taken up as an issue by the international arms control community in the framework of the UN CCW in 2014.[60] After a series of annual informal discussions, a Group of Governmental Experts (GGE) debated on the subject matter for the first time as a formal meeting during a 5-day-gathering in the CCW framework in Geneva in November 2017.

The main points of discussion of the GGE were LAWS's potential legality under IHL, questions of accountability and responsibility for the use of LAWS during armed conflict, potential (working) definitions of LAWS, as well as the need for emerging norms, since LAWS highly challenges both existing IHL as well as normative principles. However, this first GGE on LAWS brought no agreement on a political declaration and also no path toward a new regulatory international treaty. The only common denominator was the general will of states to continue conversations in 2018.[61]

The UN CCW's debate highlights at least five severe challenges to a comprehensive understanding of the risks of LAWS and AT.

(1) To date, states have not agreed on a definition of LAWS or the concept of autonomy, or on whether increasingly autonomous weapons systems, or precursor technologies, already exist. Moreover, national as well as international policy debates on LAWS have lacked precise terminology.[62]

Bearing in mind the above-described thoughts on the technological concept of 'autonomy', this is no surprise. However, it is claimed that definitions will most likely play a key role in the international deliberation on the issue of LAWS.[63] In order to comprehensively discuss, and reach agreement on, a topic, it is crucial to base the debate on a common understanding of the issue. It is also important in light

[58] See e.g. Davison, Neil, 2017, A legal perspective: Autonomous weapon systems under international humanitarian law, UNODA Occasional Papers No. 30, New York: United Nations, 12, 16.
[59] Brehm, 2017; Heyns, Christof, 2016, Human Rights and the use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement, Human Rights Quarterly 38, 350-378; Heyns, Christof, 2014, Autonomous Weapons Systems and Human Rights Law, Presentation made at the informal expert meeting organized by the state parties to the Convention on Certain Conventional Weapons, May 13-14, 2017, Geneva, Switzerland.
[60] CCW/MSP/2014/3.
[61] CCW/GGE.1/2017/CRP.1, 4, 5.
[62] Ibid., 13. See above on Autonomous Technology.
[63] Nakamitsu, Izumi, 2017, Foreword to the Perspectives on Lethal Autonomous Weapons Systems, UNODA Occasional Papers No. 30, New York: United Nations, V.

of the fact that there exists a not negligible movement of some states and NGOs to ban LAWS.[64]

However, since AT and the concept of 'autonomy' for technological artefacts may be a proxy term for an ongoing trend in human technological endeavours to give away control to technological agents thereby relinquishing human responsibility for outcomes of autonomous systems, a premature agreement on a definition of 'autonomy' in weapons systems by the GGE on LAWS would probably hide this trend. Therefore, instead of pressing for a definition of LAWS and 'autonomy' within the GGE, it would be advisable to locate these challenges within a bigger picture of the general relationship between humans and technology, and focus on the question whether we want to continue to regard technology as a controllable tool. In this sense, the GGE framework could be deemed as unfitting. Surely, principles for responsible AI research are both a first reflection of this underlying and ongoing paradigm change, as well as a first step in the direction of responsibly addressing the seriousness of this risk. A list of existing principles is found in the ANNEX of this paper.

(2) States are generally unwilling to share information on their capacity to develop LAWS. However, in order to gain a better understanding of the lessons learned from already existing weapons with certain levels of autonomy, the sharing of information is vital.[65]

(3) The GGE's mandate comprises the discussion of '[…] emerging technologies in the area of lethal autonomous weapons systems (LAWS) in the context of the objectives and the purposes of the convention […]'.[66] However, the misuse of technology, e.g., by non-state actors, does not fall within the scope of this mandate.[67] Certainly, though, a holistic analysis and discussion of the peace and security implications of AT and new technologies requires the international community to address also the use of such by non-state actors.[68]

---

[64] See the Campaign to Stop Killer Robots, https://www.stopkillerrobots.org/ (accessed on February 15, 2018). Currently, 22 states are backing this position: Algeria, Argentina, Bolivia, Brazil, Chile, Costa Rica, Cuba, Ecuador, Egypt, Ghana, Guatemala, Holy See, Iraq, Mexico, Nicaragua, Pakistan, Panama, Peru, State of Palestine, Uganda, Venezuela, Zimbabwe, Campaign to Stop Killer Robots, 2017, Country Views on Killer Robots, November 16, 2017.

[65] ICRC, 2016, Autonomous weapons systems: Profound implications for future warfare, May 6, 2016, available at: https://www.icrc.org/en/document/autonomous-weapons-systems-profound-implications-future-warfare (accessed on February 4, 2018).

[66] CCW/CONF.V/10,10.

[67] Ambassador Amandeep Singh, 2017 GGE on LAWS, Geneva, November 13-17, 2017, Plenary Session of November 14, 2017.

[68] See e.g., the attack on Russian military facilities by a swarm of more than a dozen autonomous drones. Russia accused Turkish-backed rebel forces to be behind the attack. See e.g. Satherley, Dan, 2018, Wooden drone swarm attacks Russian forces in Syria, Newshub.com, available at: http://www.newshub.co.nz/home/world/2018/01/wooden-drone-swarm-attacks-russian-forces-in-syria.html (accessed on February 4, 2018); Embury-Dennis, Tom, 2018, Russia says mysterious armed drones are attacking its military base in Syria – and they don't know who's sending them, January 10, 2018, Independent.co.uk, available at: http://www.independent.co.uk/news/world/middle-east/russia-military-bases-drones-syria-armed-attacks-tartus-uavs-latakia-a8151066.html (accessed on February 4, 2018); Focus, 2018, Mit schwer bewaffnetem Drohnenschwarm: Terroristen greifen russischen Stützpunkt an, January 14, 2018, Focus.de, available at: https://www.focus.de/politik/ausland/drohnenschwarm-is-griff-russischen-stuetzpunkt-an-nun-naehrt-sich-ein-besorgniserregender-verdacht_id_8296804.html (accessed on January 15, 2017).

(4) LAWS represent a new category of weapons, in that their novelty lies in a formless technological capacity to recognize patterns from a continuous inflow of data. The difference between a currently existing remotely controlled drone and a 'fully autonomous' drone does not lie in the casing, but in the fact that the latter is controlled by a software with autonomous capacities. The UN CCW, established in 1983, seeks to prohibit the use of certain conventional weapons. Its protocols currently prohibit the use of weapons whose primary effect is to injure by fragments that, once within the human body, escape X-Ray detection, as well as the use of mines, booby-traps and incendiary weapons against civilians.[69] One may argue that the UN CCW's GGE on LAWS is not capable to fully understand the technological complexity of current (not to mention future) AT.

(5) In addition, the CCW is a framework underpinned by IHL, which narrows the debate's focus on weapons and their use during *armed conflict*.[70] However, increasingly autonomous weapons systems can be and are used during peace time in law enforcement operations (e.g. crowd control, hostage situations), [71] where IHRL represents the legal benchmark.

Compared to IHL, IHRL is much more restrictive on the use of force. Military technology often finds its way into law enforcement. One may assume that once the advantages of increasingly autonomous systems have been proven in the military context, they might be considered for use during domestic law enforcement, although IHRL, regulating the latter, would prohibit their use.[72] Therefore, the CCW's/ GGE's approach could be criticized as not being legally comprehensive enough due to its limited focus on the use of a weapons during times of war.

6 Further ways to weaponize AT

The CCW's discussion on LAWS has focused on conventional (physical/ robotic) systems which interact in a 3D reality with other machines or humans.[73] However, there also exist additional ways to weaponize AT.

[69] Protocol I to Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects as amended on 21 December 2001 (CCW) on Non-Detectable Fragments, Protocol II to the CCW on Prohibitions or Restrictions on the Use of Mines, Booby Traps and Other Devices, and Protocol III to the CCW on Prohibitions or Restrictions on the Use of Incendiary Weapons.
[70] Art. 1 and 2 CCW.
[71] See e.g. Opall-Rome, Barbara, 2016, Introducing: Israeli 12-Kilo Killer Robot, DefenseNews.com, May 8, 2016, available at: https://www.defensenews.com/global/mideast-africa/2016/05/08/introducing-israeli-12-kilo-killer-robot/ (accessed on February 4, 2018); Hurst, Luke, 2015, Indian Police Buy Pepper Spraying Drones To Control 'Unruly Mobs', Newsweek.com, April 7, 2015, available at: http://www.newsweek.com/pepper-spraying-drones-control-unruly-mobs-say-police-india-320189 (accessed on February 4, 2018). The 'Mozzy Wildlife Darting Copter' is promoted for wildlife capture, Desert Wolf: Leaders in Technology and Innovation, available at: http://www.desert-wolf.com/dw/products/unmanned-aerial-systems/mozzy-wildlife-darting-copter.html (accessed on February 4, 2018).
[72] Heyns, Christof, 2016, Human Rights and the use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement, Human Rights Quarterly 38, 350-378; Heyns, Christof, 2014, Autonomous Weapons Systems and Human Rights Law, Presentation made at the informal expert meeting organized by the state parties to the Convention on Certain Conventional Weapons, May 13-14, 2017, Geneva, Switzerland.
[73] See also, UNIDIR, 2017, 1.

(1) First, software with autonomous capacities can be used to act and interact entirely in cyberspace. Those sometimes-called autonomous intelligent agents[74] are of tremendous military interest for 'conventional' military operations: Autonomous intelligent agents acting in cyberspace can support the decision-making process, they can identify an adversary's vulnerabilities and they can enable an ever-greater speed of response.[75] Hence, the use of autonomy for intangible cyber operations[76] (defensive or offensive) could be decisive and much more economic in current/future warfare.[77]

Five UN GGE discussions on cyber security have taken place since 2004/ 2005, and have confirmed that international law applies to the cyber space. Moreover, those GGEs have decided on a variety of confidence building measures (CBMs), and recommended norms for responsible State behavior in the domain.[78] Since both autonomous cyber weapons as well as LAWS are characterized by AT, both the GGE on LAWS as well as those on cyber security share the thematic technological basis. Nevertheless, those international policy discussions have nearly no overlap.[79]

(2) Second, it is highly necessary to consider potentially malicious linkages of AT and other emerging technologies. Theoretically, it may be possible to create autonomous systems that control processes with the core aim of harming humans, e.g. the malicious use of biotechnology, 5G radiation,[80] or products of molecular nanotechnology.[81] [82] Current examples of such linkages do not exist. However, it is crucial to raise this concern early enough in order to trigger both research in this field as well as a *comprehensive* debate of peace and security implications of both AT and other emerging technologies.

Recent technological breakthroughs in biotechnology highlight the risks of potential malicious linkages between different emerging technologies:

The term 'biotechnology' refers to '[…] any technological application that uses biological systems, living organisms, or derivatives thereof, to make or modify products

[74] Guarino, Alessandro, 2013, Autonomous Intelligent Agents in Cyber Offence, in: Podins, K., Stinissen, J., and Maybaum, M. (eds.), 5th International Conference on Cyber Conflict, NATO CCD COE Publications, 2013.
[75] Ibid, 4.
[76] There exists no standard terminology yet in this field.
[77] Meissner, Christopher, 2016, The Most Military Decisive Use of Autonomy You Won't See, DefenseOne, November 7, 2016, available at http://www.defenseone.com/ideas/2016/11/most-militarily-decisive-use-autonomy-you-wont-see-cyberspace-ops/132964/ (accessed on November 25, 2017). See e.g. the United States' cyberwarfare program MonsterMind. This software could constantly be on the lookout for traffic patterns indicating known or suspected cyberattacks. When it detected an attack, it would automatically block it from entering the country. This is regarded as a "kill" in cyber terminology. See e.g. Zetter, Kim, 2014, Meet Monstermind, The NSA Bot That Could Wage Cyberwar Autonomously, Wired, August 13, 2014, available at https://www.wired.com/2014/08/nsa-monstermind-cyberwarfare/ (accessed on November 28, 2017).
[78] UN Doc. A/70/174, 7-10.
[79] For a good discussion on the questions of interaction between the GGEs on LAWS and cyber space, see UNIDIR, 2017.
[80] For health risks of 5G radiation, see e.g. Puzzanghera, Jim, 2016, Is 5G technology dangerous? Early data shows a slight increase of tumors in male rats exposed to cellphone radiation, Los Angeles Times, August 8, 2016, available at: http://www.latimes.com/business/la-fi-cellphone-5g-health-20160808-snap-story.html (accessed on February 15, 2018).
[81] Helbing, 2018.
[82] For dangers of molecular nanotechnology and molecular manufacturing, see e.g. Dangers of Molecular Manufacturing, Center for Responsible Nanotechnology, http://www.crnano.org/dangers.htm (accessed on February 15, 2018).

or processes for specific use.'[83] One specific use of biotechnology is the creation of biological weapons. Biological weapons are designed to spread disease among people, animals and plants by introducing microorganisms and toxins, such as bacteria and viruses.

Using so-called DNA synthesis, which enables the artificial creation of DNA molecules, it may soon be possible to synthesize any virus whose DNA sequence is known.[84] Similarly, it is possible to insert small bacterial DNA fragments into another bacteria's DNA in order to increase its virulence, which would create a so-called 'binary biological weapon'.[85] Moreover, biotechnology could be used to manipulate cellular mechanisms to cause a disease. An agent could, e.g., be designed to induce cells to multiply uncontrollably, as in cancer, or induce programmed cell death (apotosis).[86] Further, in coming years it might be possible to design a pathogenic agent that targets a specific person's genome. When spread through a population that generally shows no or only minimal symptoms, it could nevertheless be fatal for the targeted person.[87]

Recently, scientists have been able to transform the four DNA nucleotid's letters into binary code, which now makes genetic engineering a matter of electronic manipulation and decreases the technique's cost.[88] Moreover, as of today, the European Nucleotide Archive of the European Bioinformatics Institute published sequences of 17075 genomes, including the genomes of 3316 bacteria and 4026 viruses.[89]

It is argued that biowarfare using genetically engineered pathogens can be considered as a potential revolution of military affairs.[90] Moreover, the exponential increase in computational power, the growing accessibility of genetic information and biological tools for the general public as well as the lack of governmental regulations also raise concerns about the non-state use of biowarfare.[91]

It is potentially possible to link AT and bioweapons, in that pathogens could be spread by autonomous systems.[92]

[83] Art. 2, Convention on Biological Diversity, of Rio de Janeiro of June 5, 1992.
[84] Hessel, A., Goodman, M., Kotler, S., 2012, Hacking the President's DNA, The Atlantic, available at http://www.theatlantic.com/magazine/archive/2012/11/hacking-the-presidents-dna/309147/?single_page=true (accessed on February 6, 2017).
[85] Ainscough, M., 2002, Next Generation Bioweapons: Genetic Engineering and Biowarfare, available at: http://www.au.af.mil/au/awc/awcgate/cpc-pubs/biostorm/ainscough.pdf (accessed on February 6, 2018), 256.
[86] Ibid., 273.
[87] Hessel et al, 2012.
[88] Ibid.
[89] European Bioinformatics Institute, Access to Completed Genomes, https://www.ebi.ac.uk/genomes/index.html (accessed on February 6, 2018).
[90] Ainscough, M., 2002.
[91] See e.g. Kay, D., 2003, Genetically Engineered Bioweapons, available at: https://www.aaas.org/sites/default/files/migrate/uploads/ch17.pdf, (accessed on February 6, 2018); Breakingnews.ie, 2011, Advances in Genetics Could Create Deadly Biological Weapons, Clinton Warns, July 7, 2011, available at: http://www.breakingnews.ie/world/advances-in-genetics-could-create-deadly-biological-weapons-clinton-warns-531347.html (accessed on February 6, 2018).
[92] Helbing, 2018.

(3) Moreover, based on the above-mentioned understanding of 'autonomy' for artefacts as any *result of a technological process for which the human cannot or does not want to bear responsibility,* any intentionally harmful use of a technology whose causes for harm cannot be traced back to a human 'trigger' may be deemed an autonomous weapon.[93]

7 Peace-time threats of not-weaponized AT

The risks of AT for society are not limited to its weaponized use during an armed conflict. AT also bears risks for global society during peace-time, when it is not weaponized.

7.1 Mass disinformation generated by intelligent technology

Both fake news (deliberate misinformation via traditional or online media with the intent to mislead the readers) and internet trolls (the posting of erroneous, extraneous and off-topic messages in order to manipulate public opinion) could potentially be generated by autonomous intelligent agents, which could lead to mass disinformation guided by AT. Not only news portals that deliberately and automatically spread fake information, but also social bots on twitter have an immense potential for mass manipulation. Moreover, bots that deceive us are currently already more numerous than those that tell us the truth, and they hardly cost anything.[94]

In addition to *general* mass manipulation through widely spread disinformation by bots, research on AI makes it possible to generate *individualized* information.[95] In this case, people do not share a common reference point for information anymore. The borders between reality and artificial creation with regards to knowledge through individual research would blur.

Further, AI research is able to create so-called 'generative adversarial networks' (GAN) that can currently generate fake images and videos whose quality is such that humans are incapable of telling that they are not real shots.[96] Moreover, it is said that GANs could soon generate speech, language and behavior.[97]

---

[93] Ibid.

[94] Laukenmann, Joachim, Der Nutzen von Lügenbots überwiegt: Interview mit Wirtschaftsinformatiker Oliver Bendel, #12 – Die Story des Tages, available at: https://mobile2.12app.ch/articles/29735653 (accessed on February 15, 2018).

[95] Cambridge Analytica has made lucrative use of those technological developments, see e.g. Hall, Jessica, 2017, Meet the weaponized propaganda that knows you better than yourself, Extremetech.com, March 1, 2017, accessible at: https://www.extremetech.com/extreme/245014-meet-sneaky-facebook-powered-propaganda-ai-might-just-know-better-know (accessed on February 15, 2018).

[96] See e.g., Leary, Kyree, 2017, An AI that makes fake videos may facilitate the end of reality as we know it, Futurism, December 8, 2017, available at: https://futurism.com/ai-makes-fake-videos-facilitate-end-reality-know-it/ (accessed on February 15, 2018).

[97] Karras, T., Aila, T., Laine, S., and Lehtinen, J., 2018, Progressive Growing of GANs for Improved Quality, Stability, and Variation, NVidia Rsearch, submitted to ICLR 2018, available at: http://research.nvidia.com/sites/default/files/publications/karras2017gan-paper-v2.pdf (accessed on February 3, 2018); Future of Life Institute, 2018, Podcast: Top AI Breakthroughs and Challenges of 2017 with Richard Mallah and Chelsea Finn, January 31, 2018, available at: https://futureoflife.org/2018/01/31/podcast-top-ai-breakthroughs-and-challenges-of-2017-with-richard-mallah-and-chelsea-finn/ (accessed on February 2, 2018).

With Adobe's application 'Project Voco' it is also possible to rapidly alter an existing voice recording to include words and phrases that the original speaker has never said.[98] One may assume that an altering of a recording by a machine or software instead of a human may soon be possible too. When real videos, images and voice recordings become indistinguishable from fake ones, fake news will become even more prevalent, and video, image, and voice evidence could become inadmissible in court.[99]

7.2 Autonomously generated profiles

Computerized pattern and correlation recognition in order to identify and represent people, for example during criminal investigations, could be performed by AT. The detection and capture of potential (pre-emptive profiling) and actual criminals could be outsourced to increasingly autonomous machine calculation based on Big Data – uncontrollable for humans. Already today, deep learning software allows for increasingly perfected facial recognition. Facial recognition technology is a computer application capable of identifying and verifying a person from a digital image or video. It is currently installed in public surveillance cameras, i.a., in Russia and China and used in order to continuously track potential criminals or public dissidents.[100]

Through increasingly autonomous criminal profiling the border between a criminal and a legally innocent person would be drawn exclusively by an algorithm, and vulnerable to incorrect data due to bad sensor-technologies, incompleteness or noise. Furthermore, categorizing potential criminals based on computational inferences somehow turns the presumption of innocence upside down, assuming a general potential for criminal conduct.[101]

Often, AI systems are claimed to be more 'objective' in their 'behavior' than humans, because they are not influenced by human feelings and prejudices. However, as 'intelligent' software and machines need to be 'fed' by a huge amount of data in order to 'learn' (a trait that we deem 'intelligent'), there exists the risk that they learn human prejudices from biased data. And so-called machine biases constitute a danger for AI-controlled or autonomous systems that some experts regard as far more acute than LAWS.[102] Based on the data a bot is fed in

[98] BBC News, 2016, Adobe Voco 'Photoshop-for-voice' causes concern, November 7, 2016, BBC News Technology, available at: http://www.bbc.com/news/technology-37899902 (accessed on February 15, 2018).
[99] Leary, 2017.
[100] See e.g. Chin, Josh, and Lin, Lisa, 2017, China's All-Seeing Surveillance State Is Reading Its Citizen's Faces, The Wall Street Journal, June 26, 2017, available at https://www.wsj.com/articles/the-all-seeing-surveillance-state-feared-in-the-west-is-a-reality-in-china-1498493020 (accessed on November 27, 2017); Fischer, Sophie-Charlotte, 2018, Künstliche Intelligenz: Chinas Hightech-Ambitionen, CSS Analysen zur Sicherheitspolitik 220, Zurich: CSS ETH Zurich, 4; Mezzofiore, Gianluca, 2017, Moscow's facial recognition CCTV network is the biggest example of surveillance society yet, Mashable, September 28, 2017, available at http://mashable.com/2017/09/28/moscow-facial-recognition-cctv-network-big-brother/#kF19SB72r8qA (accessed on November 27, 2017). See also the new Israeli business 'Faception', which provides real-time facial personality analytics and personal profiling also from offline datasets, https://www.faception.com/our-technology (accessed on February 15, 2018).
[101] Hildebrandt, Mireille, 2015, Smart Technologies and the End(s) of Law, Novel Entanglements of Law and Technology, Elgar Publishing, 97.
[102] Knight, Will, 2017b, Forget Killer Robots – Bias is the real AI danger, MIT Technology Review, October 3, 2017, available at: https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/ (accessed on February 15, 2018).

order to learn, it could learn, e.g., to discriminate people of color or minorities, or gain a strict political attitude.[103]

7.3 Autonomous technology in light of emerging resource-scarcity on our planet

The current global social, economic (including financial and monetary) and environmental trends render the planet's resources scarce. This constitutes a risk to humanity and makes our present global human coexistence potentially unsustainable. Hence, some experts ask the question: In an increasingly unsustainable society in critical times, what kind of citizens should be protected, and whose lives could be sacrificed? Should the worth of people's lives be weighed according to a certain benchmark, so that we can more easily decide who could stay alive? And does a human being have the guts to decide – or should we outsource this decision to autonomous software?[104]

For example, autonomous intelligent agents could be integrated into health insurance systems and feeding from patients' data, they could determine who receives a potential treatment and who does not. This may yet be a dystopian idea. However, the idea of a rating system for citizens is already tested in China with the so-called Citizen Score Card, which represents the value of an individual citizen from a governmental perspective. [105] A rating system like this could potentially become a reference point for informing decisions that aims at limiting population figures. [106]

The emergence of AT forces us to evaluate our current economic, social and environmental systems and trends, to ensure that we do not put society at risk of being kept in quantitative borders set by algorithms and based on utilitarian calculations.

---

[103] See e.g. 'Tay', a bot created by Microsoft who should learn from humans and turned into a Nazi within 24 hours, in: Steiner, Anna, 2016, Zum Nazi und Sexisten in 24 Stunden, Frankfurter Allgemeine, March 24, 2016, available at: http://www.faz.net/aktuell/wirtschaft/netzwirtschaft/microsofts-bot-tay-wird-durch-nutzer-zum-nazi-und-sexist-14144019.html (accessed on February 15, 2018); or Google's search algorithm that spread false information with a right wing bias, in: Solon, Olivia, and Levin, Sam, 2016, How Google's search algorithm spreads false information with a right wing bias, The Guardian, December 16, 2016, available at: https://www.theguardian.com/technology/2016/dec/16/google-autocomplete-rightwing-bias-algorithm-political-propaganda (accessed on February 15, 2018).

[104] Nagler, Jan, van den Hoven, Jeroen, and Helbing, Dirk, 2018, Ethics for Times of Crisis: How not to use autonomous systems in an unsustainable world, available at AARN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3112742 (accessed on February 16, 2018).

[105] Storm, Darlene, 2015, ACLU: Orwellian Citizen Score, China's credit score system, is a warning for Americans, Computerworld, October 7, 2015, available at https://www.computerworld.com/article/2990203/security/aclu-orwellian-citizen-score-chinas-credit-score-system-is-a-warning-for-americans.html (accessed on November 25, 2017); see also India's mandatory biometric ID system 'Aadhar': Pahwa, Nikhil, 2017, How not to screw up your national ID, Medianama, November 21, 2017, available at https://www.medianama.com/2017/11/223-how-not-to-screw-up-your-national-id-india-aadhaar/ (accessed on November 27, 2017) and the British 'Karma Police', a GCHQ program by the British government that creates personality profiles of British citizens, Brandom, Russel, 2015, British 'Karma Police' program carries out mass surveillance of the web, TheVerge.com, September 25, 2015, available at: https://www.theverge.com/2015/9/25/9397119/gchq-karma-police-web-surveillance (accessed on February 7, 2018).

[106] Helbing, Dirk, Nagler, Jan, and Van den Hoven, Jeroen, 2017, Ethics for Times of Crisis: How not to use autonomous systems in an unsustainable world, available at https://www.researchgate.net/publication/320740872_Ethics_for_Times_of_Crisis_How_not_to_use_autonomous_systems_in_an_unsustainable_world (accessed on November 25, 2017).

8 Arguments for shaping an international interdisciplinary debate

8.1 The polity of the cyberspace

Code can be regarded as the regulator of the cyberspace, the way a constitution can be regarded as a regulator of society.[107] Code enables the exchange of data among networks and is currently still generally neutral regarding the content of the data and ignorant about the user. This makes regulating behavior in the cyberspace difficult. However, code is not fixed, but the architecture of the cyberspace can be changed by the people who code. The fact that it is hard to know who someone is in the Net and what the character of the content is that is delivered, can be changed. New architecture can facilitate identification and rate data content. This architecture can either be privacy-enhancing or not. This depends on the incentives that are being faced by those who set it up.[108]

In other words, there exists a choice whether to influence the 'regulability' of the cyberspace as well as a choice about what this regulation should look like. Moreover, the way a constitution represents the normative values of a society through codifying them by law, code can be said to reflect a 'choice' of values that should guide actions and inactions in the cyberspace. If code represents the law of cyberspace, and computer software potentially interferes with citizens' privacy and maybe physical integrity (LAWS), should their use be restricted and regulated by a democratic process?[109]

This argument for a value-sensitive design of any software code that does or could interfere with citizen's privacy and physical integrity approved by a democratic political process would, as a first step, require a constant and very strong interaction between technological experts and both national and international policy-makers. Only thereby could the current policy discussions on technology lose their theoretical aspect and become more practical, which is crucial in order to potentially introduce the necessary aspects into a legislative process. Creating a fixed national and international policy-technology interface would require an architectural change of national and international political institutions, similar to the United Arab Emirates new state minister for AI.[110]

Furthermore, source codes of AT and AI-controlled systems need to be open source in order to be accessible for a political discussion and potential introduction into a legislative process. This condition will require deeply considered answers on the question of property rights of source codes of autonomous systems.

[107] Lessig, Lawrence, 2000, Code Is Law, On Liberty in Cyberspace, Harvard Magazine, January-February 2000, available at http://socialmachines.media.mit.edu/wp-content/uploads/sites/27/2015/03/Code-is-Law-Harvard-Magazine-Jan-Feb-2000.pdf (accessed on November 25, 2017).
[108] Ibid.
[109] Ibid., see also Van den Hoven, Jeroen, Vermaas, Home Pieter, and Van de Poel, Ibo (Eds.), 2015, Handbooks of Ethics, Values and Technological Design: Sources, Theory, Values and Application Domains, Springer.
[110] Galeon, Dom, 2017, An Inside Look at the First Nation With a State Minister for Artificial Intelligence, Futurism, December 11, 2017, available at: https://futurism.com/uae-minister-artificial-intelligence/ (accessed on February 16, 2018).

8.2 The subtle linguistics and the human-machine analogy

The international debate on AT and LAWS contains the unexamined assumption that humans and artificially intelligent systems are different only to a degree, and that human qualities can be reproduced in a machine. This underlying belief is the reason why the international debate uses anthropomorphic language – machine 'decision-making', machine 'learning', machine 'intelligence' or 'autonomy' – to describe current technological artefacts.

On this subject it is crucial to highlight two points: First, the human-machine analogy grew out of the initial wish and claim of AI research to understand the human brain by modelling it. However, this analogy still has a mere hypothetical character. Science could not yet fully reveal what happens in the human brain when, e.g., a decision is taken,[111] or how and if 'consciousness' can be linked to a physical process.[112]

And second, a software is usually named by its purpose, and not by its structure. If the purpose of, e.g., an 'autonomous' software is to supplant the human in an area where the latter used to take a *human* decision in no way implies that the software 'takes a decision' as well.[113] Hence, by comparing humans and machines or software at a common reference point (e.g. capacity to 'decide', 'learn' or 'behave morally') we may risk falling into a linguistic trap and prematurely overestimate technological artefacts and underestimate human capacities, let alone human language.

Language frames the way we think, understand and compare. Using the same language for machines and software as for humans could lead us to make potentially false comparisons – 'machines decide *better* than humans'.[114] Keeping in mind also the above-discussed risk for terminological confusion through the term 'autonomy' or 'intelligence', the question whether we need a new language for technological artefacts may be legitimate.

8.3 A moral argument for a sustainable environment

We are on the threshold of a paradigm shift where the human being will not be the only existing 'intelligent system' on the planet with the capacity for autonomous action anymore. Depending on the features that are encoded in increasingly autonomous systems and the existing risks of unpredictable outcomes[115] and vulnerabilities to hacking (e.g.), these systems may challenge the structure of current human society and might even become a risk for humanity as a species. Some experts also argue that organic human life is merely a short

[111] See e.g. Holdgraf, Chris, 2015, Decisions in the Brain, Berkeley Neuroscience News, June 15, 2015, available at: http://neuroscience.berkeley.edu/decisions-in-the-brain/ (accessed on December 9, 2017); Neue Zürcher Zeitung, Schaltzentrale Hirn,
[112] Kesser, Eduard, 2017, Das leer Gehirn, Neue Zürcher Zeitung, November 17, 2017, available at: https://www.nzz.ch/wissenschaft/das-leere-gehirn-ld.1329199 (accessed on February 16, 2018).
[113] See McDermott, Drew, 1981, Artificial Intelligence meets Natural Stupidity, in: Haugeland, John (ed.), Mind Design – Philosophy, Psychology, Artificial Intelligence, Cambridge MA: The MIT Press.
[114] Müller, Jürg, 2017, 'Oft entscheiden Menschen sehr schlecht', Neue Zürcher Zeitung, November 1, 2017, available at: https://www.nzz.ch/wirtschaft/oft-entscheiden-menschen-sehr-schlecht-ld.1325428 (accessed on February 16, 2018).
[115] Knight, Will, 2017a.

precursor in the evolutionary history of intelligent 'life' in the universe, which might soon be represented by inorganic machines with a far more powerful intellect than humans.[116]

Some are already preparing for a potential emergence of general AI through the enhancement of human brain power through AI itself: Elon Musk's recently launched company 'Neuralink' is exploring 'neural lace' technology – the implanting of tiny electrodes into the human brain to give us direct computing capabilities.[117] He argues that a '[…] merger of biological intelligence and machine intelligence […]' would be necessary for humans to stay economically valuable in a future of general AI.[118] Another way to keep up with AI and AT systems in a potential future world could also be paved by a genetic upgrade of humans through gene editing, which can nowadays already be used to alter the DNA of embryos.[119] In other words, research is focused on technology that would not only help us *do*, but that has the potential to help us *be*.[120]

A recent survey with the aim of clarifying expert opinions on the possibility and risks of human-like machine intelligence, based on 550 AI expert opinions, revealed a view among experts that AI systems will probably (over 50%) reach overall human ability by 2040-2050, and very likely (with 90% probability) by 2075. From reaching human-level-intelligence, experts assume that artificial superintelligence will be reached within 30 years after with a probability of 75%. Moreover, the respondents say that the probability that this development may be 'bad' or 'extremely bad' for humanity is 31%.[121]

Some experts claim that there exists a moral duty to pre-emptively decide not to create an invasive artificial species of autonomous agents that could endanger the lives of human beings on the planet.[122]

9 Conclusion

The international community should not get lost in attempts to define the term 'autonomy' for technological artefacts. Years of research and four years of discussions of LAWS within the UN CCW have not lead to terminological clarification, but opinions on the scope and content of the term 'autonomy' or AT have become more diverse.

---

[116] Rees, Martin, 2015, What do you think about machines that can think?, Edge.org, available at: https://www.edge.org/response-detail/26160 (accessed February 16, 2018).
[117] See https://www.neuralink.com/ (accessed on February 16, 2018).
[118] The Guardian, 2017, Elon Musk wants to connect brains to computers with new company, March 28, 2017, available at: https://www.theguardian.com/technology/2017/mar/28/elon-musk-merge-brains-computers-neuralink (accessed on February 16, 2018).
[119] Helbing, 2018; see also Levitt, Mairi, 2015, Would you edit your unborn child's genes so they were successful?, The Guardian, November 3, 2015, available at: https://www.theguardian.com/sustainable-business/2015/nov/03/designer-baby-pgd-would-you-edit-your-unborn-child-genes-more-successful (accessed on February 16, 2018); for a list of gene editing companies, see e.g. https://www.nanalyze.com/2015/04/7-gene-editing-companies-investors-should-watch/ (accessed on February 16, 2018).
[120] Prabhakar, Arati, 2017, The merging of humans and machines is happening now, Wired, January 27, 2017, available at: http://www.wired.co.uk/article/darpa-arati-prabhakar-humans-machines (accessed on February 17, 2018).
[121] Müller, Vincent C., and Bostrom, Nick, 2016.
[122] Helbing, Dirk, 2017, Open Discussion on Presentation on Lethal Autonomous Weapons Systems, November 13, 2017, ETH Zurich, Switzerland; see also Cellan-Jones, Rory, 2014, Stephen Hawkings warns artificial intelligence could end mankind, BBC Online, December 2, 2014, available at http://www.bbc.com/news/technology-30290540 (accessed on November 27, 2017).

If the term 'autonomy' for technological artefacts was defined to include a set of clearly delineated characteristics (e.g. 'learning', 'creating or pursuing of a goal', 'independent of human control, operation or intervention'), future technological research might reveal further potential characteristics which then would be excluded from this definition.

Yet a fixed definition of 'autonomy' for technological artefacts could lead to a clear definition of LAWS within the GGE. On the one hand, this could encourage a potential outcome of the UN discussions (e.g. Code of Conduct or norms for responsible State behaviour). On the other hand, again, new and yet unknown technological developments interesting for military use might be beyond the scope of this definition of LAWS. Hence, the pressure of defining 'autonomy' in order to proceed with the GGE debate would most possibly lead to a definition that reflects the current and maybe also conceivable future technological potentials. However, the exponential pace with which AI research advances must alert us to yet unknown potentials and risks.

Consequently, the endeavour to minimize risks of AI and AT must not focus on definitional questions regarding LAWS but concentrate on binding principles for responsible AI research. This alternative track would take into account the fact that 'autonomy' for technological artefacts, e.g. LAWS, can and should be regarded as a proxy term for the loss of human control and responsibility for outcomes of technological processes. Principles guiding AI research could require programmers and engineers only to develop technological artefacts whose outcomes will stay controllable for humans, and for which the latter would, hence, always bear responsibility. Initiatives of professional organizations as well as representatives of the private sector have led to several lists of principles for responsible/ ethical research on AI and autonomy (ANNEX). It would be advisable to bundle those principles and create an international body that would supervise compliance.

Consequently, an open discussion on whether or not humanity accepts the fact that technology is already crossing a threshold after which its creations might not be controllable for humans anymore, must be encouraged. Luckily, the UN CCW's debate on LAWS has brought this crucial moment into the public spotlight. Yet, for a purposeful discussion of this broader question, both the architecture of CCW forum as well as its limited mandate of LAWS are unsuitable.
As this paper has attempted to show, AT is much more than just its representation in LAWS. If we really want to look after the future of humanity, it is a prerequisite to gain a holistic understanding of all the peace and security implications of AT and emerging technologies.

Autonomous cyber weapons and autonomous weapons during law enforcement operations are excluded from the CCW discussion, yet they reflect the seriousness of the risk of weaponized AT to the same, or even to a higher degree, than LAWS. Hence, if the international community wants to prove its serious commitment to the issue of emerging technologies, the use of autonomous cyber weapons and autonomous weapons during law enforcement operations must be included in international discussions immediately.

Second, a holistic understanding of all the peace and security implications of AT must include peace-time threats of not-weaponized AT, such as mass dis- and misinformation as well as autonomous profiling and citizen control. A fixed body of experts at the UN level should take on committed discussions of peace and security implications of not-weaponized AT during peace-time.

Third, a holistic understanding of all peace and security implications of emerging technologies is necessary. This includes, i.a., AI, biotechnology, 5G radiation, and molecular nanotechnology. This paper had a limited focus on AT. However, threats for humanity stem from many more technological endeavours, whose risks are yet to be analysed. A fixed body of experts at the UN level should take on discussions of the peace and security implications of all emerging technologies.

Moreover, this paper has pointed out that the international debate on LAWS contains the unexamined assumption of a human-machine analogy. However, the view that human qualities can be reproduced in a machine should not be accepted unconditionally. As long as science cannot fully reveal the physical representation of human intelligence, consciousness, and decision-making processes in the human brain, self-protection should force us to acknowledge human distinctiveness. The fact that 'being human' is unquantifiable for science must not mean that human distinctiveness does not exist. We have a duty to preserve an assumption of this distinctiveness by limiting potential technologies that could challenge it or even wipe it out.

One way of preserving an understanding of the distinctiveness of 'being human' is by a careful use of language. Software or machine 'autonomy', 'intelligence' or 'agency' are terms that are very problematic in this sense. A premature heroization of technology could be prevented by introducing distinct terms. By using a term such as, e.g., 'artefact with *cognitive functions*' instead of 'intelligent agent', the fact that the machine is performing a *function* would be highlighted. This would set a clear boundary to being 'human and intelligent', as humans are not only performing a function, but are always an end unto themselves. Moreover, the term 'artefact' would point out its objective character as opposed to 'agent'.

In addition, this paper has challenged the view of the inevitability of AT and LAWS, which, unfortunately, reigns in the minds of some commentators.[123] We argued that the use of any software that could potentially interfere with a citizen's privacy or physical integrity could and should be regulated by a democratic process. This is a difficult demand or expectation. However, as the future may be far closer than we might think, it is highly important to start thinking and planning for this future today.

An introduction of software codes into a legislative process would require a creation of a constant policy-technology interface through, e.g., fixed state departments for technology/ AI. A constant dialogue between tech experts and policy-makers through an institutional

[123] See e.g. 'Warfare will continue and autonomous robots will ultimately be deployed in its conduct', Arkin, Ronald, 2009, Governing Lethal Behavior in Autonomous Robots, CRC Press, 29; or 'Autonomous weapons systems are the next logical and seemingly inevitable step in the continuing evolution of military technologies', Beard, Jack M., 2014, Autonomous Weapons and Human Responsibilities, Georgetown Journal of International Law 45, 617-681, 620.

integration could limit the risk that programmers and policy-makers could palm off the responsibility of 'autonomous' systems' 'immoral' outcomes to each other. Further, such an idea would require source codes to be publicly accessible, for which deeply considered answers to the question of property rights of source codes of autonomous and other systems are a prerequisite.

Humanity is striding into a future where machines and software will have an unprecedented role in almost all aspects of our lives. Moreover, future technology may have an immense potential for humans to define what they want to become. If we want to navigate wisely through a future that we might share with artefacts with cognitive abilities, we need to discuss some serious questions on 'autonomy', 'responsibility,' 'privacy' and 'identity'– and we have to do it now. This paper represents a small contribution to those profound challenges. More will be needed.

Based on this paper's conclusions, ICT4Peace would welcome

1. A creation of a UN level body for technology, with the tasks of ensuring responsible technological research and discussing peace and security implications of emerging technologies, i.a. AI and AT, biotechnology, 5G, molecular nanotechnology. This body would also set principles for responsible research in the above-mentioned scientific fields and ensure compliance.
2. An inclusion of autonomous cyber weapons and autonomous weapons during law enforcement into international discussions. The former could be integrated into the GGE on LAWS, and the latter could be taken up by the Human Rights Council.
3. A combined UN policy position on all the peace and security implications of emerging technologies.
4. A public discussion of the human-machine analogy, and a potential introduction of new terms, replacing 'AI' and 'autonomy'. Examples are 'artefact' instead of 'agent' or '…with cognitive functions/ capabilities' instead of 'intelligent'.
5. A creation of a constant national policy-technology interface through, e.g., fixed state ministers for technology/ AI.
6. An engaged debate on property rights on source codes of AI and AT software.
7. An increased engagement of civil society, including the private sector and academia, on the questions of human control of and responsibility for technological outcomes.
8. A constant dialogue between tech experts and civil society. Technologists must learn to transfer their expert knowledge in a practical way. This could be enhanced if courses were included in university curricula.

Regina Surber (reginasurber@ict4peacefoundation.org)
ICT4Peace Foundation
Zurich, February 21, 2018

10 ANNEX: Existing guidelines on responsible AI, AT and Robotics research

This annex contains six lists of guidelines for ethical/ responsible AI and AT research. Since the research field of robotics is highly linked to the research on AI and AT, and many endeavors have already lead to lists of principles in robotics, the annex also includes four lists of guidelines for ethical/ responsible robotics research.

A. Guidelines on responsible AI research:

1. FUTURE OF LIFE INSTITUTE (FLI)

The FLI is a volunteer-run research and outreach organization in the Boston area that works to mitigate existential risks facing humanity, particularly existential risk from advanced artificial intelligence (AI). Its founders include MIT cosmologist Max Tegmark, Skype co-founder Jaan Tallinn, and its board of advisors includes cosmologist Stephen Hawking and entrepreneur Elon Musk.

https://futureoflife.org/ai-principles/

'Asilomar AI Principles of 2017

Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.

Research Issues

1) **Research Goal:** The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
2) **Research Funding:** Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies, such as:
How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
How can we grow our prosperity through automation while maintaining people's resources and purpose?
How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
What set of values should AI be aligned with, and what legal and ethical status should it have?

3) **Science-Policy Link:** There should be constructive and healthy exchange between AI researchers and policy-makers.
4) **Research Culture:** A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
5) **Race Avoidance:** Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and Values

6) **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
7) **Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.
8) **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
9) **Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
10) **Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
11) **Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
12) **Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
13) **Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
14) **Shared Benefit:** AI technologies should benefit and empower as many people as possible.
15) **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
16) **Human Control:** Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
17) **Non-subversion:** The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
18) **AI Arms Race:** An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

19) **Capability Caution:** There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.
20) **Importance:** Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.
21) **Risks:** Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

22) **Recursive Self-Improvement:** AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.
23) **Common Good:** Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.'

1. Association for Computing Machinery (ACM)

ACM, the world's largest educational and scientific computing society, delivers resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources. The ACM Code of Ethics and Professional Conduct includes, i.a., four principles relating to ethical and responsible research. Due to its length, only those four are included in this annex.

https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct#CONTENTS

'ACM Code of Ethics and Professional Conduct Adopted by ACM Council 10/16/92. […]

General Moral Imperatives

As an ACM member I will ....

1.1 Contribute to society and human well-being.

This principle concerning the quality of life of all people affirms an obligation to protect fundamental human rights and to respect the diversity of all cultures. An essential aim of computing professionals is to minimize negative consequences of computing systems, including threats to health and safety. When designing or implementing systems, computing professionals must attempt to ensure that the products of their efforts will be used in socially responsible ways, will meet social needs, and will avoid harmful effects to health and welfare.
In addition to a safe social environment, human well-being includes a safe natural environment. Therefore, computing professionals who design and develop systems must be alert to, and make others aware of, any potential damage to the local or global environment.

Avoid harm to others.

"Harm" means injury or negative consequences, such as undesirable loss of information, loss of property, property damage, or unwanted environmental impacts. This principle prohibits use of computing technology in ways that result in harm to any of the following: users, the general public, employees, employers. Harmful actions include intentional destruction or modification of files and programs leading to serious loss of resources or unnecessary expenditure of human resources such as the time and effort required to purge systems of "computer viruses."

Well-intended actions, including those that accomplish assigned duties, may lead to harm unexpectedly. In such an event the responsible person or persons are obligated to undo or mitigate the negative consequences as much as possible. One way to avoid unintentional harm is to carefully consider potential impacts on all those affected by decisions made during design and implementation.

To minimize the possibility of indirectly harming others, computing professionals must minimize malfunctions by following generally accepted standards for system design and testing. Furthermore, it is often necessary to assess the social consequences of systems to project the likelihood of any serious harm to others. If system features are misrepresented to users, coworkers, or supervisors, the individual computing professional is responsible for any resulting injury. In the work environment the computing professional has the additional obligation to report any signs of system dangers that might result in serious personal or social damage. If one's superiors do not act to curtail or mitigate such dangers, it may be necessary to "blow the whistle" to help correct the problem or reduce the risk. However, capricious or misguided reporting of violations can, itself, be harmful. Before reporting violations, all relevant aspects of the incident must be thoroughly assessed. In particular, the assessment of risk and responsibility must be credible. It is suggested that advice be sought from other computing professionals. See principle 2.5 regarding thorough evaluations.
[...]

1.7 Respect the privacy of others

Computing and communication technology enables the collection and exchange of personal information on a scale unprecedented in the history of civilization. Thus there is increased potential for violating the privacy of individuals and groups. It is the responsibility of professionals to maintain the privacy and integrity of data describing individuals. This includes taking precautions to ensure the accuracy of data, as well as protecting it from unauthorized access or accidental disclosure to inappropriate individuals. Furthermore, procedures must be established to allow individuals to review their records and correct inaccuracies.
This imperative implies that only the necessary amount of personal information be collected in a system, that retention and disposal periods for that information be clearly defined and enforced, and that personal information gathered for a specific purpose not be used for other purposes without consent of the individual(s). These principles apply to electronic communications, including electronic mail, and prohibit procedures that capture or monitor electronic user data, including messages, without the permission of users or bona fide authorization related to system operation and maintenance. User data observed

during the normal duties of system operation and maintenance must be treated with strictest confidentiality, except in cases where it is evidence for the violation of law, organizational regulations, or this Code. In these cases, the nature or contents of that information must be disclosed only to proper authorities.
[…]

3.5 Articulate and support policies that protect the dignity of users and others affected by a computing system.

Designing or implementing systems that deliberately or inadvertently demean individuals or groups is ethically unacceptable. Computer professionals who are in decision making positions should verify that systems are designed and implemented to protect personal privacy and enhance personal dignity.
[…]'

2. INSTITUTE OF ELECTRIC AND ELECTRONICAL ENGINEERS (IEEE)

IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity. IEEE and its members inspire a global community to innovate for a better tomorrow through its more than 420,000 members in over 160 countries, and its highly cited publications, conferences, technology standards, and professional and educational activities. IEEE is the trusted "voice" for engineering, computing, and technology information around the globe.

The IEEE created the Global Initiative on Ethics of Autonomous and Intelligent Systems, an incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies. It aims at ensuring that every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity. The latest version of the book can be downloaded here: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

3. IBM

https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/

'Purpose: The purpose of AI and cognitive systems developed and applied by the IBM company is to augment human intelligence. Our technology, products, services and policies will be designed to enhance and extend human capability, expertise and potential. Our position is based not only on principle but also on science. Cognitive systems will not realistically attain consciousness or independent agency. Rather, they will increasingly be embedded in the processes, systems, products and services by which business and society function—all of which will and should remain within human control.

Transparency: For cognitive systems to fulfill their world- changing potential, it is vital that people have con dence in their recommendations, judgments and uses. Therefore, the IBM company will make clear:

When and for what purposes AI is being applied in the cognitive solutions we develop and deploy.
The major sources of data and expertise that inform the insights of cognitive solutions, as well as the methods used to train those systems and solutions.
The principle that clients own their own business models and intellectual property and that they can use AI and cognitive systems to enhance the advantages they have built, often through years of experience. We will work with our clients to protect their data and insights, and will encourage our clients, partners and industry colleagues to adopt similar practices.

Skills: The economic and societal bene ts of this new era will not be realized if the human side of the equation is not supported. This is uniquely important with cognitive technology, which augments human intelligence and expertise and works collaboratively with humans. Therefore, the IBM company will work to help students, workers and citizens acquire the skills and knowledge to engage safely, securely and effectively in a relationship with cognitive systems, and to perform the new kinds of work and jobs that will emerge in a cognitive economy.'

4. DeepMind

DeepMind has created the DeepMind Ethics & Society, a research unit that aims to explore the key ethical challenges facing the field of AI, through interdisciplinary work that brings together the technical insights of it DeepMind team and the diverse range of people who will be affected by it.

https://deepmind.com/applied/deepmind-ethics-society/principles/

'DeepMind Ethics & Society Principles

Social benefit: We believe AI should be developed in ways that serve the global social and environmental good, helping to build fairer and more equal societies. Our research will focus directly on ways in which AI can be used to improve people's lives, placing their rights and well-being at its very heart.
Rigorous and evidence-based: Our technical research has long conformed to the highest academic standards, and we're committed to maintaining these standards when studying the impact of AI on society. We will conduct intellectually rigorous, evidence-based research that explores the opportunities and challenges posed by these technologies. The academic tradition of peer review opens up research to critical feedback and is crucial for this kind of work.
Transparent and open: We will always be open about who we work with and what projects we fund. All of our research grants will be unrestricted and we will never attempt to influence or pre-determine the outcome of studies we commission. When we collaborate or co-publish with external researchers, we will disclose whether they have received funding from us. Any published academic papers produced by the Ethics & Society team will be made available through open access schemes.
Diverse and interdisciplinary: We will strive to involve the broadest possible range of voices in our work, bringing different disciplines together so as to include diverse viewpoints. We recognize that questions raised by AI extend well beyond the technical domain, and can only be answered if we make deliberate efforts to involve different sources of expertise and knowledge.
Collaborative and inclusive: We believe a technology that has the potential to impact all of society must be shaped by and accountable to all of society. We are therefore committed to supporting a range of public and academic dialogues about AI. By establishing ongoing collaboration between our researchers and the people affected by these new technologies, we seek to ensure that AI works for the benefit of all.'

5. MICROSOFT

https://www.microsoft.com/en-us/ai/our-approach-to-ai

'Microsoft AI Principles

Fairness: AI must maximize efficiencies without destroying dignity and guard against bias
Accountability: AI must have algorithmic accountability
Transparency: AI must be transparent
Ethics: AI must assist humanity and be designed for intelligent privacy'

B. Guidelines on responsible Robotics research:

1. Engineering and Physical Science Research Council (EPSRC)

EPSRC is the main UK government agency for funding research and training in engineering and the physical sciences, investing more than £800 million a year in a broad range of subjects - from mathematics to materials science, and from information technology to structural engineering. Its mission is to promote and support, by any means, high quality basic, strategic and applied research and related postgraduate training in engineering and the physical sciences; to advance knowledge and technology (including the promotion and support of the exploitation of research outcomes), and provide trained scientists and engineers, which meet the needs of users and beneficiaries (including the chemical, communications, construction, electrical, electronic, energy, engineering, information technology, pharmaceutical, process and other industries).

https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

'Principles of Robotics

Note: The rules are presented in a semi-legal version; a more loose, but easier to express, version that captures the sense for a non-specialist audience and a commentary of the issues being addressed and why the rule is important.

Principles for designers, builders and users of robots

	LEGAL	GENERAL AUDIENCE	COMMENTARY
1	Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.	Robots should not be designed as weapons, except for national security reasons.	Tools have more than one use. We allow guns to be designed which farmers use to kill pests and vermin but killing human beings with them (outside warfare) is clearly wrong. Knives can be used to spread butter or to stab people. In most societies, neither guns nor knives are banned but controls may be imposed if necessary (e.g. gun laws) to secure public safety. Robots also have multiple uses. Although a creative end-user could probably use any robot for violent ends, just as with a blunt instrument, we are saying that robots should never be **designed** solely or even principally, to be used as weapons with deadly or other offensive capability. This law, if adopted, limits the commercial capacities of

			robots, but we view it as an essential principle for their acceptance as safe in civil society.
2	Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.	Robots should be designed and operated to comply with existing law, including privacy.	We **can** make sure that robot actions are designed to obey the **laws** humans have made. There are two important points here. First, of course no one is likely deliberately set out to build a robot which breaks the law. But designers are not lawyers and need to be reminded that building robots which do their tasks as well as possible will sometimes need to be balanced against protective laws and accepted human rights standards. Privacy is a particularly difficult issue, which is why it is mentioned. For example, a robot used in the care of a vulnerable individual may well be usefully designed to collect information about that person 24/7 and transmit it to hospitals for medical purposes. But the benefit of this must be balanced against that person's right to privacy and to control their own life e.g. refusing treatment. Data collected should only be kept for a limited time; again the law puts certain safeguards in place. Robot designers have to think about how laws like these can be respected during the design process (e.g. by providing off-switches). Secondly, this law is designed to make it clear that robots are just tools, designed to achieve goals and desires that **humans** specify. Users and owners have responsibilities as well as designers and manufacturers. Sometimes it is up to designers to think ahead because robots may have the ability to learn and adapt their behaviour. But users may also make robots do things their designers did not foresee. Sometimes it is the owner's job to supervise the user (e.g. if a parent bought a robot to play with a child). But if a robot's actions do turn out to break the law, it will

			always be the responsibility, legal and moral, of one or more human beings, not of the robot (We consider how to find out who is responsible in law 5, below).
3	Robots are products. They should be designed using processes which assure their safety and security.	Robots are products: as with other products, they should be designed to be safe and secure.	Robots are simply not people. They are pieces of technology their owners may certainly want to protect (just as we have alarms for our houses and cars, and security guards for our factories) but we will always value human safety over that of machines. Our principle aim here, was to make sure that the safety and security of robots in society would be assured, so that people can trust and have confidence in them. This is not a new problem in technology. We already have rules and processes that guarantee that, e.g. household appliances and children's toys are safe to buy and use. There are well worked out existing consumer safety regimes to assure this: e.g. industry kite-marks, British and international standards, testing methodologies for software to make sure the bugs are out, etc. We are also aware that the public knows that software and computers can be "hacked" by outsiders, and processes also need to be developed to show that robots are secure as far as possible from such attacks. We think that such rules, standards and tests should be publicly adopted or developed for the robotics industry as soon as possible to assure the public that every safeguard has been taken before a robot is ever released to market. Such a process will also clarify for industry exactly what they have to do. This still leaves a debate open about how far those who own or operate robots should be allowed to protect them from e.g. theft or vandalism, say by built-in taser shocks. The group chose to delete a phrase that had ensured the right of manufacturers or owners to include "self defence" capability into a robot. In other

			words we do not think a robot should ever be “armed” to protect itself. This actually goes further than existing law, where the general question would be whether the owner of the appliance had committed a criminal act like assault without reasonable excuse.
4	Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.	Robots are manufactured artefacts: the illusion of emotions and intent should not be used to exploit vulnerable users.	One of the great promises of robotics is that robot toys may give pleasure, comfort and even a form of companionship to people who are not able to care for pets, whether due to rules of their homes, physical capacity, time or money. However, once a user becomes attached to such a toy, it would be possible for manufacturers to claim the robot has needs or desires that could unfairly cost the owners or their families more money. The legal version of this rule was designed to say that although it is permissible and even sometimes desirable for a robot to sometimes give the impression of real intelligence, anyone who owns or interacts with a robot should be able to find out what it really is and perhaps what it was really manufactured to do. Robot intelligence is artificial, and we thought that the best way to protect consumers was to remind them of that by guaranteeing a way for them to "lift the curtain" (to use the metaphor from The Wizard of Oz). This was the most difficult law to express clearly and we spent a great deal of time debating the phrasing used. Achieving it in practice will need still more thought. Should all robots have visible bar-codes or similar? Should the user or owner (e.g. a parent who buys a robot for a child) always be able to look up a database or register where the robot's functionality is specified? See also rule 5 below.
5	The person with legal responsibility for a robot should be attributed.	It should be possible to find out who is responsible for any robot.	In this rule we try to provide a practical framework for what all the rules above already implicitly depend on: a robot is never legally

			responsible for anything. It is a tool. If it malfunctions and causes damage, a human will be to blame. Finding out who the responsible person is may not however be easy. In the UK, a register of who is responsible for a car (the "registered keeper") is held by DVLA; by contrast no one needs to register as the official owner of a dog or cat. We felt the first model was more appropriate for robots, as there will be an interest not just to stop a robot whose actions are causing harm, but people affected may also wish to seek financial compensation from the person responsible. Responsibility might be practically addressed in a number of ways. For example, one way forward would be a licence and register (just as there is for cars) that records who is responsible for any robot. This might apply to all or only operate where that ownership is not obvious (e.g. for a robot that might roam outside a house or operate in a public institution such as a school or hospital). Alternately, every robot could be released with a searchable online licence which records the name of the designer /manufacturer and the responsible human who acquired it (such a licence could also specify the details we talked about in rule 4 above). There is clearly more debate and consultation required. Importantly, it should still remain possible for legal liability to be shared or transferred e.g. both designer and user might share fault where a robot malfunctions during use due to a mixture of design problems and user modifications. In such circumstances, legal rules already exist to allocate liability (although we might wish to clarify these, or require insurance). But a register would always allow an aggrieved person a place to start, by finding out who was, on first principles, responsible for the robot in question.

Seven High-Level Messages

In addition to the above principles the group also developed an overarching set of messages designed to encourage responsibility within the robotics research and industrial community, and thereby gain trust in the work it does. The spirit of responsible innovation is, for the most part, already out there but we felt it worthwhile to make this explicit. The following commentary explains the principles.

	PRINCIPLE	COMMENTARY
1	We believe robots have the potential to provide immense positive impact to society. We want to encourage responsible robot research.	This was originally the "0th" rule, which we came up with midway through. But we want to emphasize that the entire point of this exercise is positive, though some of the rules can be seen as negative, restricting or even fear-mongering. We think fear-mongering has already happened, and further that there are legitimate concerns about the use of robots. We think the work here is the best way to ensure the potential of robotics for all is realised while avoiding the pitfalls.
2	Bad practice hurts us all.	It's easy to overlook the work of people who seem determined to be extremist or irresponsible, but doing this could easily put us in the position that GM scientists are in now, where nothing they say in the press has any consequence. We need to engage with the public and take responsibility for our public image.
3	Addressing obvious public concerns will help us all make progress.	The previous note applies also to concerns raised by the general public and science fiction writers, not only our colleagues.
4	It is important to demonstrate that we, as roboticists, are committed to the best possible standards of practice.	as above
5	To understand the context and consequences of our research we should work with experts from	We should understand how others perceive our work, what the legal and social consequences of our work may be. We must figure out how to best integrate our robots

	other disciplines including: social sciences, law, philosophy and the arts.	into the social, legal and cultural framework of our society. We need to figure out how to engage in conversation about the real abilities of our research with people from a variety of cultural backgrounds who will be looking at our work with a wide range of assumptions, myths and narratives behind them.
6	We should consider the ethics of transparency: are there limits to what should be openly available	This point was illustrated by an interesting discussion about open-source software and operating systems in the context where the systems that can exploit this software have the additional capacities that robots have. What do you get when you give "script kiddies" robots? We were all very much in favour of the open source movement, but we think we should get help thinking about this particular issue and the broader issues around open science generally.
7	When we see erroneous accounts in the press, we commit to take the time to contact the reporting journalists.	Many people are frustrated when they see outrageous claims in the press. But in fact science reporters do not really want to be made fools of, and in general such claims can be corrected and sources discredited by a quiet & simple word to the reporters on the byline. A campaign like this was already run successfully once in the late 1990s.

2. RoboLaw Project

The main goal of the RoboLaw project is to achieve a comprehensive study of the various facets of robotics and law and lay the groundwork for a framework of "Robolaw" in Europe. The RoboLaw project aims at understanding the legal and ethical implications of emerging robotic technologies and of uncovering (1) whether existing legal frameworks are adequate and workable in light of the advent and rapid proliferation of robotics technologies, and (2) in which ways developments in the field of robotics affect norms, values and social processes we hold dear.

The problem of regulating new technologies has been tackled in Europe almost by every legal system: Therefore, it is possible to rely on a background which includes a large amount of studies on the relationship between law and science and between law and technology. Nevertheless, the RoboLaw project is focused on the extreme frontiers of technological advance, confronting the legal "status" of robotics, nanotechnologies, neuroprostheses, brain-computer interfaces, areas in which very little work has been done so far.

This project is the first in-depth investigation into the requirements and regulatory framework(s) of "robolaw" in the age of the actualization of advanced robotics, and the first study to combine the many different legal themes that have been investigated in isolation before. Moreover, it is the first research to delve into the legal and ethical consequences of developments in robotics within specific legal systems within the EU and to compare these with the US and the Far East, Japan in particular.

The complete book of **Guidelines on Regulating Robotics** can be found here: http://www.robolaw.eu/

3. European Robotics Research Network (EURON)

EURON is a network of excellence in robotics, that is aimed at coordination and promotion of robotics research in Europe. The network is sponsored by the European Commission through the Future and Emerging Technologies Programme.

Its **Roboethics Roadmap** can be found here:
http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf

4. European Civil Law Rules in Robotics

The European Parliament's Legal Affairs Committee commissioned this study to evaluate and analyse, from a legal and ethical perspective, a number of future European civil law rules in robotics.

The full report can be found here: http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf