

GUGLIELMO TAMBURRINI

EMERGING GLOBAL ISSUES IN AI ETHICS

1. Introduction

There is hardly any aspect of human life that artificial intelligence (AI) has not changed or may not change in the near future. The pervasive impact of AI over the last decade has been powered by machine learning (ML) technologies in general, and deep neural networks (DNN) in particular. AI learning systems are having an increasing role to play in finance, commerce, industry, the management of public and private services, communication and entertainment, security and defense.

AI ethics has been mostly concerned with the fundamental rights and the good of groups of people selectively affected by the operation of AI systems. Exemplary cases of this prevalent approach concern automatic decision-making, classifications, and predictions performed by AI systems in a wide variety of application domains, including access to education, loans, career development, insurance and other vital public or private services. This approach is characterized by a piecemeal analysis of tasks performed by AI systems and their ethical implications. The significance of this approach in AI ethics is unquestionable. However, the growing pervasiveness and networking of AI systems within and across application domains impels one to broaden the AI ethics agenda. Indeed, ethical issues are emerging that have a more global reach, insofar as they affect the fundamental rights and the good all human beings at once. A significant case in point concerns the double-edged role that AI is having in the climate crisis. On the one hand, AI helps one to identify effective interventions to reduce greenhouse gases emissions. On the other hand, however, there is growing apprehension about the carbon footprint of large AI models obtained by means of ML techniques, and their potential impact on climate warming. The latter affects the fundamental rights and the good of present and future generations of human beings, thereby raising genuinely global ethical issues. AI-powered autonomous weapons systems, including autonomous cyber weapons, are giving rise another global ethical issue, concerning worldwide peace and related cyber threats to nuclear defense systems.

This article analyzes the growing AI impact on both global ethical issues. Section 2 introduces the distinction between local and global ethical issues. Section 3 recalls some present efforts to achieve a better understanding of the environmental impact of AI research. Section 4 examines distinctive environmental responsibilities of the AI research community. Section 5 explores specific vulnerabilities affecting AI systems and corresponding epistemic responsibilities of AI scientists in connection with AI-enabled warfare. Section 6 highlights the local ethical issues so far dominating ethical and legal debates about AI-powered autonomous weapons systems. The final section 7 examines threats to international peace and stability posed by the maturing of autonomous cyber weapons that are powered by AI technologies, focusing on existential threats arising from the possibility of autonomous cyber weapons attacking nuclear command, control, and communication systems, and concluding on the moral responsibilities of AI scientists facing the development of these cyberweapons.

2. Local and global ethical issues

The EU proposal for regulating AI (European Commission, 2021) classifies a wide range of AI application domains as ‘high risk’ from ethical and legal standpoints. These application domains notably include access to education, vocational training and employment, the management of migration, asylum and border control, access to essential services and public benefits. Most of these ethical issues are *local*, in the sense that an AI system operating in one of these domains affects, at each given point in time, the good and the fundamental rights of some proper subset of the human population. Thus, for example, information processing biases present in an AI system supporting college admissions procedures may lead to discriminations affecting the good and the rights of rejected college applicants (O’Neil, 2017). Similarly, AI decision-making concerning bank loans, job hirings, career advancement, migration and asylum management, access to unemployment benefits, and other public or private services affect a proper subset of the humankind at each given point in time.

In contrast with this, here are called *global* those ethical issues involving at some point in time the good and the rights of all members of the human species. Pandemic infections like SARS-CoV-2 raise – in addition to formidable medical and economic problems – genuinely global ethical issues. These concern the physical integrity, life, well-being, right to work and education of all members of the human species. By the same token,

global ethical issues are presently at stake in connection with both the anthropically induced climate crisis (IPCC, 2021) and the very existence of nuclear arsenals (Cerutti 2008). The history of the humankind is scattered with the waxing and waning of other global ethical issues. The Spanish flu pandemic posed a global ethical issue back in 1918-19. The Rowland-Molina hypothesis advanced in 1976 identified the major cause of atmospheric ozone layer depletion in the use of chlorofluorocarbons (CFCs). This anthropically induced effect might have deprived the humankind and other living entities of protection from exposure to solar UV radiation, thereby triggering another potentially global ethical issue. Effective international efforts to decrease the use of chlorofluorocarbons (CFCs) from the 1980s onward appear to have successfully curbed this specific global threat.

Scarce attention has been given so far in AI ethics to the increasing impact of AI on global ethical issues, as one can readily glean from the outstanding state-of art survey of the field in the entry *Ethics of AI and robotics* (Müller, 2021) of the Stanford Encyclopedia of Philosophy. In contrast with this, the focus is placed here on the growing impact of AI on two ethical issues that have a global reach, insofar as they affect the fundamental rights and the good all human beings at once: (i) AI and the climate crisis; (ii) threats to worldwide peace emerging from AI-powered autonomous weapons systems, notably including autonomous cyber threats to nuclear defense systems.

To begin with, let us consider the increasing role that AI is playing in connection with global ethical issues raised by the climate crisis.

3. *The AI carbon footprint*

The 2020 *White Paper on AI* released by the European Commission recommends achieving a sufficiently detailed knowledge of the AI carbon footprint and its sources: “Given the increasing importance of AI, the environmental impact of AI systems needs to be duly considered throughout their lifecycle and across the entire supply chain, e.g., as regards resource usage for the training of algorithms and the storage of data.” (EU, 2020, p. 2). This knowledge is needed to guide effective countervailing actions and to allocate related responsibilities – including prospective responsibilities for actions to undertake, and retrospective responsibilities for significant omissions. Achieving the required knowledge, however, presupposes the solution of a variety of conceptual and empirical problems that are now capturing widespread attention within the AI research community.

Discussion of the AI carbon footprint have been stimulated by estimates of computational and electricity resources required to train AI models by ML methods. Strubell and co-workers focused on AI models for natural language processing (NLP) considering, in addition to the training of a variety of off-the-shelf AI models, downstream training processes that one needs to adapt and fine-tuning AI models to perform a variety of new NLP tasks (Strubell et al., 2019). Then, electricity consumption and emissions of greenhouse gases (GHG) were extrapolated from an estimate of computational training costs. In particular, training an NLP Transformer model, based on a DNN architecture, was estimated to produce GHG emissions equivalent to those of 5 average automobiles throughout their lifecycle. And the GHG emissions of BERT_{LARGE} trained using GPU (Graphic Processing Units) as specialized hardware accelerator were estimated to be equivalent to those of a commercial flight between San Francisco and New York (Strubell et al., 2019).

Patterson and co-workers provided reduced figures for the carbon footprint of these AI models. At the same time, however, they emphasized that major digital firms – unlike AI academic research focusing on model training – attribute to post-training phases up to 90% of computational and energy consumption costs of their AI models. Accordingly, to evaluate more accurately the carbon footprint of AI learning models, one must attend to their post-training operational uses for inference, prediction, and decision-making (Patterson et al., 2020).

Another source of environmental concern is the steadily growing size of both AI learning models and training datasets. The size of a DNN is usually measured by reference to the total number of weights associated to the connections between its neural nodes. In the NLP domain, the large version of the BERT model mentioned above contains about 350 millions of these parameters, the system GPT-2 about 1.5 billion, and 175 billion its more recently released successor GPT-3. This trend is explained by the improved performances achieved by larger models and the availability of computing resources. These improved performances, however, come with increased energy consumption associated to the use of larger models and their training. Similar trends are observed in connection with the increasing size of training sets and the massive recourse to hyperparameter experiments to compare and improve performances. Hyperparameter experiments explore how the performances of learning systems change by tuning the number of learning cycles, the network learning rate or other hyperparameters (Schwartz et al., 2020, p. 58).

Additional factors to consider in credible estimates of the AI carbon footprint include the energetic efficiency of datacenters and their energy

supply mix, the computational costs of algorithms and programs used for AI model training and inference, the energetic efficiency of the processors on which these programs run (Patterson et al., 2020, p. 2). Accordingly, “a concerted effort by industry and academia” is invoked by many to attain substantive reductions of the AI environmental cost (Strubell et al., p. 5). This concerted effort requires the participation of both individual and institutional actors, including researchers and engineers, universities, and AI firms. However, one should be careful to note that an environmental instance of the many-hands problem is likely to arise in these circumstances. As many different actors contribute to the overall AI carbon footprint, it becomes difficult – both in practice and in principle – to set apart which responsibilities pertain to which individual actor. The insurgence of a many-hands problem affords an excusing circumstance to each involved actor, often resulting into ineffective calls for concerted efforts of such breadth. Political discussion about climate warming mitigation actions has often floundered in similar buck-passing games.

4. Prospective responsibilities of AI scientists

To prevent the insurgence of a many-hands problem in connection with the AI carbon footprint, one may try and pin down roles and responsibilities for each one of the involved actors. These include both prospective responsibilities for GHG reduction actions and retrospective responsibilities for corresponding omissions. What are the distinctive roles and responsibilities of AI researchers? How are these disentangled from roles and responsibilities of other involved actors, including private and public data-centers, providers of cloud computing resources, electricity producers, and commercial firms relying on already trained AI models for inference and prediction? How are climate warming mitigation efforts by AI researchers fairly compared to each other within the AI community?

It is not easy to provide sensible answers to these questions, insofar as AI researchers tap from a variety of energy supply sources, software systems, and hardware resources to develop or adapt their AI learning models. To begin with, CHG emissions of AI research “results” are neither temporally nor spatially invariant, for the relative proportions of fossil fuel, alternative and renewable sources in the electricity supply mix depend on both location and time. Since direct estimates of CHG emissions are unsuitable to draw fair comparisons between the carbon footprints of different AI researchers working asynchronously and in distinct locations, one may try

and use electricity consumption estimates which are blind to the energy supply mix. But these are sensitive to the kinds of processors and, more generally, of hardware resources that scientists use to train AI learning models. Thus, not even gross electricity consumption is suitable to identify and compare fairly the efforts that AI researchers make or ought to make toward the reduction of CHG emissions.

To overcome these pitfalls, it has been suggested that a suitable metrics should concentrate on AI research computational efficiency. For example, Schwartz and co-workers proposed AI researchers to report “the total number of floating-point operations (FPO) required to generate a result”, on the ground that FPO estimates of the amount of work performed by a computational process are agnostic with respect to both hardware and energy supply infrastructures (Schwarz et al., 2020, p. 60). More in general, it is claimed that sensible measures of computational efficiency to correlate to the AI carbon footprint will enable the AI research community to take on distinctive responsibilities for climate warming mitigation actions based on the reduction of current GHG emissions.

Additional actions that AI researchers may undertake depend on what datacenters administrators, cloud computing providers, electricity producers and other involved actors do in the way of climate warming mitigation initiatives. These include choosing energy-efficient computing hardware, training AI models in low carbon intensity regions, choosing suitable training times, as the carbon intensity of any given region may change throughout the day (Anthony et al. 2020).

Sensible efficiency measures for AI results provide a basis to carry cost-benefit analyses considering climate warming concerns. Software tools are being made available to predict and estimate computational and energy efficiency costs of AI results (Anthony et al. 2020, Henderson et al, 2020, Lacoste et al., 2019). One may encourage environmentally thrifter research by prizing results which combine better accuracy with greater energy and computational efficiency. One may set up leaderboards and isolate environmentally significant benchmarks (Henderson et al, 2020). These ingredients are sufficient to spur AI sustainable research competitions, which may eventually achieve the reputation of other AI research programmes taking the form of competitions, such as the Grand Challenges promoted by DARPA, Robocup, computational chess, Go, and Poker (Tamburrini and Altiero, 2021). Promoting computationally thrifter AI competitions may attract the interest of industrial actors too, in view of economic advantages flowing from the reduction of electricity costs.

How distant is the prospective scenario of an environmentally virtuous AI research – promoted by nudging interventions, research awards and recognitions – from the current reality of AI research? The search for computationally efficient solutions to research problems is not a prevailing goal in today's ML research. This much can be gleaned from a random sample of 60 papers presented at top-level AI conferences: a large majority of these papers were found to target model accuracy only, without taking computational efficiency into account (Schwartz et al. 2020, p. 56). In another random sample of 100 papers from the 2019 NeurIPS proceedings, one paper only “measured energy in some way, 45 measured runtime in some way, 46 provided the hardware used, 17 provided some measure of computational complexity (e.g., compute-time, FPOs, parameters), and 0 provided carbon metrics.” (Henderson 2020, p. 6). These findings suggest that improvements in task performance accuracy – including the large amount of minor, practically inconsequential accuracy results that fail to make it into major AI conferences – are pursued without taking notice of environmental costs. Accordingly, a significant departure from prevalent research goals and standards is required to turn environmentally virtuous competitions into a widespread practice in the AI community.

One may doubt that environmentally sensitive research goals will gain sufficient traction within AI research communities on a temporal scale which is meaningfully related to the goal of limiting global warming in the XXI century to 1.5° C (IPCC, 2018). Assuming that each community must do its share to implement this goal, one may consider mandatory environmental policies for AI research as an alternative to, say, nudging interventions and voluntary participation into environmentally virtuous AI competitions. For example, one may consider introducing AI-research carbon quotas, in the absence of a swift and widespread endorsement of climate warming mitigation actions by the AI research community. Clearly, this policy would have a negative impact on research freedom, raising the additional ethical challenge of an equitable distribution of bounded computing resources among AI stakeholders (Lucivero, 2019).

In concluding this section, let us take a glance at AI's potential support to climate warming mitigation actions in *other* sectors. AI learning systems may substantively contribute to identify approximate solutions to a wide variety of energy use optimization problems, ranging from the energetic efficiency of buildings to the reduction of transportation needs, the planning of travel routes, and the efficiency of supply chains in industry production and food consumption (Rolnick et al., 2019). But what is the environmental

impact of the very AI models serving these energy efficiency goals? Over and above training and inference, one must consider energy consumption of sensor networks and other connected cyberphysical systems. Even more important, one must consider rebound effects. Jevon's paradox (Alcott, 2005) suggests that interventions successfully improving the efficiency of some given resource often end up encouraging more extended uses of that resource. This outcome eventually defeats the initial motivation of reducing consumption by improving efficiency. Suitable norms and incentives may be needed to avoid the positive feedback loop leading from improved efficiency to increased consumption.

To sum up. Applications of AI research promise to drive climate warming mitigation actions across a variety of economic and social activities. At the same time, however, AI research is an integral part of the climate crisis problem. This much is conveyed by recent – admittedly perfectible – estimates of the AI carbon footprint. Suitable measures of computational costs arising from AI research are needed to foster a better understanding of AI's environmental impact and the distinctive environmental responsibilities of the AI research community. AI competitions prizing computational efficiency and the establishment of leaderboards may encourage environmentally virtuous research attitudes by the AI research community. But the need for mandatory policies may emerge too, if the prevailing goal of prizing accuracy in current AI research will not be effectively and timely replaced by more comprehensive goals combining accuracy with energy and computational efficiency.

Let us now turn to consider AI-enabled warfare, related epistemic responsibilities of AI scientists, and the gradual emergence of a global ethical issue from current developments in AI-powered autonomous weapons systems.

5. AI-enabled warfare and the epistemic responsibilities of AI scientists

AI is expected to have a pervasive and disruptive effect on warfare: "AI-enabled warfare will not hinge on a single new weapon, technology, or operational concept; rather, it will center on the application and integration of AI-enabled technologies into every facet of warfighting. AI will transform the way war is conducted in every domain from undersea to outer space, as well as in cyberspace and along the electromagnetic spectrum. It will impact strategic decision-making, operational concepts and planning, tactical maneuvers in the field, and back-office support"

(NSCAI 2021, p. 81). In connection with this sweeping vision, AI scientists and engineers carry special responsibilities – distinct from those of political and military decision-makers or voters in democratic countries – which flow from their epistemically privileged position. AI scientists and engineers have the duty to raise awareness in the public opinion and decision-makers about limitations and vulnerabilities of AI systems that are relevant in warfare contexts.

Relevant limitations and vulnerabilities that are specific of AI systems have emerged from adversarial machine learning experiments. Notably, visual perceptual systems based on DNN architectures were found by adversarial testing to mistake images of school buses for ostriches (Szegedy et al., 2013) and 3-D renderings of turtles for rifles (Athalye et al., 2018). A human operator would not incur in such mistakes, for the small adversarial input perturbations inducing the machine to err are undetectable to the human perceptual system. Accordingly, visual recognition systems developed by means of state-of-art machine-learning technologies undergo unexpected and counterintuitive mistakes that a human operator may easily detect and avoid.

Fragilities of AI systems are potentially catastrophic in warfare. Indeed, normal uses of school buses are protected by the principles of just war theory (Walzer, 1978) and International Humanitarian Law (IHL); someone carrying a harmless object in the hand may be mistakenly taken to be wielding a weapon, thereby triggering an unjustified use of force. One can exclude neither the occurrence of similar malfunctioning during warfare operations nor intentional exploitations of AI systems sensitivity to small input perturbations to induce information-processing errors. Another form of intentional attack to AWS involves the *poisoning* of its AI learning modules, by corrupting data sets, the learning algorithm or even the resulting AI model. There are no patches available to avoid once and for all either input manipulation or poisoning attacks, insofar as these are based on inherent features of the ML methods and systems that are prevalent in current AI research (Comiter 2019).

Another problematic aspect of AI systems powered by ML methods is their black-box nature. Sub-symbolic data representations and information processing characterizing DNN models are opaque to human users. However, AI systems are not celestial oracles. Their outcomes should not be taken at face value in ethically sensitive domains. If humans in command-and-control positions are expected not to blindly trust machine decisions or recommendations – from strategic decision-making and planning to battlefield maneuvers – they should get a sufficient amount

of humanly understandable information about machine data processing. To fulfil this *interpretability* condition, machine data and information processing must be mapped into perceptual and cognitive domains that humans can make sense of.

Fulfilling the interpretability condition is necessary but largely insufficient for commanders to achieve adequate situational awareness. For their strategic and tactical decision-making, commanders should additionally obtain an account of *why* the AI system is suggesting or going to take a certain course of action. And to fulfil this more properly *explainability* condition, AI systems should be endowed with the capability to provide humanly understandable explanations of courses of actions that they are suggesting or undertaking. Meeting the explainability condition presupposes a conceptual clarification of what it is to provide correct and humanly understandable explanations of AI information-processing results. Moreover, the actual development of AI systems that are capable providing explanations raises formidable research problems, which now characterize the goals of the eXplainable AI (or XAI in brief) research area (<https://www.darpa.mil/program/explainable-artificial-intelligence>). Pending significant breakthroughs in XAI, one cannot but acknowledge the difficulty of fulfilling the system interpretability and explainability conditions that are necessary for a commander to achieve the required situational awareness.

A weakness of AI systems emerging from the newly developing research area of *adversarial* XAI casts additional doubts about the delivery of decisive solutions to the situational awareness problem. Indeed, adversarial XAI techniques enable one to attack explanation modules of AI learning systems, perturbing their inputs in ways that are not perceptible to humans, and yet radically altering the interpretations and the explanations provided by the machine (Ghorbani et al. 2019, Heo et al. 2019). Unreliable XAI interpretations and explanations jeopardize the achievement of the required situational awareness. More important, if adversarial XAI techniques become systematically exploitable by enemies, iterative runs of the competition between XAI and adversarial XAI will feed an AI arms race spiral: red teams of AI engineers may resort to adversarial XAI activities to patch identified weaknesses of AI explanation modules; enemies may respond by developing more powerful adversarial XAI techniques, thereby posing new challenges to the more robust AI systems developed by red team engineers; and so on.

Let us now turn to consider these and other sources of concern about AI-powered warfare in connection with the development, deployment and use of autonomous weapons systems (AWS).

6. *The AWS normative debate: local ethical issues*

The US Department of Defense identifies the more distinctive traits of AWS in their capability to select and engage targets without further intervention by a human operator after their activation (DoD 2012). Using slightly different wordings – but embracing the same functional approach to the autonomy of weapons systems – the International Committee of the Red Cross (ICRC) describes AWS as “weapons that can independently select and attack targets” (ICRC 2016). AI technologies play a crucial role in the development of ever more sophisticated AWS, by enabling perceptual, deliberative, and action planning capabilities that an AWS needs to perform its critical tasks of target selection and attack.

The Stockholm International Peace Research Institute (SIPRI) proposed a classification of existing AWS (Boulanin and Verbruggen, 2017). Some notable classes of existing AWS are (i) *robotic sentries*, like the Super aEgis II stationary robotic platform, which operates either as a decision-support system to human sentinels or else in autonomous mode to identify and target trespassers in the demilitarized zone between North and South Korea; (ii) *guided munitions*, which autonomously identify and engage targets that are not in sight of the attacking aircraft (e. g., the Dual-Mode aircraft Brimstone); (iii) *loitering munitions*, such as the Harpy NG, which overfly an assigned area in search of targets to dive-bomb and destroy. The classification proposed by SIPRI stands in need of continual expansion on account of ongoing military research projects on ever more sophisticated autonomous weapons systems. Notably, research work based on swarm intelligence technologies is paving the way to the development of large cohorts of small-size AWS, coordinating their action in the absence of centralized control, and aiming to overwhelm enemy forces by their numbers and firepower.

The main ethical issues addressed so far in the normative debate about AWS concern (1) AWS causing breaches of *jus in bello* norms of just war theory and IHL, thereby affecting the rights and the welfare of their victims, (2) the difficulty of selectively attributing responsibilities for IHL breaches to the persons taking on responsibilities and decision-making roles in AWS operation, and (3) affronts that AWS targeting decisions make to the human dignity of its victims (Amoroso and Tamburrini 2020, Amoroso and Tamburrini 2021). Claims (1)-(3) refer to *local* ethical issues, according to the distinction introduced in section 2. Indeed, only the good and the rights of AWS potential victims are selectively at stake, in addition to the

duties of persons who are responsible for AWS operation. The limitations and vulnerabilities of AI systems discussed in the previous section play a central role in the arguments supporting claims (1) and (2) that we now turn to consider.

In connection with claim (1), it has been observed that present and foreseeable AWS may incur violations of the IHL principle of distinction, which establishes that parties involved in a conflict must at all times distinguish between military objectives on the one hand, and civilians and civilian objects on the other hand, directing their attacks only against military objectives, and refraining from directing attacks against civilian objectives. Vivid examples of breaches of the IHL principle of distinction that AWS systems may induce are afforded by adversarial manipulations that go unnoticed to the human perceptual system, bringing an AWS to take images of school buses (that are normally protected by the distinction principle) for images of ostriches, and 3-D models of turtles for rifles. Thus, granting for the sake of argument that the overall performance of AWS may come to match or even statistically surpass the performance of human combatants does not entail that this will occur *in each and every situation*. Even the most sophisticated AWS may commit (or be induced to commit) disastrous mistakes from an IHL perspective—mistakes that might have been avoided if a human operator had been substantively involved in the decision-making loop. This is a major reason offered for endorsing the view that all weapons systems, including AWS, should be subject to meaningful human control (MHC), and that human control is truly meaningful only if human operators are put in the position to act as *fail-safe actors*, contributing to prevent the weapon from delivering indiscriminate attacks in breach of IHL (Amoroso and Tamburrini, 2021).

Let us now turn to claim (2), that is, to the difficulty of selectively attributing responsibilities for IHL breaches to human operators taking designated responsibilities and decision-making roles in AWS deployment and use. To avoid such responsibility gaps, MHC must secure ethical conditions for moral responsibility ascriptions and legal conditions for international criminal law (ICL) responsibility ascriptions, whenever breaches of IHL that are materially caused by AWS occur. To exert a genuine MHC, human operators must be put in the position to play the role of *accountability attractors* (Amoroso and Tamburrini, 2021). In other words, one must preserve the accountability of human operators after the AWS is activated. But specific weaknesses and vulnerabilities of AI learning systems jeopardize the fulfilment of this condition. Let's see.

According to Johannes Himmelreich, the accountability requirement is preserved if the commander of an AWS has “robust tracking control” over the outcomes of its action. Himmelreich proposes the following definition of robust tracking control:

an *a* has control over whether an outcome *x* occurs if (1) there is an order *a* can give, such that (2) if *a* were to give this order, then *x* would occur (in all relevantly similar situations), and (3) if *a* were not to give this order, then *x* would not occur (in all relevantly similar situations). (Himmelreich, 2019, p. 736)

Let us immediately note that to infer that outcome *x* of an AWS action is under their robust tracking control, commanders must be in the position to assert with a high level of confidence that the battlefield situation is *similar* in all relevant aspects to their own cognitive model of the battlefield situation and its predicted outcomes. But confidence in this similarity judgment is jeopardized by the findings of AI adversarial testing: AWS perceptual judgments may be blatantly wrong in ways that are IHL-relevant, unexpected, and counterintuitive for human operators. Even answers provided by an AWS explanation module to a commander’s explanation request does not put at rest these doubts, for adversarial XAI suggests that one may artfully manipulate AI explanations.

Epistemic limitations on the commander’s capability to make robust predictions about the behaviors of AWS are the chief source of the “accountability gap” problem. Suppose that an autonomous weapons system commits a material act that would be equivalent to a war crime had this act been performed by human beings. In other words, the AWS targeting decisions, had they been taken by human agents, would trigger individual criminal responsibility. Since, however, the direct targeting decision was taken by the AWS, who should be held responsible for its conduct? Commanders might complain that an interruption of their robust tracking control condition occurred due to unexpected epistemic predicaments, such as bizarre perceptual errors of the AI learning system embedded in the AWS or hostile interactions on the battleground. Under these circumstances, commanders would plausibly claim that their responsibility was mitigated, or altogether excluded, due to epistemic predicaments. Cases may therefore occur where one cannot ascertain the existence of the mental element (intent or knowledge), which is required under ICL to ascribe and punish criminal responsibilities. Consequently, no one would be held criminally responsible, notwithstanding that the conduct at stake materially amounts to an international crime.

Even more relevant to mitigate responsibility is the epistemic limitation affecting a commander's model of cluttered and unstructured battlefield situations, involving sustained interactions among AWS and other artificial agents. Robust tracking control wanes as AWS loiter for extended periods of time on dynamic battlefields, possibly involving interactions between AWS accelerating the pace of conflict beyond human cognitive capabilities (Altmann and Sauer, 2017). It is this acceleration that looms large on the global ethical issue which is now emerging in connection with the development of autonomous cyber weapons.

7. The ascent from local to global ethical issues in AWS normative debates

Normative debates about AWS have mostly focused on existing and foreseeable AWS that operate in the warfare domains of land, air, and sea – leaving largely aside the cyber domain. However, “parties to armed conflicts frequently deploy cyber weapons and, recognizing the competitive advantages afforded by autonomy, States are developing – or perhaps have already developed – autonomous cyber weapons for use in armed conflict” (Buchan and Tasgourias, 2020, p. 646). Are autonomous cyber weapons genuine AWS? What happens to the normative debate about AWS if one counts autonomous cyber weapons as AWS? Are the involved normative issues bound to stay local?

To begin with, let us note that the 2012 Directive of the US Department of Defense (DOD 2012) – which first introduced the functional condition on a weapons system to count as autonomous (see section 5) – leaves aside any consideration of machine autonomy in the cyberspace. At the same time, however, no explicit restriction in terms of warfare domains is introduced there. AI-powered cyber weapons can in principle use their adaptive intelligence and learning capabilities to identify and exploit without human intervention the software vulnerabilities of other digitalized military systems. These autonomous systems and their targets inhabit the cyberspace, thereby differing from autonomous robotic sentries, loitering munitions or swarms of autonomous armed drones in the way of their operational domain. But they are no different from other AWS in their capability to independently select and attack targets. It is therefore reasonable to conclude – on the basis of the functional approach espoused by the DoD and the ICRC – that an autonomous cyber weapon (ACW from now on) is a genuine AWS.

Like cyber weapons operated by humans, ACW can potentially target surveillance and reconnaissance military systems, weapons system requiring software resources to be operated, software systems serving intelligence and command-and-control purposes at military headquarters. However, by replacing teams of skilled engineers in orchestrating cyber attacks, ACW are likely to accelerate the pace of cyberwarfare beyond human response capabilities, enabling the delivery of cyber attacks on larger scales, and making cyber threats more persistently available in the cyberspace (Heinl, 2018). ACW may target AWS releasing their kinetic force in traditional warfare domains. Accelerating the pace of cyberwarfare may lead to runaway interactions between ACW and AWS. Most significantly, the rise of ACW may aggravate existing cyberthreats to nuclear weapons and related nuclear command, control and communication (NC3) systems. Cyberattacks directed to nuclear defense systems could lead to false warnings of nuclear attacks by the enemy, disrupt access to information and communication, damage nuclear delivery systems, and even enable the hacking of a nuclear weapon (<https://www.nti.org/newsroom/news/new-report-finds-nuclear-weapons-and-related-systems-increasingly-vulnerable-cyberattack/>). Therefore, cyberattacks on nuclear defense systems raise new daunting threats for peace and a global ethical issue concerning the very persistence of human civilizations.

The maturing of these technological possibilities has far-reaching ethical implications involving the responsibilities of AI scientists on account of their privileged epistemic position. Right after World War II, many physicists felt it was their moral obligation to make public opinion and political decision-makers aware of the existential threat posed by nuclear weapons and the nuclear arms race starting during the cold war. Later, chemists and biologists played a pivotal role in international debates and diplomatic efforts leading to international treaties banning the development, production, stockpiling and use of chemical and biological weapons of mass destruction. Today, AI scientists must make public opinion and political decision-makers aware of the threats to peace and stability posed by the maturing of ACW, up to and including their impact on NC3 systems, and the existential threats for human civilization that may emerge from ACW targeting nuclear defense systems. And they must face hard moral choices concerning their active participation in or support of ACW research.

Acknowledgments: This research was partially funded by Italian National Research Project PRIN2020, grant 2020SSKZ7R.

References

- Alcott B. (2005) Jevon's Paradox, *Ecological Economics* 54, 9-21.
- Altmann Jürgen and Frank Sauer, "Autonomous Weapon Systems and Strategic Stability," *Survival* 59, no. 5 (September 2017), pp. 117-42
- Amoroso D., Tamburrini G. (2020) 'Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues', *Current Robotics Reports* 1(7), pp. 187-194. 10.1007/s43154-020-00024-3
- Amoroso, D., & Tamburrini, G. (2021). Toward a Normative Model of Meaningful Human Control over Weapons Systems. *Ethics & International Affairs*, 35(2), 245-272. doi:10.1017/S0892679421000241
- Athalye, A., Engstrom, L., Ilyas, A., Kwok, K. (2018). Synthesizing Robust Adversarial Examples. in Proceedings of Machine Learning Research 80, 284-293 Available at <https://proceedings.mlr.press/v80/athalye18b.html>.
- Boulanin V, Verbruggen (2017). N. Mapping the development of autonomy in weapon systems. Solna: SIPRI Report; 2017.
- Buchan R. and Tsagourias N. (2020) Autonomous Cyber Weapons and Command Responsibility, *International Law Studies* 96, 645-673.
- Cerutti F. (2008) *Global challenges for the Leviathan. A political philosophy of nuclear weapons and global warming*, New York, Lexington Books.
- Comiter M. (2019). Attacking Artificial Intelligence. AI's security vulnerability and what policymakers can do about it. Belfer Center for Science and International Affairs, Harvard Kennedy School, August 2019.
- DOD (2012) U.S. Department of Defense, "Autonomy in Weapons Systems," Directive 3000.09 (November 21, 2012).
- European Commission (2020). *White paper on AI. A European approach to excellence and trust*, Bruxelles, 19 February 2020. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en
- European Commission (2021). *Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, Brussels, 21.4.2021
- Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 3681-3688. <https://doi.org/10.1609/aaai.v33i01.33013681>
- Heo Juyeon, Joo, Sunghwan, Moon, Taesup (2019) Fooling Neural Network Interpretations via Adversarial Model Manipulation, in H. Wallach, H. Larochelle, A. Beygelzimer, F. Alchêbuc, E. Fox, R. Garnett (eds) *Advances in Neural Information Processing Systems*, Curran Associates, <https://proceedings.neurips.cc/paper/2019/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf>
- Heinl C. (2017) Maturing autonomous cyber weapons systems: Implications for international security cyber and autonomous weapons systems regimes,

Acknowledgments: This research was partially funded by Italian National Research Project PRIN2020, grant 2020SSKZ7R

- to appear in Paul Cornish (ed), Oxford Handbook of Cyber Security, Oxford University Press. Retrieved online https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3255104
- Henderson Peter, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, Joelle Pineau (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning, *Journal of Machine Learning Research* 21, 1-43.
- Himmelreich Johannes (2019), "Responsibility for Killer Robots," *Ethical Theory and Moral Practice* 22, no. 3 (June 2019), pp. 731–47.
- ICRC (2016). International Committee of the Red Cross, "Views of the International Committee of the Red Cross on Autonomous Weapon System" (paper submitted to the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons, Geneva, April 11, 2016)
- ICRC (2019) International Committee of the Red Cross, *Autonomy, artificial intelligence and robotics: Technical aspects of human control* Geneva, August 2019, Geneva.
- IPPC (2018). Intergovernmental Panel on Climate Change. *Global Warming of 1.5°* August 2018, Geneva. <https://www.ipcc.ch/sr15/>
- IPPC (2021). Intergovernmental Panel on Climate Change. *Climate Change 2021. The physical science basis*, August 2021, Geneva
- Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the Carbon Emissions of Machine Learning. Technical report <https://deepai.org/publication/quantifying-the-carbon-emissions-of-machine-learning>
- Lucivero, F. (2019) 'Big data, big waste? A reflection on the environmental sustainability of big data initiatives', *Science and Engineering Ethics*, 26, pp. 1009–30. <https://doi.org/10.1007/s11948-019-00171-7>
- Müller, Vincent C. (2021), "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/>
- NSCAI (2021). National Security Commission on Artificial Intelligence. *Final Report*. Washington, DC, <https://www.nscai.gov/2021-final-report/>
- O'Neil, C. (2017). *Weapons of math destruction*. Penguin Books.
- Patterson David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean (2020) Carbon Emissions and Large Neural Network Training, <https://arxiv.org/pdf/2104.10350>
- Rolnick D. et al. (2019) 'Tackling Climate Change with Machine Learning', *arxiv.org.1906.05433*. <https://arxiv.org/abs/1906.05433>
- Santoni de Sio Filippo and Jeroen van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," *Frontiers in Robotics and AI* 5

Acknowledgments: This research was partially funded by Italian National Research Project PRIN2020, grant 2020SSKZ7R

- Schwartz R., Dodge J., Smith N. A., Etzioni O. (2020) Green AI, *CACM* 63(12), pp. 54-63.
- Strubell E., Ganesh A., McCallum A. (2019) ‘Energy and Policy Considerations for Deep Learning in NLP’, *arxiv.org.1906.02243*. <https://arxiv.org/abs/1906.02243>
- Szegedy Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, “Intriguing Properties of Neural Networks,” *arXiv.org*, last updated February 19, 2014; <https://arxiv.org/abs/1312.6199v4>
- Tamburrini G., Altiero F. (2021). Research Programs Based on Machine Intelligence Games, in S. Chiodo, V. Schiaffonati (eds.), *Italian Philosophy of Technology*, Springer, pp. 163-179. https://link.springer.com/chapter/10.1007/978-3-030-54522-2_11
- United Nations (1945). United Nations Charter, <https://www.un.org/en/about-us/un-charter>
- United Nations (2000). United Nations Millennium Declaration, <https://digitallibrary.un.org/record/422015>
- Walzer, Michael (1978). *Just and Unjust Wars*. Basic Books.
- Wolff Anthony L. F., Kanding B., Selvan R. (2020). Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models, 2020 ICML Workshop on “Challenges in Deploying and monitoring Machine Learning Systems”, <https://arxiv.org/pdf/2007.03051>

Acknowledgments: This research was partially funded by Italian National Research Project PRIN2020, grant 2020SSKZ7R