Chapter 2
The Human Control Over Autonomous Robotic Systems: What Ethical and Legal Lessons for Judicial Uses of AI?

Daniele Amoroso and Guglielmo Tamburrini

Abstract This contribution provides an overview of normative problems posed by increasingly autonomous robotic systems, with the goal of drawing significant lessons for the use of AI technologies in judicial proceedings, especially focusing on the shared control relationship between the human decision-maker (i.e. the judge) and the software system. The exemplary case studies that we zoom in concern two ethically and legally sensitive application domains for robotics: autonomous weapons systems and increasingly autonomous surgical robots. The first case study is expedient to delve into the normative acceptability issue concerning autonomous decision-making and action by robots. The second case study is used to investigate the human responsibility issue in human-robot shared control regimes. The convergent implications of both case studies for the analysis of ethical and legal issues raised by judicial applications of AI enable one to highlight the need for and core contents of a genuinely *meaningful* human control to be exerted on the operational autonomy, if any, of AI systems in judicial proceedings.

2.1 Introduction

Recent advances in robotics and artificial intelligence (AI) have paved the way to robots autonomously performing a wide variety of tasks[1] that may significantly affect individual and collective interests, which are worthy of protection from both ethical and legal perspectives. Exemplary cases are the application of lethal force by

[1]A robotic system may be counted as “autonomous” at given tasks if, once activated, it is able to carry out those tasks without further human intervention.

D. Amoroso (✉)
University of Cagliari, Cagliary, Italy
e-mail: daniele.amoroso@unica.it

G. Tamburrini
University of Naples Federico II, Naples, Italy
e-mail: guglielmo.tamburrini@unina.it

X. Kramer et al. (eds.), *New Pathways to Civil Justice in Europe*,
https://doi.org/10.1007/978-3-030-66637-8_2

autonomous weapons systems (AWS) and the circulation of autonomous vehicles on public roads. But one may also think of increasingly autonomous surgical and care robots.[2] To this list of robotic systems one may certainly add the judicial use of AI software systems, notwithstanding the lack of their (direct) kinetic interaction with the physical world. Indeed, the use of AI in the Court is aimed at replacing or supporting the human judge in decision-making processes and tasks that, *by their very definition*, are supposed to have an impact on legal rights and duties.[3]

These technological developments have fuelled, together with new machine autonomy issues, longstanding discussions on Ethical, Legal and Socio-Economic implications of robotics and AI (ELSE issues in robotics and AI); the origins of which can be traced back at least to Norbert Wiener's seminal reflections on the ethics of information technologies and robotics.[4] Present debates about machine autonomy in ethically and legally sensitive domains have now reached out well beyond academic and specialist circles, entering the political debate and receiving considerable media coverage.[5] Over and above the specificities of each technological application domain, there are a few overarching issues arising in connection with most artificial systems endowed with autonomy in the execution of tasks that are ethically and legally sensitive. First, there is the *technical question* concerning whether artificial agents are inherently unable to carry out properly certain functions governed by law (e.g. making proportionality assessments in the context of a military attack), insofar as they would (allegedly) require uniquely human capabilities. Second, there is the (strictly) *legal problem* of determining how to allocate responsibility if harmful events caused by the machine occur. Third, it is debated—from the perspective of *deontological ethics*[6]—whether it is morally acceptable to remove human agency from decision-making processes that are likely to impinge on individual rights and duties. Fourth and finally, from a *consequentialist* perspective in normative ethics,[7] there is the question concerning the opportunity, or perhaps even the moral and legal duty, to replace human operators with autonomous machines, whenever the latter's performances ensure better protection of the interests at stake (e.g. by reducing the number of road accidents and fatalities).

Although these issues have been raised in recent times especially with regard to autonomous robots, they appear to be at least equally relevant when one considers the use of AI in the courtroom. To begin with, one may question whether it will ever be *technically possible* to program software that is able to reliably replace human

[2]With regard to autonomous vehicles and care robots, which will not be examined in this contribution, see respectively Lin (2015), Decker (2008).

[3]See, on this issue, the recent monograph by Nieva Fenoll (2018).

[4]Wiener (1950, 1964).

[5]See, among many others, The Guardian (2019), Schwarzman (2019), Metz (2019).

[6]Broadly speaking, deontological ethics identifies moral duties as guides for acting and judging the moral worth of choices.

[7]Unlike deontological ethics, consequentialism focuses on criteria to distinguish between morally good and bad consequences of choices, and prescribes to judge the moral worth of choices in the light of consequences only.

judges in performing tasks involving discretionary reasoning and/or equitable evaluations. Also, not differently from what happens in relation to robotic systems, the problem may arise as to who is to be blamed for wrongful AI-based judicial decisions that might trigger disciplinary responsibility or compensatory proceedings. Finally, even in relation to judicial applications of AI, one may detect an ethical tension between consequentialist reasons favouring the use of these applications (e.g. the need to avoid decisions tampered by typically-human biases) and the view, ultimately based on deontological ethics, whereby human agency should always be retained in the judicial decision-making process.

Against this background, the present contribution provides an overview of normative—both ethical and legal—problems posed by increasingly autonomous robotic systems, with the aim of drawing some lessons for the use of AI technologies in judicial proceedings, especially focusing on lessons concerning the shared control relationship between the human decision-maker (i.e. the judge) and the software system. After briefly expanding on the notion of *operational* machine autonomy (Sect. 2.2), we will focus on two case studies: autonomous weapons systems (Sect. 2.3) and increasingly autonomous surgical robots (Sect. 2.4). These case studies enable us to highlight—respectively—the issue of acceptability of autonomous decision-making in ethically and legally sensitive domains, and the issue of human responsibility for harmful events caused by autonomous artificial agents. Finally, we zoom in on the implications of this analysis for addressing ethical and legal issues raised by judicial applications of AI (Sect. 2.5). Section 2.6 concludes.

2.2 A Preliminary Distinction: Personal vs. Operational Autonomy

It is important to allay a possible source of confusion deriving from the prevalent use of the word "autonomy" in both legal and philosophical parlance. Definitions employed therein in relation to so-called "*personal* autonomy" are not helpful to grasp ascriptions of "autonomy" to robotic systems and other artificial agents. Personal autonomy is attributed only to self-aware entities, conscious of the surrounding world, and capable to act on their own genuine intentions.[8] No educated guess can be presently made as to whether a machine will be built meeting the conditions to enjoy personal autonomy. Whether this will ever happen in some undetermined technological future may be the object of stimulating speculations; but this is of no avail for significant ethical and legal questions arising today, in relation to already existing or imminent autonomous systems.

As anticipated in the introductory section, a notion of "autonomy" more suitable for our purposes is that of "*operational* autonomy". In this perspective, the "autonomy" of a robotic system stems from (and is solely relative to) its capacity to carry

[8] See, among many others, Buss and Westlund (2018).

out one or more tasks without requiring any intervention by human operators/users.[9] Systems that are endowed with operational autonomy form a broad and heterogeneous class. Indeed, the repertoire of tasks that a machine can autonomously perform includes both complex activities, such as the execution of parts of a surgical intervention, and very simple ones, like a thermostat switching a boiler on or off. Analogously, the autonomous performance of judicial functions by some software system may—at least as a matter of principle—either concern repetitive activities and tasks (e.g. the issuance of injunctive reliefs based on electronic invoices) or very demanding ones, including evidence assessment or the working out of the grounds for a judicial decision.[10]

We are presently interested in novel ethical and legal problems raised by some of these systems. These problems usually arise in connection with the performance of tasks requiring perceptual, cognitive and evaluative capabilities that, until recent times, were in the exclusive purview of human beings. It is clear, therefore, that in the present contribution we will mainly look at the current technological frontier of systems endowed with operational autonomy, the realization of which has been now made possible by the technological advances in the fields of AI and robotics and their synergic confluences. Systems of this kind are the autonomous weapons systems, which we now turn to consider. These systems raise novel ethical and legal issues concerning the conduct of warfare operations and the forms of control that humans ought to exert on their action.

2.3 Autonomous Weapons Systems and the Boundaries of Normatively Acceptable Autonomy

According to a widely held view, which is consistent with the broader notion of operational autonomy set out above, a weapons system is autonomous only if it is able to select and engage targets without any human intervention.[11] On the initiative of civil society, the international community has recently begun debating the legality and ethical acceptability of autonomous weapons systems (AWS), mainly within the institutional framework of the Convention on Conventional Weapons (CCW), first in the context of informal meetings (2014–2016) and then, starting in 2017, within a Group of Government Experts.[12]

[9]Vamvoudakis et al. (2015).

[10]Nieva Fenoll (2018), pp. 24–30.

[11]See, in almost identical terms, US Department of Defence, Directive 3000.09, 'Autonomy in Weapons Systems', 21 November 2012, 13–14; International Committee of the Red Cross (2016), p. 1.

[12]Awareness about this topic has been raised by the Campaign 'Stop Killer Robots', which was launched in 2013 by an international coalition of NGOs with the primary goal of banning lethal robot weapons. For a full chronology of Campaign's activities, see https://www.stopkillerrobots.org/action-and-achievements/. The proceedings of the CCW debates are available at https://www.

The reason why this technological development has created so much interest (unlike, for example, a drone that autonomously executes navigation tasks) is that AWS autonomy concerns the critical functions of target selection and engagement in warfare operations. These functions are "critical" because their performance (i) is crucially regulated by international law of armed conflicts (or international humanitarian law, IHL); (ii) is a key factor for the purposes of individual and state responsibility; (iii) implies moral choices that affect, and even profoundly so, ethically relevant and legally protected individual positions (the right to life and physical integrity, the right to housing, and so on).

The ethical and legal acceptability of autonomous weapons systems (AWS) is questioned from both deontological and consequentialist standpoints, while arguments for the ethical and legal acceptability of AWS are mainly framed in consequentialist terms.[13] Let us briefly summarize the AWS acceptability debate cast in deontological and consequentialist terms, starting from three deontological arguments against AWS.

(1) The first deontological argument supports the claim that AWS would be unable to comply with the principles of distinction, proportionality and precaution under IHL.[14] The development of AWS fulfilling distinction and proportionality requirements at least as well as a competent and conscientious human soldier presupposes the solution of many profound research problems in artificial intelligence (AI) and advanced robotics.[15] Furthermore, it is questionable whether the elimination of human judgment and supervision is compatible with the obligation to take all feasible precautions to prevent (disproportionate) damage to the civilian population, insofar as the regular behaviour of AI and robotic systems is perturbed by unpredicted dynamic changes occurring in warfare environments. Notably, systems developed by means of advanced machine learning technologies (e.g. deep learning) have been extensively demonstrated by adversarial testing to be prone to unexpected, counter-intuitive and potentially catastrophic mistakes, which a human operator would easily detect and avoid.[16]

---

unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument.

[13]For a first, comprehensive exposition of the ethical and legal problems at stake, see Heyns (2013), paras 63–97.

[14]See, respectively, Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, Articles 48, 51(2) and 52(2) (distinction), 51(5)(b) (proportionality), and 57 (precaution).

[15]This is acknowledged also by roboticists who, in principle, are in favour of autonomy in weapons systems. See Arkin (2009), pp. 211–212 (listing the "daunting problems" to be addressed in order to develop an IHL-compliant AWS).

[16]Szegedy et al. (2014), who showed how a change of a few pixels into a schoolbus input image forced a neural network to change its initially correct "schoolbus" classification into an "ostrich" classification. The small image perturbations causing this surprising switch in the neural network classification go completely unnoticed to the human visual system. Indeed, the latter is not similarly

(2) The upshot of the second deontological argument is that AWS are likely to determine an accountability gap. One cannot exclude that AWS will assume targeting decisions that, were they taken by human agents, would trigger individual criminal responsibility. Who will be held responsible for this conduct? The list of potentially responsible persons in the decision-making chain includes the military commander in charge and those overseeing the AWS operation, in addition to manufacturers, robotics engineers, software programmers, and those who conducted the AWS weapons review. People in this list may cast their defence against responsibility charges and criminal prosecution in terms of their limited decision-making roles, as well as of the complexities of AWS systems and their unpredictable behaviour in the battlefield. Cases may occur where it is impossible to ascertain the existence of the mental element (intent, knowledge or recklessness), which is required under International Criminal Law (ICL) to ascribe criminal responsibilities. Consequently, no one would be held criminally liable, notwithstanding the conduct at stake materially amounts to an international crime. This outcome is hardly reconcilable with the moral duty of military commanders and operators to be accountable for their own actions, as well as with the related principle of individual criminal responsibility under ICL.[17]

(3) The third deontological argument supports the claim that AWS would run counter to the principle of human dignity, which would dictate that decisions affecting the life, physical integrity and property of human beings should be entirely reserved to human operators and cannot be entrusted to an autonomous artificial agent. Otherwise, people subject to AWS' use of force would be placed in a position where any appeal to the shared humanity of persons placed on the other side—and thus their inherent value as human beings—would be a priori and systematically denied.[18]

As regards consequentialist approaches to the AWS debate, a distinction should be drawn between *narrow* and *wide* arguments. Narrow consequentialist arguments focus on AWS battlefield performances and their immediate expected outcomes. Wider approaches bring into the picture expected geopolitical consequences for peace and stability of AWS development and deployment. The main argument for permitting AWS on consequentialist grounds is based on a narrow appraisal of expected consequences:

fooled to change its initial "schoolbus" classification. Clearly, mistaken classifications of this sort by an AWS perceptual system may lead to an AI disaster in warfare scenarios, whereby an object normally protected by international humanitarian law (such as a schoolbus) is mistakenly classified as an object which normally is not protected in the same way (such as an ostrich). Additional adversarial testing results that are relevant from an international humanitarian law perspective were reported by Athalye et al. (2018). Specifically, a model of a turtle obtained by a 3D printing process was initially classified correctly as a turtle by a suitably trained AI system. However, by slightly modifying the 3D model - in ways that go unnoticed to the naked human eye - the AI system was induced to classify the newly produced object as a rifle.

[17] See, also for further references, Amoroso and Giordano (2019).

[18] See, also for further references, Amoroso (2020), pp. 161–215.

(4) AWS have the potential to bring about reduced casualties in each single battlefield. This expectation is grounded in the expectation that AWS will perform more accurate targeting than human soldiers, and will take more conservative firing decisions, insofar as they can be programmed to have no self-preservation concerns. When these conditions are obtained, the choice of permitting AWS deployment ought to be preferred on consequentialist grounds.[19]

The force of argument (4) depends on the other-things-being-equal assumption that the deployment of AWS will not have a significant impact outside battlefield scenarios. This *ceteris paribus* assumption is challenged by argument (5), which is based on a broader approach to consequence appraisal:

(5) The spreading of AWS is likely to bring about comprehensive consequences for international security and peace, which outweigh any local and short-term advantage one may envisage on the battlefield. Destabilization risks include AWS proliferation with oppressive regimes and terrorists, a new arms race among state actors, fewer disincentives to start wars on account of reduced numbers of involved soldiers, unpredictable interactions between AWS and their harmful outcomes, cyber-vulnerabilities leading to unintended conflicts, acceleration in the pace of war beyond human reactive abilities.[20]

Diplomatic discussions about the ethical and legal acceptability of granting autonomy to weapons systems are essentially based on arguments (1–5). In this regard, however, it should be noted that, over the years, the debate has been progressively focusing on the so-called "human element", that is to say on the identification of a normatively acceptable human-weapon shared control policy. Notably, a growing consensus developed during the CCW proceedings around the idea that all weapons should be subject to a "meaningful human control" (MHC) and that their use should be regulated accordingly.[21]

But then, what is it that makes human control truly "meaningful"? The ethical and legal principles appealed to in the above arguments—and especially in deontological arguments (1–3)—go a long way towards shaping the content of MHC, by providing criteria that enable one to distinguish perfunctory from truly meaningful human control: human control over weapons systems should ensure compliance with the IHL law of targeting; should avoid responsibility gaps in case of harmful events; should ensure that it is a *moral agent*, and not an artificial one, to take decisions affecting the rights of the human beings involved in an armed conflict.

The application of these broad principles in concrete situations must be facilitated by considering a variety of contextual factors guiding human judgments about the

---

[19] Arkin (2009), pp. 29–36.

[20] Tamburrini (2016), Altmann and Sauer (2017).

[21] The UK-based NGO Article 36 must be credited for putting MHC at the centre of AWS debates by a series of reports and policy papers making the case for MHC over individual attacks as a legal requirement under international law (see the documents available at http://www.article36.org/issue/autonomous-weapons.).

presence of conditions for exercising MHC over weapons systems. These factors include defensive or offensive operational goals, anti-personnel or anti-materiel character of the mission (more generally, the *what* of targeting actions), temporal and spatial frames of the attacks to deliver, dynamical features of the environment and its overall predictability (the *where* of targeting actions), and the perceptual and cognitive capabilities that the weapons system is endowed with (the *how* of targeting actions). By piecing together combinations of these factors, one should be put in a position to evaluate what kind of control would be ethically and legally required on each single use of a weapons system. Following a taxonomy proposed by Noel Sharkey (only slightly modified below),[22] one may sensibly consider five basic types of human-machine interaction, ordered according to decreasing levels of human control and increasing levels of machine control in connection with the critical target selection and engaging tasks:

- (L1) A human engages with and selects targets, and initiates any attack;
- (L2) A program suggests alternative targets and a human chooses which to attack;
- (L3) A program selects targets and a human must approve before the attack;
- (L4) A program selects and engages targets, but is supervised by a human who retains the power to override its choices and abort the attack;
- (L5) A program selects targets and initiates an attack based on the mission goals as defined at the planning/activation stage, without further human involvement.

As we argued elsewhere,[23] the ethical and legal calling for MHC examined above dictates that, as a general default policy, the higher levels of human control (L1 and L2) be exerted on AWS. Under this proviso, lower levels of human control may become acceptable only as internationally agreed on exceptions, provided that the fail-safe, accountability, and moral agency conditions for exercising a genuinely MHC over weapons systems can be actually satisfied at those lower levels of human control. In this way, the residual autonomy of weapons systems, if any, would be purified of its ethically and legally problematic aspects concerning humanly uncontrolled target selection and attacking functions. Defensive systems autonomously targeting incoming missiles and ballistic projectiles are a significant case in which lower levels of human control may be granted without jeopardizing MHC.

2.4 Machines' Autonomy and Human Responsibilities: The Case of Surgical Robots

Ethical and legal motivations for applying MHC over increasing robotic autonomy emerge in medical robotics too. In setting up a hierarchy of autonomy levels for medical robots, Yang and co-authors advanced at the same time the requirement that

[22] Sharkey (2016). Deviations concern, notably, levels L4 and L5.

[23] Amoroso and Tamburrini (2019).

treating physicians should be "still in control to a significant extent".[24] The proposed levels of autonomy (which partly overlap Sharkey's taxonomy) are as follows:

– L0: The robot has no autonomy but only responds to and follows the user's commands;
– L1: The robot provides mechanical assistance, by constraining or correcting human action;
– L2: The robot autonomously carries out tasks that humans designate and supervise;
– L3: The robot generates task execution strategies under human supervision;
– L4: The robot performs an entire medical procedure with human supervision;
– L5: The robot performs an entire medical procedure without human supervision.[25]

In the medical domain of Robot-Assisted Surgery, L0 autonomy systems are used as slave devices for scaling motion, attenuating tremor and enhancing the precision of surgical gestures.[26] The MHC requirement is unproblematically satisfied when these settings are in place. More subtle MHC issues arise at L1–L3.[27]

Various surgical robots deployed in operating rooms are already granted L1 autonomy. A significant case in point are robotic systems assisting surgeons to move the manipulator along desired workspace paths or preventing robotic manipulators from entering selected workspace regions. Robotic systems identifying and applying these active constraints (a.k.a. as Virtual Fixtures) are more than slave devices, as they on occasion correct the surgeon's intended motions. To exert MHC at this autonomy level, one must have the option to override robotic corrections, by means of second-level human control privileges enabling the surgeon to prevail on first-level robotic corrections.

At L2, humans select a task for surgical robots to perform. The surgeon's supervising role consists in hands-free monitoring and possible overriding of robotic execution. Thus, the robotic system is under the surgeon's discrete (rather than continuous) control. The ROBODOC system for orthopaedical surgery is a relatively early example of a system deployed in operating rooms and endowed with L2 autonomy, insofar as it carries out bone milling preoperative plans under human supervision.[28] A more recent research prototype endowed with L2 autonomy is the experimental Smart Tissue Autonomous Robot (STAR) platform, which carries out intestinal suturing (anastomosis) on pig tissue. In experimental tests on this animal

---

[24] Yang et al. (2017).

[25] Be it noted that L4 surgical robots are technologically more distant and L5 ones are currently in the realm of science fiction. Yang et al. (2017).

[26] The da Vinci robotic system for laparoscopic surgery is typically configured as a teleoperated system with L0 autonomy, where surgeons exercise direct control over the entire surgical procedure, including data analysis, preoperative and intraoperative planning, decisions and actual execution. Ackerman (2014).

[27] Ficuciello et al. (2019).

[28] Netravali et al. (2016).

model, STAR was found to outperform expert human surgeons in manual laparoscopic surgery conditions on account of various clinically used suturing metrics.[29]

The ROBODOC and STAR surgical systems are presently characterized by different Technology Readiness Levels. The former system is used for clinical standard procedures, while the latter is still at the research level. This disparity crucially depends on the nature of their respective operational environments and predictability properties. ROBODOC's surgical sites are rigid anatomic structures, whereas STAR operates on deformable soft tissues. The structured environments where ROBODOC operates allow for safe autonomous task execution due to the possibility of making accurate measurements and scene changes predictions. In contrast with this, the soft and deformable surgical sites where STAR operates raise more severe challenges for the accurate detection and tracking of both surgical tools and anatomical parts. These differences in the ROBODOC and STAR operational environments suggest that the human perceptual and cognitive vigilance must be suitably modulated to achieve MHC of individual surgical robots that one brings together under the broad category of L2 autonomous robots. Discrete perceptual sampling and cognitive evaluation of robotic task execution are arguably more demanding in the case of STAR-like systems, in view of scene changes due to physiological blood flow and respiration, and the corresponding need to assess the robot's adaptive response. Accordingly, one size of discontinuous MHC control does not fit all L2 autonomous surgical robots.

L3 autonomous surgical robots generate task strategies under human supervision, and conditionally rely on humans to select from various generated strategies or to approve an autonomously selected strategy. To a limited extent, STAR achieves this level of conditional autonomy as far as anastomosis strategies generation is concerned, along with systems dynamically identifying virtual fixtures and generating optimal control parameters or trajectories.

MHC for L3 autonomy distinctively requires surgeons to decide competently whether to approve one of the robot-generated strategies. This decision presupposes that surgeons understand the rationale for proposed strategies, are in the position to compare their respective merits, and to make up their mind in due time about which strategy to prefer over alternatives. Depending on the complexity of proposed strategies and surgical sites, MHC may incrementally raise human interpretability and decision-making challenges about robot-generated strategies. Similar issues may emerge in connection with strategies that surgical robots may *learn* to propose on the basis of machine learning methods, in view of interpretability and explainability problems affecting learning systems.[30] Today, the learning of surgical strategies is bound to be based on data sets formed by humanly generated strategies. In a more distant future, interpretability and explanation issues arising in the context of MHC for L3 robotic autonomy may become increasingly acute if datasets for learning how to generate intervention strategies progressively shift from data concerning

[29] Shademan et al. (2016).

[30] On this problem, as well as on the attempts to address it, see Chakraborty et al. (2017).

human-generated strategies to robot-generated strategies and corresponding clinical outcomes.

Similarly to the AWS case examined in the previous section, to identify proper MHC policies for robot autonomies one has to consider the functionalities that are appealed to define the tasks assigned to increasingly autonomous surgical robots (the *what* of autonomy), the bodily environments in which these robots operate (the *where* of autonomy), and the system capabilities that are deployed, e.g. learning, to undertake given autonomous actions (the *how* of autonomy). From an ethical standpoint, the identification and application of MHC policies on increasingly autonomous surgical robots are motivated by the bioethical principles of beneficence and non-maleficence,[31] and by the prospective deontological responsibilities of surgeons that these principles entail.

A thorough analysis of prospective responsibilities induced by the MHC requirement is needed to shape training programs for surgeons in Robot-Assisted Surgery. In particular, the non-maleficence bioethical principle requires proper training to provide conceptual tools countervailing positive machine biases, which may wrongly induce human surgeons to trust more what the robot does or proposes to do rather than their own contrasting judgment. Consideration of MHC-related duties plays an equally significant role in evaluating what are the surgeon's retrospective responsibilities, if any, when something goes wrong. Indeed, a surgeon might be held responsible for damages caused by an autonomously performing robot if she failed to exert MHC properly and the harm in question might have been averted had she carefully complied with her MHC duties. By the same token, retrospective responsibility allegations against surgeons for damages caused by an autonomously performing robot might be rebutted and possibly diverted towards other human agents by showing that the specified MHC duties were judiciously complied with.

2.5 Lessons for Judicial Applications of AI

The ethical and legal debate on autonomous robotic systems and the need to exert properly modulated MHC on them may contribute to better frame the discussion on judicial applications of AI in several respects, which we elaborate on in this Section.

2.5.1 Assessing the Normative Landscape for Human-Machine Shared Control in Judicial Affairs

As mentioned in the introductory paragraph, and similarly to what happens with regard to AWS, discussions about the ethical and legal desirability of increasing

[31] Beauchamp and Childress (2013), Chapters 5 and 6.

reliance on AI technologies and their operational autonomy in the courtroom are characterized by deontological and consequentialist arguments pulling into different directions.

In a *deontological perspective*, replacing human judges by algorithms is viewed as a questionable move, insofar as it runs against the moral/ethical prescription to guarantee the fundamental "right not to be subject to a decision based solely on automated processing", especially when such a decision "significantly" affects individuals, which is enshrined in the EU General Data Protection Regulation[32] as well as in the Protocol amending the Council of Europe (CoE) Convention on the Automatic Processing of Personal Data,[33] and which *a fortiori* may be deemed to apply also to decisions having a judicial character.[34]

In a *consequentialist perspective*, the promise of better performances of AI in terms of impartiality[35] and uniformity of decisions[36]—and thus the societal benefits deriving therefrom—provides an argument in favour of its pervasive application in judicial proceedings. Yet, and we may find here another analogy with the AWS debate, favourable consequentialist arguments capture only a fragment of the overall picture (*narrow* consequentialism), as they fail to consider the negative backlashes, on the judicial system as a whole, of systematic resort to AI (*wide* consequentialism).

From a *wide consequentialist perspective*, a less optimistic view of judicial applications of AI ensues from an assessment of present and foreseeable limits of artificial agents' moral and legal autonomy. Indeed, artificial agents (be them robotic or not) are generally endowed with the capability to learn a moral or legal rule from their experience with relevant cases (possibly from the vast amounts of relevant cases processed by means of big data techniques), and to apply the learned rule uniformly to settle new and previously unseen cases. This form of *performative*

[32]European Parliament and Council Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1, 71st preambular paragraph and Article 22 (providing exceptions to this right, but maintaining "the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision" (para 3).

[33]Protocol amending the Council of Europe Convention on the Automatic Processing of Personal Data, Strasbourg 10 October 2018, not yet into force, CETS 203, Article 11. Interestingly enough, the Explanatory Report to the CoE Protocol establishes a connection, which is again reminiscent of the debate on AWS, between this fundamental right and the principle of human dignity, having particular regard to the need to put in place adequate safeguards "when processing personal data, in order for *individuals not to be treated as mere objects*" (Explanatory Report to the Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Strasbourg 10 October 2018, para. 10, emphasis added).

[34]See, in this sense, European Commission for the Efficiency of Justice (CEPEJ) (2018), Appendix I, fn 58.

[35]See, for instance, the way Public Safety Assessment (an algorithm helping judges in deciding on requests for pre-trial release on bail) is advertised by its own developers. https://www.psapretrial.org/about/what-is-psa.

[36]Van Ettekoven and Prins (2018), p. 435.

autonomy of moral or legal kind enables the machine to apply a uniform policy in its decisions. The latter, in other words, is stable over time, insofar as they are based either on explicitly formalized ethical and legal rules implemented by the programmers or on rules extrapolated by machine learning methods from past decisions processed by the machine. What artificial agents cannot presently do (and will plausibly not be able to do, at least for a long time to come) is to autonomously modify these uniform policies based on their own comprehensive judgments about the dynamic evolution of societal beliefs and moral judgments (one may sensibly call *reflective* this form of moral and legal autonomy). Accordingly, while judicial applications of AI may foster the uniformity of decisions and therefore legal certainty, they will also bring about the risk of an undesirable stagnation of the case law,[37] which would become irresponsive to changes in public opinion, due to the lack of reflective autonomy in present and foreseeable AI systems, so failing to fulfil its crucial function to "bridge the gap" between law and society.[38]

2.5.2 Preserving Meaningful Human Control in the Courtroom

Deontological and consequentialist arguments (of both narrow and wide scope) must be properly amalgamated to identify appropriate forms and levels of MHC over AWS, increasingly autonomous surgical robots, and other sorts of artificial agents, including AI systems for use in the courtroom, and more generally judicial applications of AI. A significant landmark in this discussion is provided by the *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment*, adopted by the European Commission for the Efficiency of Justice of the Council of Europe on December 2018 (from now on "Ethical Charter"), whose Principle 5—significantly entitled "under user control"—is aimed at ensuring "that users are informed actors and in control of their choices", by precluding a "prescriptive approach", i.e. one that considers AI decisions as binding and final.[39] One should be careful to note that this deontological principle does not rule out that some operational autonomy of AI systems in the courtroom may still be allowed and beneficial. However, a more fine-grained approach is needed to outline which tasks are admissible for autonomous machine execution, and which functions ought to be prescriptively reserved to human control in relation to artificial agents operating in

[37]See CEPEJ (2018), p. 23 and Nieva Fenoll (2018), pp. 31–32. On the need to avoid the immutability of case law, see *Greek Catholic parish Lupeni and Others v Romania* App Nos 76943/11 (ECtHR [GC], 29 November 2016), para. 116, where it is contended that "a failure to maintain a dynamic and evolutive approach" to case law "would risk hindering reform or improvement", being ultimately detrimental to the proper administration of justice.

[38]On this function of the judiciary, see Barak (2006), pp. 3–19.

[39]CEPEJ (2018), p. 12.

ethically and legally sensitive domains. In other words, one has to give a more precise content to the idea of MHC over judicial applications of AI.

General properties of truly meaningful human control over intelligent machines can be extrapolated from the previous case studies. However, these must be made more specific for the judicial domain. First, human control must afford a *fail-safe mechanism*, where human responsibilities and corresponding control privileges come into play; to prevent (or remedy) erroneous legal interpretations and/or fact assessments that autonomous systems may bring about, in accordance with the principle of good administration of justice, as well as—to the extent that the parties are allowed to contest the machine's decision and to obtain a human revision thereof—with the fundamental right to a fair trial. Second, it must serve as a *catalyst for accountability*, in that it avoids responsibility gaps, and facilitates the distribution of moral and legal responsibilities in case of harmful decisions; which may prove crucial to ascertain the civil and/or disciplinary liability of judges. Third, it must ensure compliance with an overarching ethical requirement, already set forth in the European legal framework on personal data protection: genuine *moral agents*, rather than artificial systems, must be the recognizable (and ultimate) sources of decisions concerning freedom, welfare, and material properties (not to speak of life and death) of persons.[40] This characteristic feature of MHC in the courtroom context might well be viewed as the expression of a fundamental right to have access to *human* justice.[41]

But how to preserve human agency in connection with these various properties, and in front of increasingly autonomous (and complex) artificial agents? The analysis above shows that two problems have to be addressed and solved: (i) how to ensure a proper *quality* of human involvement; (ii) how to establish the kind of human-machine distribution of control privileges that is normatively demanded on ethical and legal grounds.

2.5.2.1 The Quality of Human Control

It is crucial that the human operator does not blindly trust the machine, but takes advantage of AI technology without forfeiting human judgement and critical sense, and without succumbing to so-called automation biases.[42] This is ensured, on the one hand, by training the final users with a view to making them aware of both

[40]See above footnotes 32–34 and the accompanying text. See also Organisation for Economic Co-operation and Development (OECD) (2019), 1.2(b): 'AI actors should implement mechanisms and safeguards, *such as capacity for human determination*, that are appropriate to the context and consistent with the state of art' (emphasis added). More explicitly the French text of the Recommendation states that 'les acteurs de l'IA devraient instituer des garanties et des mécanismes, *tels que l'attribution de la capacité de décision finale à l'homme*, qui soient adaptés au contexte et à l'état de l'art' (emphasis added).

[41]See, in a similar vein, CEPEJ (2018), p. 47.

[42]*Ibid*., p. 23 ('User autonomy must be increased and not restricted through the use of artificial intelligence tools and services').

ascertained and likely limits in the functioning of the artificial systems; and, on the other hand, by putting them in a position to get a sufficient amount of humanly understandable information about machine data processing (*interpretability* requirement), and to additionally obtain an account of the reasons why the machine has taken (or is suggesting or is going to take) a certain course of action (*explainability* requirement).

Meeting the explainability requirement might prove particularly demanding in relation to systems endowed with machine-learning capabilities. Indeed, currently used learning technologies are often based on sub-symbolic data representations and other information processing methods that are not transparent to human users. Notably, deep neural networks are achieving statistically excellent classification and decision-making results, but are mostly unable to fulfil the interpretability and explainability requirements. Moreover, adversarial testing of these learning systems shows that some advanced AI might able to generate excellent decisions (in the case under analysis, solutions to legal problems) in a wide majority of cases; but may occasionally incur into mistakes that no competent human decision-maker would ever make (so-called AI-disasters).[43] The combination of these factors (lack of interpretability for machine decisions and occasional but serious mistakes) does not put human users in the position to understand what happened when the machine goes (badly) wrong. Significantly enough, the demand for AI systems that are capable of providing humanly understandable explanations for their decisions and actions is addressed by and is the focus of the rapidly expanding XAI (eXplainable AI) research area.[44]

The importance of ensuring interpretability and explainability as legal requirements has been well underscored in a recent judgment by the Italian Council of State,[45] where the acceptability of algorithmic decision-making has been scrutinized in relation to a field where the use of AI has already gone a good deal further than the judicial one: that of public administration.[46] The Italian Court, while acknowledging that the automatization of standard procedures through algorithms may be beneficial to the efficiency of public administration, stressed the need to ensure that the artificial decision-maker justifies its choices in terms that are intelligible by citizens and judges. Such a translation of the "algorithmic technical formula" into a "legal rule", in the Court's view, is indeed instrumental to ensure both the transparency of

[43]See, in this regard, the remarkable results by Szegedy et al. (2014) and Athalye et al. (2018), discussed above.

[44]Scientifically challenging issues in XAI are, by no coincidence, central themes of research programs supported by the US Defense Advance Research Project Agency (DARPA). See Dickson (2019).

[45]In the Italian judicial system, the Council of State (*Consiglio di Stato*) is the highest administrative jurisdiction.

[46]Council of State, *Mariateresa Altomare and others v. Ministero dell'Istruzione e della Ricerca and others*, Judgment of 8 April 2019, No. 2270.

the administrative procedures and the effectiveness of judicial review.[47] In a similar vein, but this time with reference to the use of AI by judicial institutions, Principle 4 of the Ethical Charter posits the duty to foster the transparency and understandability of data processing methods.[48]

2.5.2.2 A Normative Approach to Human-Machine Distribution of Control Privileges

The aforementioned training and transparency requirements are crucial to an exercise of human control that is compliant with the MHC requirement. Here we may draw another significant lesson from the analysis of various autonomous robotic systems: one size of human control does not fit all judicial applications of AI. After all, it is intuitively obvious that the issuance of an injunctive relief regarding a small amount of money should not necessarily be subject to the same level of human control required for a decision impinging on personal freedom.

The identification of the level of human control normatively required for each application of AI could be facilitated by the formulation of a set of rules bridging the gap between ethical and legal principles on the one hand, and specific software and their concrete uses on the other hand. These "if-then" bridge rules should be able to express the fail-safe, accountability, and moral agency conditions for exercising *in context* a genuinely MHC in the courtroom.

Analogously to what we already said in relation to both AWS surgical robots, the "if-part" of these rules should include properties concerning *what* task the software is entrusted to, *where* (that is, in which judicial domain) it will be employed and *how* it will perform its tasks. The "what-properties", in particular, must concern the judicial activities that the machine is expected to perform autonomously or support (formal verifications, evidence assessment, elaboration of legal grounds for the decision, and so forth). Unlike robotic systems, the "where-properties" must here be understood as related to the areas of law *where* the machine operates (e.g. small claims, family law, or criminal law) rather than to the physical areas and contexts where the robot operates (battlefields, human body in operating theatres, and so on). The "how-properties", finally, must regard the information processing that the system puts at work to carry out its tasks and that may affect its overall controllability, predictability and explainability. Machine-learning capabilities, which may be increasingly implemented on future legal software, are a significant case in point of one kind of how-property that may raise serious concern from an MHC perspective.

---

[47] *Ibid.*, para 8. See also General Data Protection Regulation (n 32), 71st preambular paragraph ('[data] processing should be subject to suitable safeguards, which should include [. . .] the right [. . .] to obtain an explanation of the decision reached after such assessment and to challenge the decision'). On this point, see the stimulating analysis by Sileno et al. (2018).

[48] CEPEJ (2018).

The "then-part" of bridge rules should establish what kind of human control would be ethically and legally required on each single judicial application of AI. To this end, one may imagine four basic types of human-machine shared control interaction, ordered according to decreasing levels of human control privileges:

– L0. The judicial activity is carried out by a human with the support of a program;
– L1. A program suggests a course of action and a human decides whether to approve it or not. Deviations from the solution recommended by the program must be justified;
– L2. A program issues a binding judicial decision and a human verifies its legality/correctness only if requested by the concerned party;
– L3. A program issues a judicial decision, which is both binding and final.

Against this background, suitable bridge rules should be formulated to establish what level is required to grant the fulfilment of a genuinely *meaningful* human control, as well as the values of the what/where/how properties (or combinations thereof) that justify the identification of some specific level in the above list. To this end, one should take into account (at least) the following observations:

– To the extent that the freedom of human deliberation is unaffected by machine biases, the L0 level of human control should be considered as unproblematic, provided that the training and transparency requirements are met;
– The use of capabilities that may reduce the overall predictability of the software system functioning, such as deep learning-based decision-making (*how-property*), should be treated as a compelling factor pushing towards the application of the higher level (L0) of human control;
– Obviously enough, routine activities (e.g. those concerning the admissibility of evidence; *what-property*) are best candidates for being carried out at lower levels of human control (L1 or L2);[49]
– Decisions impinging on fundamental rights (such as those regarding personal freedom or family relationships; *where-property*) must be taken at L0 (or at most L1) level of human control. L2 human control, instead, should not be considered adequate, in that it would let an artificial agent decide—although provisionally—with respect to individual interests worthy of enhanced protection from both an ethical and legal perspective;
– L3 level of human control should be in principle deemed contrary to the MHC requirement. Exceptions may be allowed in relation to so-called "disposable rights" (*where-property*), provided that there is informed consent by all concerned parties.

[49] Nieva Fenoll (2018), pp. 33–42.

2.6 Concluding Remarks

Principles for moral judgment and action that normative ethical theories make available do not come with a recipe that one applies mechanically to derive ready-made solutions to moral problems. As we saw above, moral choices based on deontological ethics and consequentialism may occasionally come into conflict with each other. Indeed, the pursuit of collective well-being may well motivate, from a consequentialist perspective, the warrant of greater autonomy to artificial systems in ways that are, at least *prima facie*, at odds with deontological imperatives to preserve human decision-making in the execution of certain critical tasks. Therefore, one must think through each moral problem under scrutiny, with the aim of identifying moral norms that appear to be relevant, interpreting them in context and figuring out their situational implications. Suitable prioritization of moral principles and thoughtful compromise is often needed to advance conflict resolution proposals and defuse moral tensions, in order to lay the ground for identifying consistent ethical policies and adopting proper legal regulations.

This intellectual effort is well epitomized by current attempts to understand how to keep AI-based judicial software under MHC (or, in the words of the Ethical Charter, "under user control") without giving up the benefits of automation of decision-making, or at least of certain aspects thereof. In the present contribution, we have tried to show that the debate on autonomous *robotic* systems, such as AWS and surgical robots, may provide useful insights in this regard and may contribute to shape the content of a possible regulation of judicial uses of AI (be it legally binding or soft law).

In the first place, such analysis brings out a three-fold role (fail-safe actor, accountability attractor, and moral agency enactor) that the human operator ought to perform in relation to *any* artificial system operating in legally and ethically sensitive domains, including AI-based judicial software. Also, it enables us to pinpoint distinctive human obligations in the way of human control over judicial artificial agents, which may be differentiated in primary and ancillary ones. Primary obligations concern control functions, with their attending privileges and duties, that must be carried out by human controllers of artificial system, and that no machine should ever be entrusted with. Ancillary obligations (which include training and design requirements) are aimed at ensuring that human-machine partnership conditions are fulfilled for the informed exercise of primary human obligations. Finally, it has been observed that what human controllers of judicial artificial systems must do to fulfil their primary and ancillary obligations depends in significant ways on what task the system is entrusted with, in what legal context it is employed and how it performs its tasks. The what-where-how dimensions of the judicial uses of AI suggest that the MHC formula (or, equivalently in our perspective, the "under user control" principle) does not admit a one-size-fits-all solution, but requires a differentiated approach that is based on the unifying ethical and legal framework that we tried to sketch out above.

References

Ackerman E (2014) New Da Vinci Xi surgical Robot is optimized for complex procedures. IEEE Spectrum

Altmann J, Sauer F (2017) Autonomous weapon systems and strategic stability. Survival 59 (5):117–142

Amoroso D (2020) Autonomous weapons systems and international law. A study on human-machine interactions in ethically and legally sensitive domains. ESI/Nomos, Napoli/Baden-Baden

Amoroso D, Giordano B (2019) Who is to blame for autonomous weapons systems' misdoings? In: Lazzerini N, Carpanelli E (eds) Use and misuse of new technologies. Contemporary challenges in international and European law. Springer, Cham, pp 211–232

Amoroso D, Tamburrini G (2019) What Makes Human Control Over Weapon Systems "Meaningful"? Report submitted by the International Committee for Robot Arms Control (ICRAC) to the GGE on lethal autonomous weapons systems of the CCW, Geneva, 20-21 August 2019 https://www.icrac.net/wp-content/uploads/2019/08/Amoroso-Tamburrini_Human-Control_ICRAC-WP4.pdf

Arkin R (2009) Governing lethal behavior in autonomous robots. CRC Press, Boca Raton

Athalye A et al (2018) Synthetizing robust adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm Sweden, 10-15 July 2018, pp 284–293

Barak A (2006) The Judge in a democracy. Princeton University Press, Princeton

Beauchamp TL, Childress JF (2013) Principles of biomedical ethics, 7th edn. Oxford University Press, Oxford

Buss S, Westlund A (2018) Personal autonomy. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.

Chakraborty S et al (2017) Interpretability of deep learning models: a survey of results. In: IEEE, Smart World Congress, San Francisco, 7–8 August 2017. https://orca.cf.ac.uk/101500/1/Interpretability%20of%20Deep%20Learning%20Models%20-%20A%20Survey%20of%20Results.pdf

Decker M (2008) Caregiving robots and ethical reflection: the perspective of interdisciplinary technology assessment. AI Soc 22(3):315–330

Dickson B (2019) Inside DARPA's effort to create explainable artificial intelligence. TechTalks, 10 January 2019

European Commission for the Efficiency of Justice (CEPEJ), European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment, Strasbourg, 3–4 December 2018

Ficuciello F et al (2019) Autonomy in surgical robots and its meaningful human control. Paladyn 10 (1):30–43

Heyns C (2013) Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions. UN Doc. A/HRC/23/47

International Committee of the Red Cross (2016) Views on autonomous weapon system', paper submitted to the Informal meeting of experts on lethal autonomous weapons systems of the Convention on Certain Conventional Weapons (CCW), Geneva, 11 April 2016

Lin P (2015) Why ethics matters for autonomous cars. In: Maurer M et al (eds) Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte. Springer, Heidelberg, pp 69–85

Metz C (2019) Is ethical A.I. even possible? The New York Times. https://www.nytimes.com/2019/03/01/business/ethics-artificial-intelligence.html

Netravali NA et al (2016) The use of ROBODOC in total hip and knee arthroplasty. In: Ritacco LE et al (eds) Computer-assisted musculoskeletal surgery thinking and executing in 3D. Springer, Cham, pp 219–234

Nieva Fenoll J (2018) Inteligencia Artificial y Proceso Judicial. Marcial Pons, Madrid

Organisation for Economic Co-operation and Development (OECD), Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (22 May 2019)

Schwarzman SA (2019) Can we make artificial intelligence ethical? The Washington Post. https://www.washingtonpost.com/opinions/2019/01/23/can-we-make-artificial-intelligence-ethical/?noredirect=on

Shademan A et al (2016) Supervised autonomous robotic soft tissue surgery. Sci Transl Med 8(337)

Sharkey NE (2016) Staying the Loop: human supervisory control of weapons. In: Bhuta N et al (eds) Autonomous weapons systems: law, ethics, policy. Cambridge University Press, Cambridge, pp 34–37

Sileno G, Boer A, van Engers T (2018) The role of Normware in Trustworthy and Explainable AI. *arxiv.org*

Szegedy C et al (2014) Intriguing properties of neural networks. https://arxiv.org/

Tamburrini G (2016) On banning autonomous weapon systems: from deontological to wide consequentialist reasons. In: Bhuta N et al (eds) Autonomous weapons systems: law, ethics, policy. Cambridge University Press, Cambridge, pp 127–128

The Guardian (2019) The Guardian view on the future of AI: great power, great irresponsibility. The Guardian. https://www.theguardian.com/commentisfree/2019/jan/01/the-guardian-view-on-the-future-of-ai-great-power-great-irresponsibility.

Vamvoudakis K et al (2015) Autonomy and Machine Intelligence in Complex Systems: A Tutorial. In: IEEE (ed) Proceedings of the American Control Conference (1-3 July 2015, Chicago). Curran Associates, New York, pp 5062–5089

van Ettekoven BJ, Prins C (2018) Data analysis, artificial intelligence and the judiciary system. In: Mak V et al (eds) Research handbook in data science and law. Edward Elgar, Cheltenham, p 435

Wiener N (1950) The human use of human beings. Houghton Mifflin, Boston

Wiener N (1964) God & Golem, Inc. – A comment on certain points where cybernetics impinges on religion. MIT Press, Cambridge

Yang GZ et al (2017) Medical robotics – regulatory, ethical, and legal considerations for increasing levels of autonomy. Sci Robot 2(4):2

Case Law

ECtHR [GC], *Greek Catholic parish Lupeni and Others v Romania* App Nos, Judgment of 29 November 2016, 76943/11

Council of State, *Mariateresa Altomare and others v. Ministero dell'Istruzione e della Ricerca and others*, Judgment of 8 April 2019, No. 2270