AI standards and regulations?

A small new generation of ethical standards is emerging as the ethical, legal, and societal impacts of artificial intelligence and robotics. Whether a standard clearly articulates explicit or implicit ethical concerns, all standards embody some kind of ethical principle. The standards that do exist are still in development and there is the limited publicly available information on them. Perhaps the earliest explicit ethical standard in robotics is BS 8611 Guide to the Ethical Design and Application of Robots and Robotic Systems. BS8611 is not a code of practice, but guidance on how designers can identify potential ethical harm, undertake an ethical risk assessment of their robot or AI, and mitigate any ethical risks identified. It is based on a set of 20 distinct ethical hazards and risks, grouped under four categories: societal, application, commercial & financial, and environmental.

British Standard BS 8611 assumes that physical hazards imply ethical hazards, and defines ethical harm as affecting 'psychological and/or societal and environmental well-being.' It also recognises that physical and emotional hazards must be balanced against the user's expected benefits. The standard highlights the need to involve the public and stakeholders in development of robots and provides a list of key design considerations including:

- Robots should not be designed primarily to kill humans;
- Humans remain responsible agents;
- It must be possible to find out who is responsible for any robot;
- Robots should be safe and fit for purpose;
- Robots should not be designed to be deceptive;
- The precautionary principle should be followed;
- Privacy should be built into the design;
- Users should not be discriminated against, nor forced to use a robot

Particular guidelines are provided for roboticists, particularly those conducting research. These include engaging the public, considering public concerns, working with experts from other disciplines, correcting misinformation, and providing clear instructions. Specific methods to ensure the ethical use of robots includes: user validation (to ensure the robot can/is operated as expected), software verification (to ensure the software works as anticipated), involvement of other experts in ethical assessment, economic and social assessment of anticipated outcomes, assessment of any legal implications, compliance testing against relevant standards.

# IEEE 'human standards' with implications for AI

The IEEE Standards Association has also launched a standard via its global initiative on the Ethics of Autonomous and Intelligent Systems. Positioning 'human well-being' as a central precept, the IEEE initiative explicitly seeks to reposition robotics and AI. It aims to educate, train and empower AI/robot stakeholders to 'prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.' Currently, 14 IEEE standards working groups are working on drafting so-called 'human' standards that have implications for artificial intelligence.

Standard		Aims/Objectives
P7000	Model process for addressing ethical concerns during system design	To establish the process of **ethical design of an autonomous and intelligent system**
P7001	Transparency of Autonomous Systems	To ensure the transparency of autonomous systems to a range of stakeholders. It specifically will address: • Users: ensuring users understand what the system does and why with the intention of building trust; • Validation and certification: ensuring the system is subject to scrutiny; • Accidents: enabling accident investigators to undertake investigation; • Lawyers and expert witnesses: ensuring that, following an accident, these groups are able to give evidence; • Disruptive technology (e.g. driverless cars): enabling the public

		to assess technology (and, if appropriate, build confidence).
P7002	Data Privacy process	To establish standards for the ethical use of personal data in software engineering processes. It will develop and describe privacy impact assessments (PIA) that can be used to identify the need for, and effectiveness of, privacy control measures. It will also provide checklists for those developing software that uses personal information.
P7003	Algorithmic Bias Considerations	To help algorithm developers make explicit the ways in which they have sought to eliminate or minimise the risk of bias in their products. This will address the use of overly subjective information and help developers ensure they are compliant with legislation regarding protected characteristics (e.g. race, gender). It is likely to include: • Benchmarking processes for the selection of data sets; • Guidelines on communicating the boundaries for which the algorithm has been designed and validated (guarding against unintended consequences of unexpected uses); • Strategies to avoid incorrect interpretation of system outputs by users.
P7004	The standard for Child and	Specifically aimed at educational

	Student Data Governance	institutions, this will provide guidance on accessing, collecting, storing, using, sharing, and destroying child/student data.
P7005	Standard for Transparent Employer Data Governance	Similar to P7004, but aimed at employers
P7006	standard for Personal Data Artificial Intelligence (AI) Agent	Describes the technical elements required to create and grant access to personalized AI. It will enable individuals to safely organize and share their personal information at a machine-readable level, and enable personalized AI to act as a proxy for machine-to-machine decisions
P7007	Ontological Standard for Ethically Driven Robotics and Automation Systems	This standard brings together engineering and philosophy to ensure that user well-being is considered throughout the product life cycle. It intends to identify ways to maximize benefits and minimize negative impacts, and will also consider the ways in which communication can be clear between diverse communities.
P7008	The standard for Ethically Driven Nudging for Robotic, Intelligent, and Autonomous Systems	Drawing on 'nudge theory', this standard seeks to delineate current or potential nudges that robots or autonomous systems might undertake. It recognises that nudges can be used for a range of reasons, but that they seek to affect the recipient emotionally, change behaviors, and can be manipulative,

		and seeks to elaborate methodologies for the ethical design of AI using nudges.
P7009	The standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems	To create effective methodologies for developing and implementing robust, transparent, and accountable fail-safe mechanisms. It will address methods for measuring and testing a system's ability to fail safely
P7010	Well-being Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems	To establish a baseline for metrics used to assess well-being factors that could be affected by autonomous systems, and for how human well-being could proactively be improved.
P7011	The standard for the Process of Identifying and Rating the Trustworthiness of News Sources	Focusing on news information, this standard sets out to standardize the processes for assessing the factual accuracy of news stories. It will be used to produce a 'trustfulness' score. This standard seeks to address the negative effects of unchecked 'fake' news and is designed to restore trust in news purveyors.
P7012	The standard for Machine Readable Personal Privacy Terms	To establish how privacy terms are presented and how they could be read and accepted by machines
P7013	Inclusion and Application Standards for Automated Facial Analysis Technology	To provide guidelines on the data used in facial recognition, the requirements for diversity, and benchmarking of applications and situations in which facial recognition should not be used.

# Machine Ethics:

Machine ethics is ethics for machines, for "ethical machines", for machines as *subjects*, rather than for the human use of machines as *objects.* machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable. (Anderson and Anderson 2007: 15). The basic idea of machine ethics is now finding its way into actual robotics where the assumption that these machines are artificial moral agents in any substantial sense is usually not made (Winfield et al. 2019). It is sometimes observed that a robot that is programmed to follow ethical rules can very easily be modified to follow unethical rules (Vanderelst and Winfield 2018). The idea that machine ethics might take the form of "laws" has famously been investigated by Isaac Asimov, who proposed "three laws of robotics" (Asimov 1942):

First Law—A robot may not injure a human being or, through inaction, allow a human being to come to harm. Second Law—A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Third Law—A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

# Artificial Moral Agents:

If one takes machine ethics to concern moral agents, in some substantial sense, then these agents can be called "artificial moral agents", having rights and responsibilities.

Several authors use "artificial moral agent" in a less demanding sense, borrowing from the use of "agent" in software engineering in which case matters of responsibility and rights will not arise. James Moor (2006) distinguishes four types of machine agents: ethical impact agents (e.g., robot jockeys), implicit ethical agents (e.g., safe autopilot), explicit ethical agents (e.g., using formal methods to estimate utility), and full ethical agents (who "can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent".) Several ways to achieve "explicit" or "full" ethical agents have been proposed, via programming it in (operational morality), via "developing" the ethics itself (functional morality), and finally full-blown morality with full intelligence and sentience. Programmed agents are sometimes not considered “full” agents because they are “competent without comprehension”, just like the neurons in a brain.

# Singularity:

The idea of *singularity* is that if the trajectory of artificial intelligence reaches up to systems that have a human level of intelligence, then these systems would themselves have the ability to develop AI systems that surpass the human level of intelligence, i.e., they are “super intelligent” (see below). Such super-intelligent AI systems would quickly self-improve or develop even more intelligent systems. This sharp turn of events after reaching super intelligent AI is the “singularity” from which the development of AI is out of human control and hard to predict (Kurzweil 2005: 487).

Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion”, and the intelligence of man would be left far behind. Thus the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

The optimistic argument from acceleration to singularity is spelled out by Kurzweil (1999, 2005, 2012) who essentially points out that computing power has been increasing exponentially, i.e., doubling ca. every 2 years since 1970 in accordance with “Moore’s Law” on the number of transistors, and will continue to do so for some time in the future. He predicted (Kurzweil 1999) that by 2010 supercomputers will reach human computation capacity, by 2030 “mind uploading” will be possible, and by 2045 the “singularity” will occur. Kurzweil talks about an increase in computing power that can be purchased at a given cost—but of course, in recent years the funds available to AI companies have also increased enormously: Amodei and Hernandez (2018 [OIR]) thus estimate that in the years 2012–2018 the actual computing power available to train a particular AI system doubled every 3.4 months, resulting in a 300,000x increase—not the 7x increase than doubling every two years would have created. Criticism of the singularity narrative has been raised from various angles. Kurzweil and Bostrom seem to assume that intelligence is a one-dimensional property and that the set of intelligent agents is totally-ordered in the mathematical sense—but neither discusses intelligence at any length in their books. Generally, it is fair to say that despite some efforts, the assumptions made in the powerful narrative of superintelligence and singularity have not been investigated in detail. One question is whether such a singularity will ever occur—it may be conceptually impossible, practically impossible or may just not happen because of contingent events, including people actively preventing it. Philosophically, the interesting question is whether singularity is just a “myth” (Floridi 2016; Ganascia 2017), and not on the trajectory of actual AI research. This is something that practitioners often assume (e.g., Brooks 2017 [OIR]). They may do so because they fear the public relations backlash, because they overestimate the practical problems, or because they have good reasons to think that superintelligence is an unlikely outcome of current AI research (Müller forthcoming-a). This discussion raises the question whether the concern about “singularity” is just a narrative about fictional AI based on human fears. But even if one *does* find negative reasons compelling and the singularity not likely to occur, there is still a significant possibility that one may turn out to be wrong. Philosophy is not on the “secure path of a science” (Kant 1791: B15), and maybe AI and robotics aren’t either (Müller 2020). So, it appears that discussing the very high-impact risk of singularity has justification *even if* one thinks the probability of such singularity ever occurring is very low.

# Ethics in military use of AI: use of weapons:

We begin by defining some common terms.

Autonomous: In AI and robotics autonomy simply means the ability to function without a human operator for a protracted period of time (Bekey 2005). Robots may have autonomy over the immediate decision that they make but generally do not have autonomy over their choice of goals. There is some controversy as to what “autonomous” means for lethal weapons systems.

Lethal and Harmful autonomy: A weapon can be said to be “autonomous” in the “critical functions of targeting” if it can do one or more of following without a human operator. If the weapon can decide what classes of objects it will engage, then it would be autonomous in terms of defining its targets. No current AWS has this capability. If a weapon can use sensors to select a target without a human operator, it can be said to have autonomy in the selection function of targeting. Many existing weapons can select targets without a human operator. If a weapon can fire on a target without a human operator, it can be said to have autonomy in the engage function of targeting. Many existing weapons can engage already selected targets autonomously. For example, the Patriot anti-missile system can select targets autonomously but by design requires a human operator to hit a confirm button to launch a missile. Once the missile is launched, it can hit its 11.1 Definitions 95 target without a human operator. Given the speeds involved, human control of a Patriot missile is not possible.

Non-Lethal Autonomy: An AWS may have “autonomy” in many other functions. It might be able to take off and land autonomously and it might be able to navigate autonomously. However, this non-lethal “autonomy” is generally not regarded as morally controversial. Killer robots Autonomous weapons are often called “killer robots” in mass media reports. Some object to the use of the term. Lokhorst and van den Hoven describe the phrase as an “insidious rhetorical trick” (Lokhorst and Van Den Hoven 2012). However, this is favored by the “Campaign to Stop Killer Robots”.1 This is an umbrella group of human rights organizations seeking an international ban on lethal autonomous weapons systems.

# The use of Autonomous weapons systems

Arguments can and are made against the use of an autonomous weapons system. Generally, these arguments focus on the following issues.

1. Discrimination:

   Proponents typically concede that machines cannot, in general, discriminate as well as humans. However in some particular cases they can discriminate better than humans. For example, Identification Friend or Foe (IFF) technology sends a challenge message to an unindentified object in the sky which the object must answer or risk being shot down. Typically in air war, contested airspace is known to civilian air traffic control and neutral aircraft will not enter it. However, in 2014, a civilian airliner, Malaysia Airlines flight MH 17 en route from Amsterdam to Kuala Lumpur was shot down by a Russian Surface to Air Missile (SAM) operated by Russian secessionists in the Eastern Ukraine. This

SAM system was human-operated and not equipped with IFF. Some have observed that a more advanced system would have known the target was a civilian airliner. Proponents also note that vision systems are continuously improving. Advancing technology has dramatically improved the ability of vision, auditory, LIDAR, and infrared systems which are quickly reaching parity with humans in terms of object discrimination. A possible ethical dilemma may be approaching if an autonomous system demonstrates clear superiority to humans in terms of targeting. We may be ethically obligated to consider their use. Even so, it remains difficult for machines to distinguish between different types of behavior such as acting peacefully or fighting in a conflict.

2. Proportionality:

   Proponents, on the other hand, state that “excessive” is a relative concept that is not well-defined in International Humanitarian Law (IHL). Enemark makes the point that politicians generally do not advertise their proportionality calculations (Enemark 2013). Situations in which intelligence reveals the location of a high-value target demand a decision. Opponents claim that AWS cannot calculate proportionality (Braun and Brunstetter 2013). Proportionality is the ability to decide how much collateral damage is acceptable when attacking a military target. The standard is that “collateral damage” must not be “excessive” compared to the concrete military advantage gained. Proportionality calculations typically attempt to estimate the number of civilians that may be killed versus the military necessity of the target. Generating such calculations often involves input from a variety of experts including lawyers. It is difficult to imagine how AWS could successfully complete such a calculation.

3. Responsibility:

   Opponents of AWS argue that machines cannot be held morally responsible. They then argue that this is a reason to ban AWS. It is indeed hard to imagine how a machine can be assigned moral responsibility. However, those defending the use of AWS are inclined to assign moral responsibility for the actions of the machine to those that design, build and configure it. Thus those humans deploying an AWS can be held responsible for its actions (Arkin 2008). This raises the “problem of many hands” (Thompson 1980) in which the involvement of many agents in a bad outcome makes it unclear where responsibility lies. Clearly, if an incident were to occur an investigation would result to determine fault. Legally it is easier to hold the collective entity responsible. There is a concept of “strict liability” in law that could be used to assign responsibility to the state that operates the weapon in an AWS regulation.

Opponents also argue that, unlike a human, an AWS cannot be held responsible for its actions or decisions. While machines can be grounded for performance errors there is no true way to punish these systems in a metaphysical sense. Moreover, it may not be just to punish the commanders of these systems if they utilize automatic targeting.

# Regulations governing an AWS:

States at the UN agree that an AWS must be used in compliance with existing IHL. The Convention on Certain Conventional Weapons is generally considered the appropriate forum to discuss AWS regulations. These regulating bodies require meaningful human control over the AWS. In particular, this means the following:

1. an AWS must be able to distinguish between combatants and non-combatants
2. an AWS must be able to calculate proportionality
3. an AWS must comply with the principle of command responsibility

# Ethical arguments for and against AI for military purposes:

Arguments in favour:

In IHL the doctrine of military necessity permits belligerents to do harm during the conduct of war. Moreover, the just war theory states that, although war is terrible, there are situations in which not conducting a war may be an ethically and morally worse option. For example, war may be justifiable to prevent atrocities. The purpose of just war theory is to create criteria that ensure that war is morally justifiable. Just war theory includes criteria for (1) going to war and (2) conducting war.

The criteria for going to war include: just cause, comparative justice, competent authority, right intention, probability of success, last resort, and proportionality.

The criteria for conducting war include: distinction, proportionality, military necessity, fair treatment of prisoners of war, and not using means and methods of warfare that are prohibited. Examples of prohibited means of warfare include chemical and biological weapons. Examples of prohibited means include mass rape and forcing prisoners of war to fight against their own side.

The overall intent of IHL is to protect the rights of the victims of war. This entails rules that minimize civilian harm. With respect to the use of AI and robots in warfare, some have argued that AMS may reduce civilian casualties. Unlike humans, artificially intelligent robots lack emotions, and thus acts of vengeance and emotion-driven atrocities are less likely to occur at the hands of a robot. In fact, it may be the case that robots can be constructed to obey the rules of engagement, disobeying commands to violate civilian and enemy combatant rights. If nothing else, units being observed by a military robot may be less inclined to commit such atrocities. If, in fact, robots can be used to prevent atrocities and ensure the minimization of civilian casualties, then military leaders may have an ethical obligation to use such robots. For, not using such robots, condemns a greater number of civilians to die in a morally justified war. Moreover, an AWS may be capable of non-lethal offensive action where human units must use lethal force. Others argue that AI and military robots are necessary for defensive purposes. Some research has shown that in certain circumstances, such as aerial combat, autonomous systems have clear advantages over human systems. Hence, sending humans to fight an AWS is unlikely to succeed and may result in substantial casualties. In this situation, leaders have an ethical obligation to reduce their own casualties even if this means developing AWS for their own purposes.

Arguments Against:

It has been claimed that the advent of artificial intelligence technologies for military use could lead to an arms race between nations. Vladimir Putin, the President of the Russian Federation, said in 2017 that “the nation that becomes the leader in AI will rule the world.”. China has similarly increased spending on AI and the United States has long made the development of AI for defense purposes a priority. Experts generally agree that AWS will generate a clear and important military advantage (Adams 2001).

Relatedly, some argue that the use of an AWS is unfair in that such weapons do not result in an equal risk to all combatants. Researchers also note that the possession and use of autonomous weapons systems may actually instigate wars because the human cost of war is reduced. There is some evidence for this claim based on targeted killings in Iraq by the United States. The transition in Iraq from human-piloted missions to unmanned aerial vehicles resulted in a dramatic increase in the number of targeting missions (Singer 2009). This evidence, although important, should not discount the political and technological factors that may also have contributed to the increase in targeted killings. Perhaps the most philosophically interesting argument leveled against the use of AWS is the dignity argument claiming that "death by algorithm" is the ultimate indignity. In its more complex forms, the argument holds that there is a fundamental human right not to be killed by a machine. From this perspective, human dignity, which is even more fundamental than the right to life, demands that a decision to take human life requires consideration of the circumstances by a human being. A related claim is that meaningful human control of an autonomous weapon requires that a human must approve the target and be engaged at the moment of combat.