Introduction to meaningful human control of artificially intelligent systems

Abbink, David; Amoroso, Daniele; Cavalcante Siebert, L.; van den Hoven, M.J.; Mecacci, Giulio; Santoni De Sio, F.

DOI
10.4337/9781802204131.00006

Publication date
2024

Document Version
Final published version

Published in
Research Handbook on Meaningful Human Control of Artificial Intelligence Systems

1. Introduction to meaningful human control of artificially intelligent systems

David Abbink, Daniele Amoroso, Luciano Cavalcante Siebert, Jeroen van den Hoven, Giulio Mecacci and Filippo Santoni de Sio[1]

Artificial intelligence (AI) technology aims to replicate human intelligence and behaviour, and more in general to solve problems while operating at a relatively low dependence on direct human control. This implies that a high level of autonomous capability is desired in these systems. In recent times, the same rapidly growing capabilities of AI systems that can enable breakthroughs in many services and sectors such as transportation and healthcare, have also caused increasing concerns on whether they might spin "out of control": control by individual users, control by the developers or manufacturers, control by non-users, control by other stakeholders and control by society at large. For instance, automated vehicles could behave unpredictably and create risks for their passengers and other road users alike. Decision support systems such as AI-based recruitment systems could steer human decision-making in ways that are undesirable and possibly harmful. AI-supported medical tools could amplify existing biases in patient diagnosis or treatment. War drones could engage targets without the explicit and full consensus of a responsible human agent.

From a slightly different angle, in case of harmful events involving these increasing autonomous capabilities of AI systems, the clear and unambiguous attribution of moral responsibility and legal liability to a person or a group may become complicated. For example, should the organization employing the AI system have prevented the system errors leading to the harmful event? Should the system have been designed or deployed differently? Should the end-user have intervened? Should legislation have prevented its use in the first place? AI systems, especially when designed to function with high levels of autonomous capabilities, are prone to cause what some have called "responsibility gaps": situations where it is inherently difficult to attribute responsibility to human agents for undesired events (Santoni de Sio & Mecacci, 2021).

There are several ways in which AI systems increasingly challenge human control and responsibility attribution.

First, *a mismatch in decision-making speed* can be a concern: AI systems can consider large amounts of data very rapidly, in some cases beyond human capabilities to responsibly intervene when necessary. For example, a driver of a partially autonomous vehicle might not have enough time to react to an unexpected situation where the vehicle's automation is not reliable and in need of relinquishing control. Similarly, the speed and amount of data considered by recommender systems and automated decision-making is sometimes simply too fast for human monitoring or oversight capacities. A pilot of an unmanned aerial vehicle might be challenged to make a split-second decision within a limited window of opportunity of striking

a target, in conditions of partial uncertainty, thereby heavily relying on an intelligent system's recommendation.

Second, the *limited transparency of the AI system* and its functioning may confound human control. Many AI systems, first and foremost those based on deep neural networks, are designed in such a way that it is, in principle, hard to understand how a certain system action or output is produced. This can hinder the possibility of establishing accountability, and obscure the roles played by the numerous human agents in causing a possible malfunction. In other words, many AI systems are inherently hard to be entirely understood even for their designers, which heightens the risks of unpredictable behaviour.

Third, in addition to aspects that are rooted in the very way an AI system operates, *the process and sociotechnical context in which AI systems are developed and deployed* can challenge human control at a different level (Cavalcante Siebert et al., 2023). This context can be highly complex and composed of multiple stakeholders with different interests and perspectives. Technical questions, such as how to maximize control in human–machine interaction, often become entangled with normative ones focusing on ethical and legal aspects of desirability and compliance with rules and values (Liscio et al., 2022). For instance, when developing strategies for human–machine interaction within the realm of automated vehicles, it is necessary to comply with traffic regulations, consider the manufacturer's governance protocols, integrate a nuanced comprehension of human capabilities, and recognize the potential duties associated with reducing road accidents (Heikoop et al., 2019; Beckers et al., 2022). This complexity, per se hardly avoidable in any form of technological innovation with high stakes and a major societal impact, leads to further decreased (social) control and, consequently, the potential emergence of undesirable responsibility gaps.

These three complicating factors related to appropriate human control and responsibility over AI systems are a few of the most prominent issues being addressed by scholars from different disciplines, ranging from law to engineering, design and philosophy. Philosophers have been working at the very concepts of control and responsibility, investigating the conditions that make both of them possible and desirable; lawyers have been exploring normative frameworks that add important responsibility-preserving conditions to control situations (Restrepo Amariles & Baquero, 2023; Amoroso & Tamburrini, 2021); and designers and engineers have been investigating systems that are understandable and controllable by humans and human–machine interfaces aimed at promoting human involvement when and how it matters (Abbink et al., 2018; Hadfield-Menell et al., 2016).

Of course, human control should not be overrated either. Relaxing control does not necessarily lead to malfunctions or responsibility gaps. In fact, increasing the independence of AI technologies can enhance safety. These systems have the potential to outperform humans in terms of speed and accuracy while remaining immune to certain human vulnerabilities and biases. Consider, for instance, the issue of drunk-driving accidents: it is all too obvious that this would not be a problem for an AI-enabled self-driving car. Thus, it is crucial to strike a balance between the reasons favouring human control and those advocating for providing AI systems with more autonomous capabilities in given contexts.

This *Handbook* will present the first encompassing outlook on these issues, by analysing them through the prism of the concept of "meaningful human control" (MHC). In doing so, it incorporates three main disciplinary perspectives: philosophy and ethics (Part I), law and governance (Part II) and design and engineering (Part III). In addition, cross-cutting aspects of MHC over AI systems are discussed through interdisciplinary and systemic perspectives

(Part IV). Since different application scenarios entail different requirements for control, and present at least partially different problems and context-specific approaches, the disciplinary perspectives discuss the topic of MHC concerning four main fields of application: (i) automated intelligent mobility; (ii) recommender and decision-support systems; (iii) AI and robots in cure and care; and (iv) AI in the military.

The rest of this introduction presents the history, content and dimensions of MHC and offers an overview of the themes of the book.

The aim of the *Handbook* is not to unify the debate or, even less so, to defend or settle one particular notion of MHC. Rather, the aim is to combine bottom-up insights from many different perspectives in a single handbook, offering the reader a rich tapestry of different perspectives. Hence, this *Handbook* can be read in different ways to allow for various perspectives and interests. Figure 1.1 illustrates how the reader can focus on a single discipline, explore interdisciplinary and systemic perspectives, or approach a single application field through the lenses of the three disciplines (dashed lines). Nevertheless, other connections and approaches that take a more radical transdisciplinary approach are also possible, as will be pointed out in the following subsections.

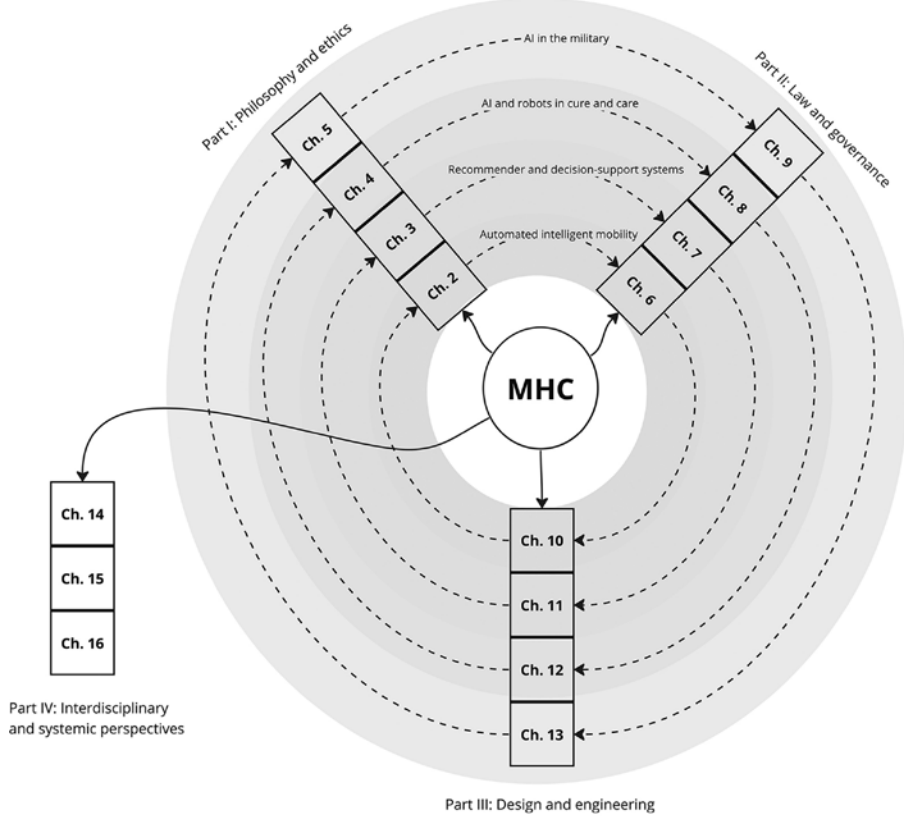

Figure 1.1 Visual outline of and possible ways to approach this handbook

1.1 RECENT HISTORY OF "MEANINGFUL HUMAN CONTROL"

The concept of "meaningful human control" (MHC) was originally proposed, in the early 2010s, in the context of the debate on autonomous weapon systems (AWS), i.e. those weapons systems that, thanks to (more or less) advanced AI technologies, are able to, once activated, select and engage targets without human intervention (such as the Super aEgis II sentry robot or, in some future, swarms of armed microdrones). The UK-based NGO Article 36 must be credited for putting the term "meaningful human control" under the limelight by circulating, since 2013, a series of reports and policy papers making the case for establishing MHC over individual attacks as a legal requirement under international law (Article 36, 2013).

The introduction of the concept of MHC in the AWS debate was grounded on two basic premises: first, it is ethically and legally problematic to let machines apply (possibly lethal) force "without any human control whatsoever"; second, and more relevantly, not every form of human control would be normatively satisfactory, since "a human simply pressing a 'fire' button in response to indications from a computer, without cognitive clarity or awareness, is not sufficient to be considered 'human control' in a substantive sense" (Roff & Moyes, 2016).

Unlike the call for a pre-emptive ban on AWS, the MHC formula was promptly met with favour by a substantial number of States. This response is explainable by at least three converging reasons. First, focusing on the requirement of control would, at least in principle, leave open the possibility of including some autonomous capabilities in some military systems – provided the controllability requirement is satisfied – rather than just trying to ban any form of autonomous technology. Second, human control is a relatively easily *understandable* concept, which "is accessible to a broad range of disciplines, governments and publics regardless of their degree of technical knowledge": it, therefore, provided the international community with a "common language for discussion" (UNIDIR, 2014). Third, the notion of MHC allowed contracting parties to avoid the old and intractable issue of defining machine "automation" and "autonomy" (not to speak of the ever-elusive concept of AI itself). The MHC debate invites to focus on a mostly *normative* problem, i.e., the identification of the forms, ways and levels of human control that ought to be exerted on weapon systems to achieve or maintain acceptable levels of safety, responsibility and other relevant values. From this perspective, it was underlined that one should not look at MHC necessarily as a "solution", as it rather indicates the right "approach" to cope with ethical and legal implications of autonomy (UNIDIR, 2014). What meaningful human control over AI systems precisely means and entails, as we will see, remains an open question.

1.2 MEANINGFUL HUMAN CONTROL BEYOND AUTONOMOUS WEAPON SYSTEMS

The debate on MHC over AWS somehow captures a broader ethical and societal issue in the development and use of systems with autonomous capabilities in sensitive domains. Over and above the autonomous delivery of lethal (or otherwise destructive) force, some key aspects of MHC also apply to most artificial systems executing tasks that may affect individual and/or collective interests. In this *Handbook*, chapters will cover four fields of applications (including but not limited to the military) that call for ethical and legal attention, namely: (i) automated

intelligent mobility; (ii) recommender and decision-support systems; (iii) cure and care robots; and (iv) AI in the military.

AI applications in these domains deal with various relevant *technical*, *legal* and *philosophical* questions that need to be addressed. *Technical questions* include whether certain functions to be assigned to AI systems require, to be properly carried out, human judgement and how these systems can be designed to effectively respect these judgements as well as endorse a number of desired societal values and human rights. *Legal* questions include the issue of determining how to allocate responsibility if harmful events caused by the machine occur (think, for instance, of damages arising from surgical robots' mishaps). A *philosophical* question – from the perspective of *deontological ethics* – is whether it is morally acceptable to remove human agency from decision-making processes that are likely to impinge on individual rights and duties, as well as on relationships that are ethical in character (like that of nursing care). From an opposite direction, via a *consequentialist* perspective in normative ethics, one concern regards the opportunity, or perhaps even the duty, to replace human operators with autonomous machines, whenever the latter's performances promise to ensure a better protection of the interests at stake (e.g., by reducing the number of road accidents).

The cross-cutting nature of these issues, which all address the general problem of the role that human agents should maintain in increasingly automated systems, explains why the language of MHC has recently popped up well beyond debates on AWS. Also, while not making explicit reference to the notion of MHC, recent international policy documents directly address some of the underlying questions related to the topic as well (e.g., CEPEJ, 2018; OECD, 2019). Such documents and existing ethical and legal handbooks on AI focus on mapping problems and challenges. By contrast, this *Research Handbook* represents an initial stepping stone for an encompassing view on how to understand, regulate, design and develop AI that remains under MHC. We believe this *Handbook* will be of interest for a range of different readers: academic scholars from philosophy, law and engineering, and others who want to deepen their understanding of the relation between AI and human control; policymakers and legislators who want to create concrete policies and rules to realize "human-centric AI"; and designers, computer scientists and engineers who want to engage in designing for MHC.

1.3 AN OLDER PHILOSOPHICAL STORY: CONTROL, TECHNOLOGY AND MORAL RESPONSIBILITY

Whereas the term "meaningful human control" is new, some of its inspiring ideas are quite old. On the one hand, MHC is connected to the so-called challenge of determinism to free will and moral responsibility in Western modern philosophy. How can individual people be "really" free or in control of their actions – as opposed to just having the illusion of being free – if their behaviour is determined either by the will of God, deterministic physical laws, human genes, the mechanisms of the human brain, or the social environment? Long before present-day concerns about AI and its MHC, the general concern that determinism in its various forms may erode or eliminate human control had fed "the spectre of creeping exculpation" (Dennett, 2004). Moral philosophers have reacted by sharpening their definition of "moral freedom", "moral agency" and "moral control" to make sense of and better define the conditions for a fair attribution of moral control and moral responsibility in a complex world shaped by natural, social and technological forces. Theories of moral responsibility are a key philosophical point

of reference in present-day philosophical approaches to MHC over AI (see Santoni de Sio & van den Hoven, 2018).

On the other hand, MHC follows from the debate over "technological determinism", the doctrine according to which technology is the key force shaping human life and social relations. While it remains a controversial issue whether Karl Marx himself was a technological determinist, this aphorism by him nicely captures the spirit of the doctrine: "The hand-mill gives you society with the feudal lord: the steam-mill, society with the industrial capitalist" (Marx, 1847/1979). In the 20th century, rapid capitalist industrialization and the fear of a military nuclear escalation leading to a tragic military deflagration, reinforced in Western culture the concern that technological development may largely determine the shape of human life and social relations (in undesirable ways). And that the efforts and the goodwill of individuals may be largely inconsequential in this respect. US sociologist of technology Langdon Winner's 1977 book describing this intellectual trend was iconically titled: *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought.*

The awareness of the power of technology in shaping society beyond individual control has prompted the creation of many different shared ethical, legal and governance frameworks for so-called "Responsible Innovation" (see e.g., Stilgoe et al., 2013). In addition, design frameworks were created to keep technological systems aligned to a broader set of societal values, democratic principles and stakeholder interests: "Design for Values" (see van den Hoven et al., 2015). The concept of MHC therefore captures two ideals: the moral ideal of maintaining and reinforcing human individual control and moral responsibility *within a complex (technical) system*, and the broader political ideal of keeping some form of social control *over technological development*, that is pushing back technological determinism. Chapters in this book cover both of these angles.

Some chapters reflect on MHC from the point of view of individual actors involved in specific domains affected by automation. From a philosophical point of view, in Chapter 3, Haselager and Mecacci discuss the overreliance of human decision-makers on decision support systems, and the consequent dilution of control and responsibility. Peeters also takes on an individual actors' perspective in Chapter 4, focusing on MHC as an enabler and promoter of caretakers' virtues in medical contexts. Finally, in Chapter 5, Devitt discusses accountability of military personnel in warfare scenarios. A similar actors-centred perspective is also taken at different points in the legal section of this book to analyse issues of individual liability: one example is Susanne Beck et al. in Chapter 8, discussing legal responsibility distribution in a medical context with AI support. From the engineering perspective, control by individual actors is discussed in Chapter 12 by Ficuciello et al., who consider technical enablers and constraints to maximize accountability in robotic surgery scenarios.

Other chapters take instead a more societal (design, regulatory) perspective, and discuss control and its enablers more broadly. In the philosophy part, Nyholm opens the book in Chapter 2 with a broad discussion on the nature of control itself, building on and expanding one of the leading theories of MHC. From a legal responsibility perspective, Bo discusses in Chapter 9 States' obligations to ensure MHC under international criminal law in the context of autonomous warfare. Calvert et al., in Chapter 10, start the engineering part of the book with a design framework meant to inform vehicle manufacturers and governments on vehicle and traffic system design for MHC.

1.4 CONTROL AND RESPONSIBILITY IN THE LAW

The notion of control is pervasively present in legal disciplines. We may find references to control in as diverse areas of law as tax law, corporate law, international law and criminal law. It is no surprise, therefore, that this notion is relied on for different aims. If one only considers international law, one will run into – for instance – the so-called "effective control test" for the purposes of State responsibility, the relevance of State control in justifying the extraterritorial application of human rights treaties, or the relationship of control that triggers superior responsibility under international criminal law.

Notwithstanding their variety, references to "control" in legal texts and doctrines feature a common thread: they conceive of it as a source of responsibilities for the (natural or legal) person held to be "in control" (Fischer and Ravizza, 1998). Responsibility can be either forward-looking, in that it creates duties upon the control-bearer, or backward-looking, to the extent that it ensures her accountability in case of wrongful harm.

This nexus between control and responsibility is clearly present in the requirements of MHC as well. On the one hand, from a forward-looking perspective, MHC aims to keep humans in a position to prevent harmful mistakes by AI systems – a "fail-safe" role that is expressly recognized as the main goal of the "human oversight" requirement under Article 14 of the draft EU AI Act (European Commission, 2021). On the other hand, the "meaningfulness" of human control is associated with its ability to function as an "accountability attractor", that is in its capacity to ensure fair venues for human (backward-looking) responsibility in case of wrongful acts (Amoroso and Tamburrini, 2020).

Traditionally, law views relationships of control as situations to which a number of normative, legal consequences are attached. However, control is usually considered, as it were, as something to be defined independently from its normative consequences. This means that one cannot arbitrarily decide or stipulate that an actor is in control only to achieve a desirable normative consequence, for instance, to be able to hold that agent accountable. They must *be* in control according to some independently established criterion, to be *legitimately held* accountable. This is one of the concerns behind the legal debate on MHC over AI; if people are not "really" (i.e., "meaningfully") in control of AI, then they will not be legally accountable for the behaviour of the AI, no matter how socially undesirable this lack of accountability may be.

At the same time, general normative principles can influence the choice of the criteria for defining control in specific, new domains or applications of the law. As seen above, the qualifier "meaningful" is precisely intended to convey its inherently normative dimension. More specifically, "meaningful human control over AI" here means control over AI that is compatible with the criteria for control typically required to attribute duties and responsibility in existing legal domains. So, while the law cannot *arbitrarily* define MHC just to avoid undesirable social consequences, the law must certainly contribute to defining MHC over AI by looking at the specific legal principles in relation to the domain where the concept will be used.

Chapters in this book cover many of these facets in the connection between control and the law. For instance, in Chapter 6, Contissa explores the way(s) the MHC requirement could be shaped in the field of automated intelligent mobilities, so as to ensure that both forward-looking and backward-looking responsibilities of the humans involved are established on a firm legal basis. Chapter 7 by Burri and Juliane Beck, on the other hand, considers the "fail safe" role of humans with regard to the use of recommender and decision-support systems, by zooming in on the abovementioned notion of "human oversight" as per the draft

EU AI Act. The role of MHC as an "accountability attractor" is finally at the core of Chapters 8 and 9. The former, by Susanne Beck, Gerndt, Samhammer and Dabrock, deals with the use of AI technologies in the field of cure and care. Based on empirical research, it shows how MHC can play a key role in reconciling AI-controlled clinical decision support systems with the doctor and patient sovereignty. In the latter, Bo provides an international criminal law account of MHC in relation to the commission of war crimes through the use of AI in the military, with a view to demonstrating the existence, at least as far as international criminal law is concerned, of an international legal obligation to ensure MHC over weapons systems, by additionally pinpointing its key features.

1.5 DESIGNING AND ENGINEERING CONTROL IN THE AGE OF AI

Hundreds of thousands of years ago, humans were already developing tools to help in their daily lives. And from early on, the pursuit of control is there. From ancient tools like spears and hammers to autonomous vehicles, companionship robots and recommender systems, the overarching goal of this pursuit is to ensure that artefacts do "what we want them to do".

The sudden increase in complexity of designed artefacts propelled by the industrial revolution called for a more formal analysis of how to keep them under control. Control theory, an engineering field that deals with the control of dynamical systems, provided a framework on how to understand and influence tools and machines, with huge relevance and applications up to today. In the 1950s, the emergence of the field of AI added yet another layer to the technical discourse around control by introducing some form of "agency" to designed artefacts. Alongside, the field of cybernetics, which is concerned with closed-loop control of complex systems in technological but also biological, cognitive and social systems, emerged as a response to challenges in understanding and designing control. A quote by Norbert Wiener, a key figure in AI and cybernetics, captures the essence of the control challenge in AI: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere, [...] we had better be quite sure that the purpose put into the machine is the purpose which we really desire" (Wiener, 1960).

As AI systems advance, particularly with the rise of deep learning in the 2010s, capabilities that enabled the efficient processing of large amounts of unstructured data and decision-making in complex and dynamic environments were developed, giving rise to even more challenges in maintaining control and responsibility. Widespread media attention has been devoted to speculative doomsday scenarios of AI systems somehow becoming conscious and taking over. Although groups of AI researchers focusing on the future challenge of controlling highly autonomous so-called "superintelligent" AI systems from a technical standpoint have made relevant advances (Russel, 2019; Bostrom, 2014), the issues of control and responsibility are certainly not exclusive to future scenarios. As a matter of fact, today's widespread application and testing in the wild of AI systems – i.e., testing in real-world, uncontrolled and unpredictable environments or situations – have demonstrated negative impacts and harms multiple times, often making it difficult to achieve a fair attribution of responsibility (Cavalcante Siebert et al., 2023). Think, as an example, of the Uber incident in Tempe in 2018: a test vehicle killed a pedestrian on a public road as the technical system did not brake in time and the operator did not intervene to prevent the crash. The legal trial found only the human opera-

tor liable, but authoritative organizations such as the US National Transport and Safety Board explicitly mentioned the lack of a sufficient "culture of safety" in the company as a key cause of the incident. Arguably, the Arizona regulator also bears some responsibility for the incident. Real-world incidents like this one illustrate the pressing challenges of AI control and responsibility, emphasizing the necessity for comprehensive solutions that need to be discussed and agreed upon by all stakeholders.

Chapters 10 to 13 comprise the design and engineering section part of this *Handbook*. They present pioneering approaches to address the multifaceted challenge of ensuring control and responsibility through the design and engineering of AI systems. Chapters 11 and 12 focus on control by interaction i.e., the intricate interplay between human and AI agents in the context of decision support systems and surgical robots, respectively. In Chapter 11, Jonker, Cavalcante Siebert and Murukannaiah propose the concept of self-reflective AI systems and self-reflective hybrid systems (human + AI) with the goal of empowering human moral reasoning in the context of decision support. In Chapter 12, Ficuciello, Hamedani and Tamburrini investigate the implications of different levels of autonomy and MHC on the interaction between surgical robots and surgeons, acknowledging possible tensions and pointing out possible extensions. From a philosophical perspective, in Chapter 3, Haselager and Mecacci discuss how interaction between humans and decision support systems can lead to overreliance and why looking into their organizational embedding might alleviate this issue. From a legal perspective, Chapter 7 by Burri and Juliane Beck looks into the interaction between human and AI systems through the lenses of international law (via the concept of "human control") and the proposed EU Draft AI Act (via "human oversight").

In Chapter 13, van Diggelen, van den Bosch, Neerincx and Steen focus not only on the interaction but also on the macrolevel design options of military human–machine teaming through a thorough sociotechnical analysis that considers multiple phases of the lifecycle of a system and that enables both prior and real-time control by multiple actors. In Chapter 10, Calvert, Johnsen and George delve into the establishment of operational control for automated vehicles through an integrated system proximity framework and operational process design. Taking a multiple-actors perspective, with a keen focus on vehicle designs and governments, this chapter sheds light on the concepts of dynamically updating system design to ensure MHC in the evolution of automated vehicle technologies. From an interdisciplinary and systemic perspective, in Chapter 16, Flemisch et al. sketch a cybernetic model for MHC, starting from a dyadic relationship between a human and an AI system, to an increasingly bigger system of systems, organizations, societies and our global environment.

1.6 INTERDISCIPLINARY AND SYSTEMIC PERSPECTIVES ON MHC

The evolving discourse on control, responsibility and AI also necessitates a holistic and sociotechnical perspective that discusses questions beyond a single disciplinary focus and application field. It is crucial that designers and engineers of AI systems address the challenges of keeping AI systems under MHC as not solely a technical, philosophical or legal issue but an interdisciplinary effort that requires collaboration between technologists, philosophers, social scientists, policy-makers and society at large.

Such broader questions and approaches are also discussed in this *Handbook*. Chapter 14, by Di Nucci, delves into the intriguing question of whether there could be such a thing as "too much control" by exploring the potential negative impact on human people of high levels of technological control in fields such as healthcare and the military. Shifting to a systemic viewpoint, Chapters 15 and 16 offer valuable insights that enable stakeholders to envision a broader picture of achieving and maintaining MHC. In Chapter 15, Pendleton-Jullian presents a novel approach to re-envision the design challenges of a "human-with-AI" future. Rather than human and AI agents doing what they are "best suited for", i.e, a clear but possibly unachievable separation of machine and human, the concept of coevolution is discussed as a possible way forward that allows for a systemic and adaptive evolution. Another contribution of the chapter is a framework for "worldbuilding", a process that aims to ask speculative "what if" questions to instigate discussions and build a "provisional world" with enough texture, complexity and coherence to support the design of future AI systems. In Chapter 16, Flemisch et al. take a systemic approach to conceptualize a comprehensive cybernetic model that intricately maps the critical relationships between MHC and its related concepts over a holistic "big picture" map. This map, represented through a bow-tie diagram, contributes to a better balance between global and local perspectives (e.g., from societies to operators) when designing, engineering and evaluating human–AI systems.

NOTE

1. Authors are listed in alphabetical order.

REFERENCES

Abbink, D. A., Carlson, T., Mulder, M., De Winter, J. C., Aminravan, F., Gibo, T. L., and Boer, E. R. (2018). A topology of shared control systems—finding common ground in diversity. *IEEE Transactions on Human-Machine Systems*, *48*(5), 509–525.

Amoroso, D., and Tamburrini, G. (2020). Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues. *Current Robotics Reports*, *1*, 187–194.

Amoroso, D., and Tamburrini, G. (2021). The Human Control Over Autonomous Robotic Systems: What Ethical and Legal Lessons for Judicial Uses of AI?, In X. Kramer et al. (eds), *New Pathways to Civil Justice in Europe. Challenges of Access to Justice* (pp. 23–42). Cham.

Article 36. (2013). *Killer Robots: UK Government Policy on Fully Autonomous Weapons*. https://article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf accessed 1 March 2024.

Beckers, N., Cavalcante Siebert, L., Bruijnes, M., Jonker, C., and Abbink, D. (2022). Drivers of partially automated vehicles are blamed for crashes that they cannot reasonably avoid. *Scientific Reports, 12*(1), 16193.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Cavalcante Siebert, L.C., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., … and Lagendijk, R. L. (2023). Meaningful human control: actionable properties for AI system development. *AI and Ethics, 3*(1), 241–255.

Dennett, D. C. (2004). *Freedom Evolves*. Penguin UK.

European Commission for the Efficiency of Justice (CEPEJ). (2018). *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment*. Council of Europe. https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c accessed 1 March 2024.

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and*

Amending Certain Union Legislative Acts. Brussels, 21/04/2021, COM (2021) 206 final. 2021/0106 (COD). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 accessed 1 March 2024.

Fischer, J. M., and Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative inverse reinforcement learning. In D. Lee, et al. (Eds.), *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. NeuroIPS.

Heikoop, D. D., Hagenzieker, M., Mecacci, G., Calvert, S., Santoni De Sio, F., and van Arem, B. (2019). Human behaviour with automated driving systems: a quantitative framework for meaningful human control. *Theoretical Issues in Ergonomics Science, 20*(6), 711–730.

Liscio, E., van der Meer, M., Cavalcante Siebert, L., Jonker, C. M., and Murukannaiah, P. K. (2022). What values should an agent align with? An empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems, 36*(1), 23.

Marx, K. (1979). *The Poverty of Philosophy*. New York. (Original work published 1847).

Organisation for Economic Co-operation and Development. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 accessed 1 March 2024.

Restrepo Amariles, D., and Baquero, P. B. (2023). Promises and limits of law for a human-centric artificial intelligence. *Computer Law & Security Review, 48*, 105795.

Roff, H. M., and Moyes, R. (2016). *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons*. Briefing Paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons. https://article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf accessed 1 March 2024.

Russel, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books Ltd.

Santoni de Sio, F., and Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*, 1057–1084.

Santoni de Sio, F., and Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 15.

Stilgoe, J., Owen, R., and Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008.

United Nations Institute for Disarmament Research (UNIDIR). (2014). *The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward*. UNIDIR Resource No. 2.

Van den Hoven, J., Vermaas, P. E., and Van de Poel, I. (2015). *Handbook of Ethics, Values, and Technological Design. Sources, Theory, Values and Application Domains*. Springer.

Winner, L. (1977). *Autonomous technology: Technics-out-of-control as a theme in political thought*. MIT Press.

Wiener, N. (1960). Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science, 131*(3410), 1355–1358.