1

Introduction to Applied Philosophy of AI: Foundations, Contexts, and Perspectives

MARTIN HÄHNEL AND REGINA MÜLLER

The increasing integration of artificial intelligence (AI) into our social and everyday practices highlights the necessity of an "applied" philosophy of AI. It raises complex and fascinating questions, including the moral and legal status of AI, the relationship between humans and machines, the status of knowledge and issues of responsibility. These questions are relevant across all societal domains, for example, political decision making, medicine and healthcare, transportation, education and legal systems. The far-reaching impacts of AI-based technologies on individuals and communities underscore the need for careful philosophical consideration. Additionally, interdisciplinary collaboration is crucial in order to address the challenges posed and opportunities offered by AI in an effective, comprehensive, and responsible way. This book will therefore shed light on the topic of AI from a philosophical perspective in the practical areas of epistemology, ethics, politics, and law. This multi-perspective approach is more necessary than ever since the topic of AI can neither be adequately understood by one discipline alone nor selectively as far as individual areas of its application are concerned.

The currently very fast-changing epistemological, ethical, political, and legal conditions have an enormous impact on the responsible implementation of AI. They form the basis for the development of a contemporary and responsible approach to emerging AI technologies. AI issues are interdependent in disciplinary terms and a future, above all normative-ethical, approach to AI can only succeed against the background of taking into account the multi-perspective approach of an applied philosophy of AI presented in this companion.

The approach of an applied philosophy does not sacrifice basic philosophical reflection to the rash prioritization of practice. Applied philosophy sparks critical impulses and transfers them into a coherent approach, with which it is possible to analyze current technological developments in a realistic and problem-conscious way. This companion is methodologically oriented toward an understanding of applied philosophy developed and promoted by Borchers (2014) and Lippert-Rasmussen (2016). Lippert-Rasmussen posits that applied philosophy is distinguished by its consideration of seven general theoretical and practical prerequisites for a thorough analysis of problems and methods within the scope of examining a subject or issue: *activism*, *audience focus on nonphilosophy*, *empirical facticity*, *practicality*, *reflectivity*, *relevance*,

A Companion to Applied Philosophy of AI, First Edition. Edited by Martin Hähnel and Regina Müller.

and *specificity*.[1] Although Lippert-Rasmussen makes important distinctions by introducing illuminating concepts in order to specify what applied philosophy is or ought to be, it remains unclear how these concepts relate to each other. Ultimately, any definition of applied philosophy in the form of a list is not distinctive and procedural enough. Although Lippert-Rasmussen acknowledges reflectivity, he does not show how this could lead to connecting the different models and unfortunately, he does not use his list of items to discuss a specific subject of investigation. It would be interesting to see how this structure could be applied to the issue of AI.

Borchers (2014) contributes another approach to the issue of applied philosophy. Her approach is more procedural and tries to explain the connections between different perspectives in more detail. Unlike Lippert-Rasmussen, Borchers reduces her approach to four main perspectives. (i) *Heterogeneity*, which means that the field of applied philosophy is internally diverse. In addition to the methods used there, internal heterogeneity also includes the self-image with which research is conducted, the orientation of research interests and the topics and questions. (ii) *Interactivity*, which means that there are complex, close interrelationships between applied research and basic philosophical research. Both areas benefit greatly from each other. Without this close connection to basic research, a scientifically reliable and fruitful applied philosophy is not possible. (iii) *Reflectivity,* which refers to the need to foster a perspectival meta-discourse on its methods and its self-understanding, among other things to develop a meta-theory of the application of philosophical thought. (iv) *Operational independence*, which means that applied philosophy is primarily not just an "application" of theories, principles, etc. from basic research, but an independent field of research that develops its own methods, theories, and concepts anew in dealing with its own questions.

In contrast to Lippert-Rasmussen, Borchers provides a kind of programme that gives instructions on how to pursue applied philosophy. In this way, applied philosophy attempts to answer the relevant questions that come from outside using methods developed specifically for this purpose. Borchers views applied philosophy as positively Janus-faced: one side focuses inwards to develop methods, while the other side engages with the public, making topics accessible to a nonphilosophical audience and collaboratively discussing ethical issues and solutions.

However, due to our special subject of investigation, artificial intelligence, some modifications and additions to an approach based on the insights of Lippert-Rasmussen and Borchers are necessary. It is very important that our approach does not take AI as *factum brutum*, but asks in terms of fundamental reflection what form of technology AI is, whether it can be used globally and across application boundaries and contexts, or whether it has the character of a purpose-bound and locally limited tool, what status this technology has in relation to humans and to what extent this technology is particularly epistemic in nature. This shows that our approach to developing methods draws on other philosophical subdisciplines and promotes networking. This also applies beyond the field of philosophy, where an applied philosophy of AI must ensure that concise philosophical conceptual analysis and the concrete, often non-philosophical application perspective remain in dialogue with each other. It aims to comprehend the impact of AI systems on human existence, societal structures and our perceptions of fundamental philosophical notions, such as intelligence, consciousness, and moral responsibility. This is especially pertinent due to the intricate and unintended effects of AI on human rights, personhood, and our concept of the self. The cross-disciplinary field of AI requires the collective efforts of philosophers, computer scientists, ethicists, social scientists, and policy makers, among others, to tackle the intricate issues and challenges that arise from AI's evolution and integration.

Conducting applied philosophy of AI is also essential for developing tailored ethical frameworks, such as ethics-by-design and embedded ethics. It allows for a meta-discourse that is

crucial for contemplating and scrutinizing the philosophical underpinnings of AI discussions. As the demand for applied philosophy increases, so does the necessity to evolve new philosophical methodologies alongside traditional ones. Gimmler et al. (2023) put it this way: "Applied philosophy is a branch of philosophy that applies traditional philosophical concepts, theories and methods to problems originating from situations that arise in practices outside academia itself. This, we believe, is a good starting point for thinking about what applied philosophy is and what it might become. The future of applied philosophy might also include developing new philosophical theories and even methods; a case in point would be empirical philosophy" (ibid., 108).

Applied philosophy effectively reconnects with empirical and nonphilosophical facts, while also tackling key questions and challenges related to AI. This approach equips philosophers with knowledge of nonphilosophical methods (such as empirical and political ones) and nonphilosophers with philosophical insights. Such an exchange of expertise has become increasingly crucial, especially in the field of AI, involving human experts, artificial systems, and the general public. By using the approach of an applied philosophy of AI, the companion responds to the growing need for a normative-ethical discussion on questions relating to AI to not be conducted from an ethical perspective alone but to provide a realistic picture of the general value of AI in the interplay of the approaches gathered in this companion. Consequently, in accordance with the approach of an applied philosophy, the book is not only directed at (applied) philosophers and students of (applied) philosophy but also at practitioners and all stakeholders already working with AI or planning to work with AI who are interested in its epistemological, ethical, political, and legal aspects.

The individual chapters in this companion are characterized for the most part through basic philosophical contributions that emphasize AI's practical relevance. Some have introductory characters; some go deeper into detail. In sum, the diverse chapters represent the different discourses that AI penetrates and the connections that exist between disciplines that all try to understand and analyze AI issues in their own way. In addition to the discussion of individual fields of AI, the basic methodology and central concepts are presented at the beginning. In the main section, currently and hitherto underexposed topics are discussed and problematized from the perspectives of applied epistemology, ethics, and legal and political philosophy. In the last chapter, an outlook on future challenges and problem areas of an applied philosophy of AI is ventured.

The overview article by **Vincent C. Müller** opens the series of contributions. His contribution presents the main topics, arguments, and positions in the philosophy of AI at present (excluding ethics). Apart from the basic concepts of intelligence and computation, Müller suggests that the main topics of artificial cognition are perception, action, meaning, rational choice, free will, consciousness, and normativity. Through a better understanding of these topics, the philosophy of AI contributes to our understanding of the nature, prospects, and value of AI. Furthermore, these topics can be understood more deeply through the discussion of AI, so Müller provides a sketch of how what we call "AI philosophy" provides a new method for an applied philosophy of AI.

Luciano Floridi argues that current ways of thinking about society, economics, law, and politics are based on an outdated "Ur-philosophy" rooted in Aristotelian and Newtonian concepts. This paradigm views society as composed of individual units (people or legal entities) that interact in absolute time and space, focusing on actions as the key drivers of social change. The chapter contends that this framework, while historically useful, is no longer adequate for understanding and addressing the challenges of mature information societies. Instead, Floridi proposes a shift toward a *relational* paradigm, drawing inspiration from

developments in mathematics and physics. This new approach to conceptualizing an applied philosophy of AI emphasizes relationships rather than individual entities as the fundamental building blocks of society. It conceptualizes society as a network of interconnected relations rather than a mechanical assembly of discrete parts. This shift allows for a more flexible, inclusive, and comprehensive analysis of social phenomena, encompassing people, institutions, artifacts, and nature. The transition is necessary to address complex contemporary issues that defy simple, intuitive explanations. The relational approach implies a reconceptualization of political space and time. Political space becomes defined by social relations rather than geographical boundaries, while political time is understood in terms of the temporality of relations rather than absolute chronology.

This methodology and research programme, the development of which has not yet been completed, gives rise to numerous research-relevant and application-oriented questions with major philosophical and ethical implications. Hence, it is widely recognized that technologies, AI included, both bear traces of ourselves as humans and, in turn, influence us in various ways. **Paula Boddington** analyzes how AI impacts how we might gain knowledge of and experience the self, asking how AI in its different forms, and also in how we imagine it, might influence how we even conceptualize "the self." We need to investigate both technology and the self, in both imagined and concrete contexts, noting that conceptions of the self have developed over time, are complex and have multiple ramifications for issues such as agency, self-mastery, and the boundaries between individuals and the world. First, Boddington explores these issues by examining how imagined uses of AI draw upon and mold certain conceptions of the self. Second, she examines how concrete applications of AI impact our understanding of the self and agency, drawing upon work in the social sciences regarding the "digital self." Lastly, she briefly illustrates the complex impacts of AI on how we imagine and experience the self by considering possible uses of AI technology in relation to dementia.

AI systems are increasingly being integrated into our system of epistemic division of labor. This will change the traditional relationship between experts and laypeople in several ways. It is no longer uncommon for human experts to be outperformed by AI systems in certain tasks, and AI is increasingly being used to assist experts in their work or even to replace them altogether. **Rico Hauswald** discusses key issues that have been explored by philosophers working on expertise and epistemic authority, focusing on the definition problem, the identification problem, the deference problem, and the transfer problem, and applies these considerations to AI-based systems that are used for purposes previously reserved for human experts and epistemic authorities. Hauswald debates what changes will result from the emergence of AI and its integration into our system of epistemic division of labor and shows that reference to the philosophy of expertise and epistemic authority can provide valuable insights that can enhance our understanding of our relationship with epistemic AI systems.

Vincent C. Müller proposes that the "black box" becomes the "black box problem" in a context of justification for judgments and actions, crucially in the context of privacy. He suggests distinguishing between two kinds of classic opacity and introducing a third: the subjects may not know what the system does ("shallow opacity"), the analysts may not know what the system does ("standard black box opacity"), or even the analysts cannot possibly know what the system might do ("deep opacity"). If the agents, data subjects as well as analytics experts, operate under opacity, then they cannot provide some of the justifications for judgments that are necessary to protect privacy, e.g., they cannot give "informed consent" or assert "anonymity." It follows that agents in big data analytics and AI often cannot make the judgments needed to protect privacy. So big data analytics makes the privacy problems worse

and the remedies less effective. To close, Müller provides a brief outlook on technical ways to handle this situation.

Much has been made about the opacity of certain AI-based decision systems. Many have argued that in high-stakes decision contexts, a failure to be able to interpret, explain or justify the outputs of such systems results in a failure of our obligations to those over whom we deploy these decision systems. These obligations are typically understood as obligations to provide information to decision subjects (or their proxies) so they may assess whether they have been treated appropriately. Concerns about black box systems have motivated work on so-called "explainable AI," tools and techniques to render black boxes transparent. At the same time, these concerns have been met with skepticism about both the meaning and value of explainability, especially given the opaque nature of much human decision making. In their contribution, **John Basl and David G. Grant** summarize the current state of the debate between explainability proponents and skeptics. The authors then go on to articulate an alternative basis for basing explainability on appealing to duties of consideration – duties decision makers have to ensure that they are reasoning about decision subjects appropriately. Basl and Grant explain how this alternative approach helps address explainability skepticism and orient our thinking about how decision makers ought to integrate AI-based tools into their decision-making processes.

In their contribution, **Oliver Buchholz and Karoline Reinhardt** discuss the complex relationship between the epistemic and political dimensions of AI. Their specific question is how political decisions which are based on machine learning systems could be justified. Although AI is increasingly used in political fields, Buchholz and Reinhardt emphasize the different underlying rationale of politics and AI: politics is about making decisions, whereas machine learning is about making predictions. They reveal the ethical, epistemic, and methodological challenges of using machine learning-based systems for political decisions. For example, because of the opacity of the processes, these systems often fail to meet certain requirements for public justification and therefore their use in political decision making. Buchholz and Reinhardt conclude that while machine learning systems might be epistemically justified in some contexts, their deployment in politically sensitive areas requires cautious and context-sensitive consideration.

Mirjam Faissner, Janne Lenk and Regina Müller understand AI-based systems primarily as epistemic tools due to their ability to analyze vast amounts of data, identify patterns and generate insights that contribute to knowledge acquisition and decision-making processes in various social settings. The authors raise concerns regarding the integration of AI into epistemic practices, especially regarding injustices, and utilize research theorizing injustice within AI-based epistemic systems and epistemic practices. They provide an overview of epistemic injustice and its variations in the context of AI. Then they describe three forms of epistemic injustice in more detail: testimonial injustice, hermeneutical injustice, and contributory injustice, and highlight the relevant characteristics of AI regarding epistemic injustice, aspects regarding training data, the use of categories and systems of classification, opacity, and epistemic fragmentation. Various examples, such as AI-based health apps, algorithmic profiling, and automatic gender recognition, are used by the authors to illustrate how the forms of epistemic injustice they describe manifest in and through algorithmic systems.

Moving from the epistemological to the ethical is marked by a contribution written by **Martin Hähnel and Regina Müller**, who systematize ethical theories for AI, distinguishing between first-order theories and second-order approaches. First-order theories, such as consequentialism, deontology, and virtue ethics, are complex moral conceptualizations that require significant adaptation for their practical application. Second-order approaches, which

are influenced by the normative demands of specific contexts, integrate elements of first-order theories but are not reducible to them. These include regulations and guidelines (hard law, soft law, self-regulation) and value-sensitive design approaches (embedded ethics, ethics-by-design, community-led ethics). Hähnel and Müller emphasize that no single theory can address all ethical questions in the context of AI. The authors also evaluate the combination of these models, discussing their relevance and effectiveness in resolving AI's ethical challenges.

The next three contributions are devoted to the classical first-order theories of ethics: deontology, consequentialism, and virtue ethics. **Thomas M. Powers** explores duty-based ethics in intelligent systems, integrating rules into symbolic and subsymbolic AI. Symbolic AI applies formal logic, while subsymbolic AI refines ethical behavior through data-driven learning. Deontological principles can be intrinsic (built into AI) or extrinsic (externally regulated). Challenges include ensuring fairness, privacy, and mitigating biases, particularly as commercial AI prioritizes profit. Combining both AI approaches may yield optimal ethical frameworks, but urgent discourse is needed to balance ethical AI development with real-world constraints.

Jörg Schroth's contribution explores the applicability and challenges of consequentialism, particularly utilitarianism, as an ethical framework for AI. Consequentialism's focus on the outcomes of actions aligns with AI's potential to evaluate and implement complex decision-making processes aimed at optimizing welfare. However, his analysis highlights significant concerns with the theory's "negative dimension," which permits instrumental harm to achieve optimal outcomes, creating tension with conventional morality. Schroth examines the distinction between direct and indirect act-consequentialism, emphasizing how AI might overcome human cognitive and emotional limitations in applying direct consequentialist principles. Yet, ethical dilemmas, such as balancing welfare with values like dignity, justice, and autonomy, pose challenges to implementing utilitarianism in machine ethics. Schroth argues that no single ethical theory, including consequentialism, can serve as a universally accepted framework for AI, given its polarizing nature and the intrinsic controversies surrounding its principles. Consequently, while consequentialism offers valuable insights, its suitability for guiding AI ethics remains limited, requiring integration with other ethical considerations to address the diverse range of moral questions posed by AI systems.

After deontology and consequentialism, virtue ethics comes under critical scrutiny. **Kathi Beier** explores the connection between virtue ethics and AI, highlighting the diversity of virtue conceptions and current virtue ethical approaches to AI. She provides an overview of the key virtue concepts discussed in contemporary ethical debates on AI, tracing them from ancient perspectives to modern discussions. She then focuses on one specific virtue, honesty, and examines its application to both humans and AI. Finally, she offers some reflections on the potential of applied virtue ethics for AI, whereby her primary goal is to illustrate the diversity of these approaches.

In contrast to the classical ethical theories, feminist approaches to AI are relatively new. Nevertheless, in recent decades feminist approaches have emerged as a leading subfield in the scholarly examination of ethical issues regarding AI. It is informed by a wide range of theoretical and methodological approaches and enriched by scholars from different disciplines. As **Regina Müller** points out, the connection between feminist ethics and AI lies at the intersection of social justice and AI-based systems. Müller emphasizes the importance of feminist thinking on developments in the context of AI, as feminist scholars take a critical look at the social, cultural, and political conditions and effects of such systems. She introduces basic feminist ideas about moral agency and epistemology and some characteristic features that are

shared by feminist theories, such as intersectionality and context sensitivity, and contextualizes these ideas within AI. Therefore, she shows the relevance of feminist approaches in the context of AI, their connections to AI-based systems, and their specific contributions to the surrounding ethical debates.

As social robots are playing an increasingly significant role in human daily life, **Michael T. Dale** discusses how to design them in such a way that humans will respond to them positively and accept them socially. In particular, he considers to what extent reciprocity might be important in human–robot interaction and whether it should be included as a design feature in social robots. Dale uses the lens of evolutionary biology, a perspective that has remained mainly unexplored in robotics literature. He examines what we already know about the evolution of reciprocity in humans and discusses to what extent this knowledge can weigh in on discussions about social robots. He argues that the evolutionary account of reciprocity can be used as a design feature in social robots and claims that social robots should be capable of both direct and indirect reciprocity if we want them to be socially accepted by humans. As we get closer to developing robots that have social capacities at or near the levels of adult humans in the future, Dale claims that models of reciprocity will play an increasingly significant role in research literature and that robot designers, for example, should take reciprocity into account if they want to most effectively enhance human–robot relations.

Antonia Kempkens looks critically at the phenomenon of datafication. Datafication turns individuals into data sources and prevents them from controlling their data. This loss of control seems ethically questionable to Kempkens. Consequently, she argues that ethical considerations should be integrated into the design of datafication. Kempkens chooses Beauchamp and Childress's principles of biomedical ethics, although it is critically discussed whether these bioethical principles would be transferrable to the digital sector. Kempkens finds the principles an important approach in digital ethics because Beauchamp and Childress's principles support all three predominantly used moral theories. In addition, as Kempkens shows, there is a lack of alternative ethical methods suitable for evaluating datafication. Despite the difficulties of their application, Kempkens use the biomedical principles to identify ethical issues in datafication and argues that regarding an ethical design of datafication, the principles of respect for autonomy, beneficence, nonmaleficence, and justice should be realized. Kempkens highlights that this realization depends on informing and empowering data sources and shows, therefore, that despite their problems, the principles can be helpful in ethically designing datafication.

Machine intelligence is also playing a more and more important role in medicine and the healthcare sector. Since medicine is intimately intertwined with value judgments, **Lukas J. Meier** emphasizes that algorithms will come into contact with normative aspects, and if we want to prevent an algorithmic form of medical paternalism, we will need to integrate ethics into medical AI. Meier argues for two steps that are crucial to this effort: first, implementing a moral theory that forms the frame of the ethical calculations, and second, equipping algorithms with input variables that are adequate for conveying context-based value judgments and preferences to the machine. In that respect, he introduces the main moral theories – consequentialism, deontology, and virtue ethics – and assesses how well each of them could be integrated into healthcare algorithms. In the end, Meier suggests a compromise solution based on principlism and describes how medical AI could be designed to take into account the preferences of various stakeholders in healthcare.

There is a growing need to ensure that autonomous artificially intelligent systems are capable of behaving ethically. **Marten H. L. Kaas** argues that virtue ethics, but in particular the normative theory of aretaic exemplarism, can play a central role in cultivating the ethical

behavior of machines. When coupled with the value inherent in and commonplace practice of training AI systems using simulated environments, it may be possible to raise ethical machines by training them to imitate simulated exemplars of moral excellence, like a digital Jesus or a virtual Confucius. This bottom-up approach to implementing machine ethics has advantages over top-down approaches and is similarly not beset by some of the challenges that arise when attempting to use real people as the imitable training set. In short, machines may be able to imitate simulated moral exemplars and thereby exhibit virtuous behavior themselves.

AI also poses unique challenges to traditional theories of trust. **Karoline Reinhardt** explores the relationship between the philosophical concept of trust and AI. Reinhardt presents the multifaceted debates about trust, different understandings of trust, such as interpersonal trust, institutional trust, and trust in technologies, and emphasizes that whether it makes sense to talk about trust in relation to AI depends on which concept of trust we choose to use. In the end, the central questions are whether we can trust AI, what it would mean to do so, and when, if ever, such trust is justified. Reinhardt emphasizes that AI developments challenge us to rethink how trust is conceptualized and managed.

Abootaleb Safdari explores the possibility of empathetic relationships with large language models (LLMs) from the perspective of applied phenomenology. Phenomenologically informed accounts of empathy typically hinge on embodied interaction. Consequently, this chapter mainly investigates whether LLMs can be considered embodied. Following a brief review of phenomenological empathy, the chapter critiques standard accounts of embodiment prevalent in cognitive science and robotics. These accounts often neglect the crucial distinction between objective and the lived body. Safdari then argues that the concept of the lived body, understood as a form of integration achieved through harmonious interaction, provides a more insightful framework. Therefore, an entity capable of establishing a harmonious action–interaction loop with another partner could be considered embodied in this phenomenological sense. Safdari concludes by arguing that LLMs possess the capacity to participate in such harmonious interaction loops, suggesting their potential embodiment and the possibility of empathetic relationships.

Hugo Cossette-Lefebvre considers the emerging third wave of AI ethics. The third wave relies on the idea that AI should be approached structurally to fully appreciate its ethical implications. Two main tendencies emerge in contemporary discourses on just AI in this third wave. First, some insist on the importance of deliberative inclusion for the development, regulation, and deployment of AI. Second, many build on contemporary approaches in social philosophy to ensure that AI is not only consistent with but fosters and promotes social equality and personal autonomy in society. However, he argues that these two tendencies are not necessarily aligned with one another.

Rosalie Waelen and Larissa Bolte analyze narratives of technological progress in debates about AI and claim that we need to overcome false narratives about progress in AI. They give an overview of the history of AI, its contingencies and the critical analyses of the historical and moral-political narratives of progress that surround AI research and development. The authors focus on recent debates on AI and the notion of progress within critical theory to develop a critique of the ideology of technological progress that is promoted by the AI industry. They argue that if critical theory wants to remain relevant to contemporary philosophy and social theory, it needs to start to take seriously (again) the important role of technology, particularly AI, in today's social, political, and economic power dynamics. Furthermore, they argue that a contemporary critique of technology should include, among other things, critical analyses of the historical and moral-political narratives of progress that

surround AI. They promote that society's conception of technological progress should align with a notion of social progress that is determined by the members of society themselves. This call for theorizing technological progress resonates with the understanding of the field of *AI ethics* which is aimed at ensuring that technological progress goes hand in hand with social progress. The authors aim to show not only that a critical theory perspective is valuable to the applied philosophy of AI, but also that AI in general, and technological progress in particular, are important topics for social critiques of contemporary society.

In his contribution, **Mark Coeckelbergh** claims that one way of showing that and how AI is political is by using the concept of power. Applying political and social theory of power to AI, he analyzes how AI relates to power. After using Sattarov's general conceptual framework about power and technology to outline some of the main ways in which AI may impact power, he draws on three theories of power in order to elaborate on some specific relations between AI and power. Using Marxism and critical theory, Foucault and Butler, and a performance-oriented approach, he shows that and how power is exercised through AI. Coeckelbergh concludes that there is a sense in which AI is also always "artificial power" but notes that this is only possible through human technoperformances, which shape what we do and are.

Challenges posed by the integration of artificial intelligent machines into human society are also addressed by **John-Stewart Gordon**. This integration raises the question of whether AI, for example robots, could be entitled to fundamental moral rights and, if so, the grounds upon which such rights could be established. Gordon introduces the concepts of moral status and moral rights, focusing on the notion of fundamental rights. He introduces what fundamental rights entail and differentiates between two categories: rights related to humans and those applicable to machines. Using the example of sex rights, Gordon focuses on the moral conflicts between the fundamental rights of humans and humanoid robots. Based on the ethical approach, Ethics as a Method, Gordon proposes a pluralistic, context-sensitive ethical framework that emphasizes dialogue and practical judgment. By advocating for the recognition of both human and robot rights, he offers a solution through which conflicting fundamental rights can be reconciled. His approach not only aims to resolve the specific issue of sex rights but also serves as a blueprint for addressing future moral issues arising from human–AI interactions, ensuring a form of harmonious coexistence that respects the fundamental rights of all entities, human and nonhuman alike.

The risks of AI underscore the need to address these risks and thus to govern AI collectively and globally as its functions and capabilities continue to evolve. As **Pak-Hang Wong** points out, the global governance of AI is challenging because of the need to balance universal normative standards with the diversity of cultural values. Wong identifies two challenges in his contribution: the need for a normative standard that is both universally accepted and culturally adaptable and the question of the legitimacy of the normative standard for AI ethics and governance at the global level. He defends the view that human rights approaches are uniquely positioned to provide a basis for global governance of AI if their proponents take cultural values seriously and engage with them. He outlines two conceptions of human rights on which the human rights approaches can build, the moral and political conceptions of human rights, and describes how they can genuinely incorporate cultural values into their approaches.

In his contribution, **Markus Furendal** discusses whether bringing AI technology under collective ownership and control is an attractive way of counteracting this development. He presents *justice-based* rationales for collective ownership, such as the claim that, since the training of AI systems relies on a form of enclosing the data commons, the value created by

those systems should be fairly distributed. He also considers *democracy-based* rationales, like the suggestion that collective ownership is needed to ensure democratic control over the way this crucial technology is being developed and deployed. He concludes that the case for shifting to a model with collective ownership is the most compelling when based on the concern that the private model could come to concentrate economic and political power to such a degree that it threatens institutions designed to promote justice and democracy.

David J. Gunkel investigates the opportunities and challenges of extending the moral and legal status of "person" to AI. He introduces the concept of a person as developed in both the philosophical and legal literature on the subject and differentiates the idea of a natural person from a legal one. He then extends these two concepts to AI and performs a cost/benefit analysis of extending personhood to AI and other seemingly intelligent and/or socially interactive technologies. He presents good reasons for, on the one hand, resisting personification of AI and, on the other hand, for applying the category of person to cover these technological innovations. The final part then reviews a number of alternatives that have been proposed to respond to and redress this apparent impasse, providing the framework for an alternative classificatory scheme that can be scaled to the opportunities and challenges of the twenty-first century and beyond.

In the final section on the future of applied philosophy of AI, **Sven Nyholm** takes up the question of human responsibility. He addresses the concern that if we outsource more and more tasks to AI, then gaps in responsibility might open up. Nyholm highlights that when we think about AI and responsibility, we should consider both praise and blame. He therefore discusses key differences between praise and blame – and the criteria for praiseworthiness and blameworthiness – which are relevant when we consider human responsibility in a future where AI technologies do more and more for us. Furthermore, he draws two key distinctions from general discussions about human responsibility: between positive and negative responsibility, and between backward-looking and forward-looking responsibility, leading to a matrix of four different kinds of responsibility: backward-looking negative responsibility, backward-looking positive responsibility, forward-looking negative responsibility, and forward-looking positive responsibility. Nyholm argues that – corresponding to these four kinds of responsibility – there are four possible kinds of gaps in responsibility. Drawing on the asymmetries between praise and blame, also with respect to the normative criteria for praiseworthiness and blameworthiness, Nyholm discusses problems with filling the different kinds of responsibility gaps. He concludes that it is often easier for people to be blameworthy for bad outcomes caused by AI technologies than to be praiseworthy for good outcomes generated by AI technologies.

Catrin Misselhorn debates the idea of artificial moral agents – artificial systems that have the ability to take moral decisions and act on them. While the debate about artificial moral agents has long been confined to the realm of science fiction, it has been gaining increasing momentum in practical contexts due to the enormous progress made by AI over the last few decades. Misselhorn first shows risks that call for artificial moral agents. She then opens the debate about artificial morality, e.g., how artificial systems can be equipped with moral capacities. Misselhorn differentiates between different types of moral agents depending on the kind and degree of moral capacities that they possess and then focuses on "functional moral agents" and discusses LLMs as moral agents. She then discusses the important question of whether it is morally acceptable to delegate moral decisions to artificial systems. Misselhorn's point is that morality is a reciprocal practice between full moral agents who can engage with each other. In the process of coming to a moral decision, they become aware of their mutual interpretation of a situation, their needs, and feelings. This allows them to develop a shared

understanding of what to do. Artificial moral agents, in Misselhorn's view, cannot participate in such a practice and, what is more, delegating moral decisions to them even seems to infringe on moral practice.

Andrea Berber discusses the question of whether AI can be used to improve our moral decision making. She draws on debates on bioenhancement as a valuable source of general insights into the nature and implications of moral enhancement. She introduces some of the most prominent proposals for AI-based moral enhancement and evaluates them against three different parameters for evaluating moral enhancement in general: effectiveness, effects on human autonomy and agency, and influence on human moral skills development. Berber uses three different types of AI-based moral enhancement that are discussed in research literature – substitutive enhancement, value-driven enhancement, and value-open moral enhancement – and analyzes them based on the parameters mentioned above. Based on this analysis, she offers some general insights into the prospects of moral enhancement using AI-based technologies and concludes that we should encourage a diversity of potential methods for AI-based enhancement, as this may be the most methodologically sound approach to making real progress in morally enhancing human beings.

Note

1 According to Lippert-Rasmussen, applied philosophy is driven by the goal of having a tangible impact on the world; changes in commitments and objectives may lead to alterations in methods to achieve these goals (*activism*). However, applied philosophy should not be confused with "engaged philosophy." The primary audience for applied philosophy consists of nonphilosophers (*audience focus on nonphilosophy*). Applied philosophy is heavily informed by empirical evidence, particularly from the empirical sciences, which highlights its interdisciplinary nature (*empirical facticity*). Applied philosophy provides justified answers to specific questions within its relevant branch about what actions should be taken (*practicality*). To refine its profile, applied philosophy requires a meta-discourse on its methods and self-understanding, including the development of a meta-theory for applying philosophical thought (*reflectivity*). Applied philosophy pertains to significant questions of everyday life. This does not mean it must answer these questions but should philosophically explore or be relevant to them (*relevance*). Applied philosophy addresses specific questions within its branch of philosophy, such as metaphysics, epistemology or moral philosophy, establishing principles to apply and explore their implications in nonphilosophical domains (*specificity*).

References

Borchers, D. (2014). Angewandte Philosophie? Versuch einer Orientierung auf unübersichtlichem Terrain. *Applied Philosophy*, *1*: 12–31.

Gimmler, A., Højme, P., & Kristensen, J. B. L. (2023). The varieties of applied philosophy. *Danish Yearbook of Philosophy*, *56*(2): 105–111.

Lippert-Rasmussen, K. (2016). The nature of applied philosophy. In K. Lippert-Rasmussen, K. Brownlee, & D. Coady (eds (Eds.), *A Companion to Applied Philosophy* (pp. 1–17). New York: Wiley.