3

Ethics of AI

Toward a "Design for Values" Approach

Stefan Buijsman, Michael Klenk, and Jeroen van den Hoven

3.1 INTRODUCTION

Artificial intelligence can (help) make decisions and can steer actions of (autonomous) agents. Now that it gets better and better at performing these tasks, in large part due to breakthroughs in deep learning (see Chapter 1 of this Handbook), there is an increasing adoption of the technology in society. AI is used to support fraud detection, credit risk assessments, education, healthcare diagnostics, recruitment, autonomous driving, and much more. Actions and decisions in these areas have a high impact on individuals, and therefore AI becomes more and more impactful every day. Fraud detection supported by AI has already led to a national scandal in the Netherlands, where widespread discrimination (partly by an AI system) led to the fall of the government.[1] Similarly, healthcare insurance companies using AI to estimate the severity of people's illness seriously discriminated against black patients. A correlation between race and healthcare spending in the data caused the AI system to give lower risk scores to black patients, leading to lower reimbursements for black patients even when their condition was worse.[2] The use of AI systems to conduct first-round interviews in recruitment has led to more opacity in the process, harming job seekers' autonomy.[3] Self-driving cars can be hard to keep under meaningful human control,[4] leading to situations where the driver cannot effectively intervene and even

[1] Heikkilä, M. "Dutch scandal serves as a warning for Europe over risks of using algorithms" (2022) *Politico*, March 29. www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

[2] Ledford, H. "Millions of black people affected by racial bias in health-care algorithms" (2019) *Nature*, 574(7780): 608–609.

[3] Aizenberg, E. and Van Den Hoven, J. "Designing for human rights in AI." (2020) *Big Data & Society*, 7(2): 2053951720949566.

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

situations where nobody may be accountable for accidents.[5,6] In all of these cases, AI is part of a socio-technical system where the new technologies interact with social elements (operators, affected persons, managers, and more). As we will see, ethical challenges emerge both at the level of technology and at the level of the new socio-technical systems. This wide range of ethical challenges associated with the adoption of AI is discussed further in Section 3.2.

At the same time, many of these issues are already well known. They come up in the context of AI because it gets integrated into high-impact processes, but the processes were in many cases already present without AI. For instance, discrimination has been studied extensively, as have complementary notions of justice and fairness. Autonomy, control, and responsibility have likewise received extensive philosophical attention. We also shouldn't forget about the long tradition of normative ethical theories, such as virtue ethics, deontology, and consequentialism, which have all reflected on what makes an action the right one to take. AI and the attention it gets provides a new spotlight on perennial moral issues, some of which are novel and have not been encountered by humanity before and some of which are new instances of familiar problems. We discuss the main normative ethical accounts that may apply to AI in Section 3.3, along with their applicability to the ethical challenges raised earlier.

As we argue, the general ethical theories of the past are helpful but at the same time often lack the specificity needed to tackle the issues raised by new technologies. Instead of applying highly abstract traditional ethical theories such as Aristotle's account of Virtue, Mill's principle of utility, or Kant's Categorical Imperative, straightforwardly to particular AI issues it is often more helpful to utilize mid-level normative ethical theories, which are less abstract, more testable and which focus on technology, interactions between people, organizations, and institutions. Examples of mid-level ethical theories are Rawls' theory of justice,[7] Pettit's account of freedom in terms of non-domination,[8] or Klenk's account of manipulation,[9] which could be construed as broadly Kantian, Amartya Sen and Martha Nussbaum's capability approach,[10] which can be construed as broadly Aristotelian, and Posner's economic theory of law,[11] which is broadly utilitarian. These theories already address a specific

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

[6] On the so-called responsibility gap, see also Chapter 6 of this book.

[7] Rawls, J., *Justice as Fairness: A Restatement* (Harvard University Press, 2001).

[8] Pettit, P. *A Theory of Freedom: from the Psychology to the Politics of Agency* (Oxford University Press, 2001).

[9] Klenk, M. "Digital well-being and manipulation online" in C. Burr and L. Floridi (eds.), *Ethics of Digital Well-Being: A Multidisciplinary Perspective* (Cham: Springer, 2020), pp. 81–100. https://doi.org/10.1007/978-3-030-50585-1_4

[10] Robeyns, I. "The capability approach: A theoretical survey" (2005) *Journal of Human Development*, 6(1): 93–117.

[11] Posner, R. A. *Economic Analysis of Law* (Aspen Publishing, 2014).

set of moral questions in their social, psychological, economic, or social context. They also point to the empirical research that needs to be done in order to apply the theory sensibly. A meticulous understanding of the field to which ethical theory is being applied is essential and part of (applied) ethics itself. We need to know what the properties of artificially intelligent agents are, how they differ from human agents; we need to establish what the meaning and scope is of the notion of, for example, "personal data," what the morally relevant properties of virtual reality are. These are all examples of preparing the ground conceptually before we can start to apply normative ethical considerations.

We then need to ensure that normative ethical theories and the consideration to which they give rise are recognized and incorporated in technology design. This is where design approaches to ethics come in (Value-sensitive design,[12] Design for Values[13] and others). Ethics needs to be present when and where it can make a difference and in the form that increases the chances of making a difference. We discuss these approaches in Section 3.4, along with the way in which they relate to the ethical theories from Section 3.3. These new methods are needed to realize the responsible development and use of artificial intelligence, and require close cooperation between philosophy and other disciplines.

3.2 PROMINENT ETHICAL CHALLENGES

Artificial intelligence differs from other technologies in at least two ways. First, AI systems can have a greater degree of agency than other technologies.[14] AI systems can, in principle, make decisions on their own and act in dynamic fashion, responding to the environment they find themselves in. Whether they can *act* and *make decisions* is a matter of dispute, but what we can say in any case is that they can initiate courses of events that would not have occurred without their initiating it. A self-driving car is thus very different from a typical car, even though both are technological artifacts. While a car can automatically perform certain actions (e.g., prevent the brakes from locking when the car has to stop abruptly), these systems lack the more advanced agency that a self-driving car has when it takes us from A to B without further instructions from the driver.

Second, AI systems have a higher degree of epistemic opacity than other technical systems.[15] While most people may not understand how a car engine works,

[12] Umbrello, S. and De Bellis, A. F. "A value-sensitive design approach to intelligent agents" in Roman V. Yampolskiy (eds.), *Artificial Intelligence Safety and Security* (Chapman & Hall/CRC, 2018), 395–409.

[13] Van den Hoven, J., Vermaas, P., and van de Poel, I. eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (Dordrecht: Springer Netherlands, 2015).

[14] List, C. "Group agency and artificial intelligence" (2021) *Philosophy & Technology*, 34(4): 1213–1242.

[15] Durán, J. M. and Jongsma, K. R. "Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI" (2021) *Journal of Medical Ethics*, 47(5): 329–335.

there are engineers who can explain exactly why the engine behaves the way it does. They are also able to provide clear explanations of why an engine fails under certain conditions and can to a great extent anticipate these situations. In the case of AI systems – and in particular for deep learning systems – we do not know why the systems give us these individual outputs rather than other ones.[16,17] Computer scientists do understand how these systems work generally speaking and can explain general features of their behavior such as why convolutional neural networks are well suited for computer vision tasks, whereas recurrent neural networks are better for natural language processing. However, for individual outputs of a specific AI system, we do not have explanations available as to why the AI generates this specific output (e.g., why it classifies someone as a fraudster, or rejects a job candidate). Likewise, it is difficult to anticipate the output of AI systems on new inputs,[18] which is exacerbated by the fact that small changes to the input of a system can have big effects on the output.[19]

These two features of AI systems make it difficult to develop, deploy, and use them responsibly. They have more agency than other technologies, which exacerbates the challenge – though we should be clear that AI systems do not have *moral* agency (and, for example, developments of artificial moral agents are still far from achieving this goal[20]), and thus should not be anthropomorphized and cannot bear responsibility for results of their outputs.[21] In addition, even its developers struggle to anticipate (due to the opacity) what the AI system will output and why. As a result, familiar ethical problems that arise out of irresponsible or misaligned action are repeated and exacerbated by the speed, scale, and opacity that come with AI systems. It makes it difficult to work with them responsibly in the wider sociotechnical system in which AI is embedded, and also complicates efforts to ensure that AI systems realize ethical values[22] as we cannot easily verify if their behavior is aligned with these values (also known as the alignment problem[23]). It is a pressing issue to find ways to embed these values despite the difficulties that AI systems present us with.

[16] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … and Herrera, F. "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI" (2020) *Information Fusion*, 58: 82–115.

[17] Buijsman, S. "Defining explanation and explanatory depth in XAI" (2022) *Minds and Machines*, 32(3): 563–584.

[18] van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. "Evaluating XAI: A comparison of rule-based and example-based explanations" (2021) *Artificial Intelligence*, 291: 103404.

[19] Akhtar, N. and Mian, A. "Threat of adversarial attacks on deep learning in computer vision: A survey" (2018) *IEEE Access*, 6: 14410–14430.

[20] Cervantes, J. A., López, S., Rodríguez, L. F., Cervantes, S., Cervantes, F., and Ramos, F. "Artificial moral agents: A survey of the current status" (2020) *Science and Engineering Ethics*, 26: 501–532.

[21] In this regard, see also Chapter 6 of this book.

[22] Van de Poel, I. "Embedding values in artificial intelligence (AI) systems" (2020) *Minds and Machines*, 30(3): 385–409.

[23] Gabriel, I. "Artificial intelligence, values, and alignment" (2020) *Minds and Machines*, 30(3): 411–437.

This brings us to the ethical challenges that we face when developing and using AI systems. There have already been a number of attempts to systematize these in the literature. Mittelstadt et al.[24] group them into epistemic concerns (inconclusive evidence, inscrutable evidence and misguided evidence) and normative concerns (unfair outcomes and transformative effects) in addition to issues of traceability/responsibility. Floridi et al.[25] use categories from bioethics to group AI ethics principles into five categories. There are principles of beneficence (promoting well-being and sustainability), nonmaleficence (encompassing privacy and security), autonomy, justice, and explicability. The inclusion of explicability as an ethical principle is contested,[26] but is not unusual in such overviews. For example, Kazim and Koshiyama[27] use the headings human well-being, safety, privacy, transparency, fairness, and accountability, which again include opacity as an ethical challenge. Huang et al.,[28] in an even more extensive overview, again include it as an ethical challenge at the societal level (together with, for example, fairness and controllability), as opposed to challenges at the individual (autonomy, privacy, and safety) and environmental (sustainability) level. In addition to these, there are myriad ethics guidelines and principles from organizations and states, such as the statement of the European Group on Ethics (European Group on Ethics in Science and New Technologies, "Artificial Intelligence, Robotics and 'Autonomous' Systems") and EU High-Level Expert Group's guidelines that mention human oversight, technical robustness and safety, privacy, transparency, diversity and fairness, societal, and environmental well-being and accountability. Recent work suggests that all these guidelines do converge on similar terminology (transparency, justice and fairness, non-maleficence, responsibility, and privacy) on a higher level but that at the same time there are very different interpretations of these terms once you look at the details.[29]

Given these different interpretations, it helps to look in a little more detail at the different ethical challenges posed by AI. Such an examination will show that, while overviews are certainly helpful starting points, they can also obscure the relevance of socio-technical systems to, and context-specificity of, the ethical challenges that AI systems can raise. Consider, first of all, the case of generative natural language

[24] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. "The ethics of algorithms: Mapping the debate" (2016) *Big Data & Society*, 3(2): 2053951716679679.

[25] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … and Vayena, E. "AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations" (2018) *Minds and Machines*, 28: 689–707.

[26] Cortese, J. F. N. B., Cozman, F. G., Lucca-Silveira, M. P., and Bechara, A. F. "Should explainability be a fifth ethical principle in AI ethics?" (2022) *AI and Ethics*, 1–12.

[27] Kazim, E. and Koshiyama, A. S. "A high-level overview of AI ethics" (2021) *Patterns*, 2(9): 100314.

[28] Huang, C., Zhang, Z., Mao, B. and Yao, X. "An overview of artificial intelligence ethics" (2022) *IEEE Transactions on Artificial Intelligence*, 4(4): 799–819.

[29] Jobin, A., Ienca, M., and Vayena, E. "The global landscape of AI ethics guidelines" (2019) *Nature Machine Intelligence*, 1(9): 389–399.

processing of which ChatGPT is a recent and famous example. Algorithms such as ChatGPT can generate text based on prompts, such as to compose an email, generate ideas for marketing slogans, or even summarize research papers.[30] Along with many (potential) benefits, such systems also raise ethical questions because of the content that they generate.

There are prominent issues of bias, as the text that such algorithms generate is often discriminatory.[31] Privacy can be a challenge, as these algorithms can also remember personal information that they have seen as part of the training data and – at least under certain conditions – can as a result output social security numbers, bank details, and other personal information.[32] Sustainability is also an issue, as ChatGPT and other Large Language Models require massive amounts of energy to be trained.[33] But in addition to all of these ethical challenges that are naturally derived from the overviews there are more specific issues. ChatGPT and other generative algorithms may produce outputs that heavily draw on the work of specific individuals without giving credit to them, raising questions of plagiarism.[34] The possibility to use such algorithms to help write essays or formulate answers to exam questions has also been raised, as ChatGPT already performs reasonably well on a range of university exams.[35,36] One may also wonder how such algorithms end up being used in corporate settings, and whether this will replace part of the writing staff that we have. Issues about the future of work[37] are thus quickly connected to the rapidly improving language models. Finally, large language models can produce highly personalized influence at a massive scale and their outputs can be used to mediate communication between people

[30] Tabone, W. and de Winter, J. "Using ChatGPT for Human–Computer Interaction Research: A Primer" (2023) www.researchgate.net/profile/Wilbert-Tabone/publication/367284084_Using_ChatGPT_for_Human-Computer_Interaction_Research_A_Primer/links/63ca6066e922c50e99abb2c8/Using-ChatGPT-for-Human-Computer-Interaction-Research-A-Primer.pdf

[31] Hovy, D. and Prabhumoye, S. "Five sources of bias in natural language processing" (2021) *Language and Linguistics Compass*, 15(8): e12432.

[32] Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. "The secret sharer: Evaluating and testing unintended memorization in neural networks" (2019, August) in *USENIX Security Symposium* (Vol. 267).

[33] Bender, E. M., Gebru, T., McMillan-Major, A., and Mitchell, S. "On the dangers of stochastic parrots: Can language models be too big?" (2021, March) in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

[34] Lee, J., Le, T., Chen, J., and Lee, D. "Do language models plagiarize?" (2022) arXiv preprint arXiv:2203.07618.

[35] Choi, J. H., Hickman, K. E., Monahan, A., and Schwarcz, D. "ChatGPT goes to law school" (2023) Available at SSRN. doi:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4335905

[36] Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A., and Chartash, D. "How well does ChatGPT do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment" (2022) *medRxiv*, 2022–12. doi:https://doi.org/10.1101/2022.12.23.22283901

[37] Wang, W. and Siau, K. "Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda" (2019) *Journal of Database Management (JDM)*, 30(1): 61–79.

(augmented many-to-many communication[38]); they raise a peculiar risk of manipulation at scale. The ethical issues surrounding manipulation are certainly related to issues of autonomy. For example, manipulation may be of ethical relevance insofar as it negatively impacts people's autonomy and well-being.[39] At the same time, manipulation does not necessarily impact autonomy, but instead raises ethical issues all on its own; issues that may well be aggravated in their scope and importance by the use of large language models.[40,41] This illustrates our main point in this section, namely that general frameworks offer a good start, but that they are insufficient as comprehensive accounts of the ethical issues of AI.

A second and very different example is that of credit scoring algorithms that help to decide whether someone qualifies for a bank loan. A recent review shows that the more complex deep learning systems are more accurate at this task than simpler statistical models,[42] so we can expect that AI is used more and more by banks for credit scoring. While this may lead to a larger amount of loans being granted, because the risk per loan is lower (as a result of more accurate risk assessments), there are of course also a number of ethical considerations to take into account that stem from the function of distributing finance to individuals. Starting off again with bias, there is a good chance of unfairness in the distribution of loans. AI systems may offer proportionally fewer loans to minorities[43] and are often also less accurate for these groups.[44] This can be a case of discrimination, and a range of statistical fairness metrics[45] has been developed to capture this. This particular case brings with it different challenges, as fairness measures rely on access to group membership (e.g., race or gender) in order to work, raising privacy issues.[46] Optimizing for fairness can also drastically reduce the accuracy of an AI system, leading to conflicts

[38] Cappuccio, M. L., Sandis, C., and Wyatt, A. "Online manipulation and agential risk" in M. Klenk and F. Jongepier (eds.), *The Philosophy of Online Manipulation* (New York, NY: Routledge, 2022), pp. 72–90.

[39] Klenk, 2020.

[40] Klenk, M. and Hancock, J. "Autonomy and online manipulation" (2019) *Internet Policy Review*. Retrieved from https://policyreview.info/articles/news/autonomy-and-online-manipulation/1431

[41] Klenk, M. and Jongepier, F. (eds.). *The Philosophy of Online Manipulation* (New York, NY: Routledge, 2022).

[42] Dastile, X., Celik, T., and Potsane, M. "Statistical and machine learning models in credit scoring: A systematic literature survey" (2020) *Applied Soft Computing*, 91: 106263.

[43] Zou, L. and Khern-am-nuai, W. "AI and housing discrimination: The case of mortgage applications" (2022) *AI and Ethics*, 1–11.

[44] Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. "Delayed impact of fair machine learning." (2018, July) in *International Conference on Machine Learning*, pp. 3150–3158. PMLR.

[45] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. "Algorithmic fairness: Choices, assumptions, and definitions" (2021) *Annual Review of Statistics and Its Application*, 8: 141–163.

[46] Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., and Zhioua, S. "Survey on fairness notions and related tensions" (2022) *EURO Journal on Decision Processes*, 11, 100033, arXiv preprint arXiv:2209.13012.

with their reliability.[47] From a more socio-technical lens, there are questions of how bank personnel will interact with these models and rely on them, raising questions of meaningful human control, responsibility, and trust in these systems. The decisions made can also have serious impacts for decision subjects, requiring close attention to their contestability[48] and institutional mechanisms to correct mistakes.

Third, and lastly, we can consider an AI system that the government uses to detect fraud among social benefits applications. Anomaly detection is an important subfield of artificial intelligence.[49] Along with other AI techniques, it can be used to more accurately find deviant cases. Yeung describes how New Public Management in the Public Sector is being replaced by what she calls New Public Analytics.[50] Such decisions by government agencies have a major impact on potentially very vulnerable parts of the population, and so come with a host of ethical challenges. There is, again, bias that might arise in the decision-making where a system may disproportionately (and unjustifiably) classify individuals from one group as fraudsters – as actually happened in the Dutch childcare allowance affair.[51] Decisions about biases here are likely to be made differently than in the bank case, because we consider individuals to have a right to social benefits if they need them, whereas there is no such right to a bank loan. Some other challenges, such as those to privacy and reliability, are similar, though again different choices will likely be made due to the different decisions resulting from the socio-technical system. At the same time, new challenges arise around the legitimacy of the decision being made. As the distribution of social benefits is a decision that hinges on political power, it is subject to the acceptability of how that power is exercised. In an extreme case, as with the social benefits affair, mistakes here can lead to the resignation of the government.[52] Standards of justice and transparency, like other standards such as those of contestability/algorithmic recourse,[53] are thus different depending on the context.

What we hope to show with these three examples is that the different classifications of ethical challenges and taxonomies of moral values in the literature are certainly valid. They show up throughout the different applications of AI systems

[47] Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., and Chi, E. H. "Understanding and improving fairness-accuracy trade-offs in multi-task learning" (2021, August) in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1748–1757.

[48] Henin, C. and Le Métayer, D. "Beyond explainability: justifiability and contestability of algorithmic decision systems" (2021) *AI & SOCIETY*, 1–14.

[49] Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. "Deep learning for anomaly detection: A review" (2021) *ACM Computing Surveys (CSUR)*, 54(2): 1–38.

[50] Yeung, K. "The new public analytics as an emerging paradigm in public sector administration" (2023) *Tilburg Law Review*, 27(2): 1–32.

[51] Heikkilä, 2022.

[52] Ten Seldam, B. and Brenninkmeijer, A. "The Dutch benefits scandal: A cautionary tale for algorithmic enforcement" (2021) *EU Law Enforcement*, April 30, 2021, https://eulawenforcement.com/?p=7941.

[53] Venkatasubramanian, S. and Alfano, M. "The philosophical basis of algorithmic recourse" (2020) in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 284–293.

and to some extent they present overarching problems that may have solutions that apply across domains. We already saw this for bias across the different cases. Another example comes from innovations in synthetic data, which present general solutions to the trade-off between privacy and (statistical) fairness by generating datasets with the attributes needed to test for fairness, but for fake people.[54] However, even when the solution is domain-general, the task of determining when such a synthetic dataset is relevantly similar to the real world is a highly context-specific issue. It needs to capture the relevant patterns in the world. For social benefits, this includes correlations between gender, nationality, and race with one's job situation and job application behavior, whereas for a bank, patterns related to people's financial position and payment behavior are crucial. This means that synthetic datasets cannot easily be reused and care must be taken to include the context. Even then, recent criticisms have raised doubts that synthetic data do not fully preserve privacy,[55] and thus may not be the innovative solution that we hope for. Overviews are therefore helpful to remind ourselves of commonly occurring ethical challenges, but they should not be taken as definitive lists, nor should they tempt us into easily transferring answers to ethical questions from one domain to another.

Finally, we pointed already to the socio-technical nature of many of the ethical challenges. This deserves a little more discussion, as the overviews of ethical challenges can often seem to focus more narrowly on the technical aspects of AI systems themselves,[56] leaving out the many people that interact with them and the institutions of which they are a part. Bias can come back into the decision-making if operators can overrule an AI system, and reliability may suffer if operators do not appropriately rely on AI systems.[57] Values such as safety and security are likewise just as dependent on the people and regulations surrounding AI systems as they are on the technologies themselves. Without appropriate design of these surroundings we may also end up with a situation where operators lack meaningful human control, leading to gaps in accountability.[58] The list goes on, as contestability, manipulation and legitimacy also in many ways depend on the interplays of socio-technical elements rather than the AI models themselves. Responsible AI thus often involves changes to the socio-technical system in which AI is embedded. In short, even though the field is called "AI ethics" it should concern itself with more than just the AI models in a strict sense. It is just as much about the people interacting with

[54] Nikolenko, S. I. *Synthetic Data for Deep Learning* (2021) Springer Nature, Vol. 174: Springer Optimization and Its Applications (SOIA).

[55] Stadler, T., Oprisanu, B., and Troncoso, C. "Synthetic data–anonymisation groundhog day" (2022) in *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1451–1468.

[56] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. "Fairness and abstraction in sociotechnical systems" (2019, January) in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68.

[57] Schemmer, M., Hemmer, P., Kühl, N., Benz, C., and Satzger, G. "Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making" (2022) arXiv preprint arXiv:2204.06916.

[58] Santoni de Sio and Mecacci, 2021.

AI and the institutions and norms in which AI is employed. With that said, the next question is how we can deal with the challenges that AI presents us with.

3.3 MAIN ETHICAL THEORIES AND THEIR APPLICATION TO AI

The first place to look when one wants to tackle these ethical challenges is the vast philosophical literature centered around the main ethical theories. We have millennia of thinking on the grounds of right and wrong action. Therefore, since the problems that AI raises typically involve familiar ethical values, it would be wise to benefit from these traditions. To start with, the most influential types of normative ethical theories are virtue ethics, deontology, and consequentialism. Normative ethical theories are attempts to formulate and justify general principles – normative laws or principles if you will[59] – about the grounds of right and wrong (there are, of course, exceptions to this way of seeing normative ethics[60]). Insofar as the development, deployment, and use of AI systems involves actions just like any other human activity, the use of AI falls under the scope of ethical theories: it can be done in right or wrong fashion, and normative ethical theories are supposed to tell just *why* what was done was right or wrong. In the context of AI, however, the goal is often not understanding (*why* is something right or wrong?) but action-guidance: what should be done, in a specific context? Partly for that reason, normative ethical theories may be understood or used as decision aids that should resolve concrete decision problems or imply clear design guidelines. When normative ethical theories are (mis-)understood in that way, when they are construed as a decisional algorithm, for example, when scholars aim to derive ethical precepts for self-driving cars from normative theories and different takes on the trolley problem, it is unsurprising that the result is disappointment in a rich and real world setting. At the same time, there is a pressing need to find concrete and justifiable answers to the problems posed by AI and we can use all the help we can get. We therefore aim to not only highlight the three main ethical theories here the history of ethics has handed down to us but also point to the many additional discussions in ethics and philosophy that promise insights that are more readily applicable to practice and that can be integrated in responsible policymaking, professional reflection, and societal debates. Here the ethical traditions in normative ethical theory are like "sensitizing concepts."[61] that draw our attention to particular aspects

[59] Berker, Selim. "The explanatory ambitions of moral principles" 2019 *Noûs*, 53: 904–36.

[60] Dancy, Jonathan, "Moral particularism," in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), https://plato.stanford.edu/archives/win2017/entries/moral-particularism/.

[61] Zerubavel, Eviatar, "Toward a concept-driven sociology: Sensitizing concepts and the prepared mind" in Wayne H. Brekhus, Thomas DeGloma, and William Ryan Force (eds), *The Oxford Handbook of Symbolic Interactionism* (online ed., Oxford Academic, April 14, 2021), https://doi.org/10.1093/oxfordhb/9780190082161.013.10

of complex situations. Following Thomas Nagel, we could say that these theoretical perspectives each champion one particular type of value at the expense of other types. Some take agent relative perspectives into account, but others disregard the individual's perspective and consider the agent's place in a social network or champion a universalistic perspective.

The focus of virtue ethics is on the character traits of agents. Virtue ethicists seek to answer the question of "how ought one to live" by describing the positive character traits – virtues – that one ought to cultivate. Virtue ethicists have no problem talking about right or wrong actions, however, for the right action is the action that a virtuous person would take. How this is worked out precisely differs, and in modern contexts, one can see a difference between, for example, Slote who holds that one's actual motivations and dispositions matter and that if those are good/virtuous then the action was good.[62] On the other hand, Zagzebski thinks that one's actual motives are irrelevant, and that what matters is whether it matches the actions of a hypothetical/ideal virtuous person.[63] In yet another version, Swanton holds that virtues have a target at which they aim[64]: for example, courage aims to handle danger and generosity aims to share resources. An action is good if it contributes to the targets of these virtues (either strictly by being the best action to promote the different targets, or less strictly as one that does so well enough). In each case, virtues or "excellences" are the central point of analysis and the right action in a certain situation depends somehow on how it relates to the relevant virtues, linking what is right to do to what someone is motivated to do.

This is quite different from consequentialism, though consequentialists can also talk about virtues in the sense that a virtue is a disposition that often leads to outcomes that maximize well-being. Virtues can be acknowledged, but are subsumed under the guiding principle that the right action is the one that maximizes (some understanding of) well-being.[65] There are then differences on whether the consequences that matter are the actual consequences or the consequences that were foreseeable/intended,[66] whether one focuses on individual acts or rules,[67] and on what consequences matter (e.g., pleasure, preference satisfaction, or a pluralist notion of well-being[68]). Whichever version of consequentialism one picks, however, it is consequences that matter and there will be a principle that the right action leads to the best consequences.

[62] Slote, M. *Morals from Motives* (Oxford University Press, 2001).
[63] Zagzebski, L. *Divine Motivation Theory* (New York: Cambridge University Press, 2004).
[64] Swanton, C. *Virtue Ethics: A Pluralistic View* (Oxford University Press, 2003).
[65] Sinnott-Armstrong, W. "Consequentialism" in Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (2022) https://plato.stanford.edu/archives/win2022/entries/consequentialism/.
[66] Feldman, F. "Actual utility, the objection from impracticality, and the move to expected utility" (2006) *Philosophical Studies*, 129: 49–79.
[67] Emmons, D. C. "Act vs. rule-utilitarianism" (1973) *Mind*, 82(326): 226–33.
[68] Mulgan, T. *Understanding Utilitarianism* (Routledge, 2014).

The third general view on ethics, namely deontology, looks at norms instead. So, rather than grounding right action in its consequences, what is most important for these theories is whether actions meet moral norms or principles.[69] A guiding idea here is that we cannot predict the consequences of our actions, but we can make sure that we ourselves act in ways that satisfy the moral laws. There are, again, many different ways in which this core tenet has been developed. Agent-centered theories focus on the obligations and permissions that agents have when performing actions.[70] There may be an obligation to tell the truth, for example, or an obligation not to kill another human being. Vice versa, patient-centered theories look not at the obligations of the agent but at the rights of everyone else.[71,72] There is a right to not be killed that limits the purview of morally permissible actions. Closer to the topic of this chapter, we may also think of, for example, a right to privacy that should be respected unless someone chooses to give up that right in a specific situation.

All three accounts can be used to contribute to AI ethics, though it is important to remember that they are conflicting and thus cannot be used interchangeably (though they can be complementary). A philosophically informed perspective on AI ethics will need to take a stand on how these theories are understood, but for here we will merely highlight some of the ways they might be applied. First, we can look at the practices and character of the developers and deployers of artificial intelligence through the lens of virtue ethics. What virtues should be instilled in those who develop and use AI? How can the education of engineers contribute to this, to instill core virtues such as awareness of the social context of technology and a commitment to public good[73] and sensitivity to the needs of others? It can also help us to look at the decision procedure that led to the implemented AI system. Was this conducted in a virtuous way? Did a range of stakeholders have a meaningful say in critical design choices, as would be in line with value sensitive design and participatory design approaches?[74] While it is typically difficult to determine what a fully virtuous agent would do, and virtue ethics may not help us to guide specific trade-offs that have to be made, looking at the motivations and goals of the people involved in realizing an AI system can nevertheless help.

The same goes for consequentialism. It's important to consider the consequences of developing an AI system, just as it is important for those involved in the operation

[69] Alexander, L. and Moore, M. "Deontological ethics" in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/.

[70] Kamm, F. M. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (Oxford University Press, 2007).

[71] Nozick, R. *Anarchy, State and Utopia* (New York: Basic Books, 1974).

[72] Vallentyne, P. and Steiner, H. (eds.). *Left-Libertarianism and Its Critics* (Houndmills: Palgrave, 2000).

[73] Harris, C. E. "The good engineer: Giving virtue its due in engineering ethics" (2008) *Science and Engineering Ethics*, 14: 153–164.

[74] Liao, Q. V. and Muller, M. "Enabling value sensitive AI systems through participatory design fictions" (2019) arXiv preprint arXiv:1912.07381.

of the system to consider the consequences of the individual decisions made once the AI is up and running. Important as it is, it is also difficult to anticipate consequences beforehand and often the more we can still shape the workings of a technology (in the early design stages), the less we know about the impacts it will have.[75] There are, of course, options to redesign technologies and make changes as the impacts start to emerge, and consequentialism rightly draws our attention to the consequences of using AI. The point we want to make here, rather, is that in practice the overall motto to optimize the impact of an AI system is often not enough to help steer design during the development phase.

Deontology is no different in this respect. It can help to look at our obligations as well as at the rights of those who are impacted by AI systems, but deontology as it is found in the literature is too coarse-grained to be of practical assistance. We often do not exactly know what our moral obligations are on these theories, or how to weigh *prima facie* duties and rights to arrive at what we should do, all things considered. The right to privacy of one person might be overruled by someone else's right not to be killed, for example, and deontological theories typically do not give the detailed guidance needed to decide to what extent one right may be waived in favor of another. In short, we need to supplement the main ethical theories with more detailed accounts that apply to more specific concerns raised by emerging technologies.

These are readily available for a wide range of values. When we start with questions of bias and fairness, there is a vast debate on distributive justice, with for example Rawls' Justice as Fairness[76] as a substantive theory of how benefits and harms should be distributed.[77] Currently, these philosophical theories are largely disconnected from the fairness debate in the computer science/AI Ethics literature,[78] but there are some first attempts to develop connections between the two.[79] The same goes for other values, where for example the philosophical work on (scientific) explanation can be used to better understand and perhaps improve the explainability of machine learning systems.[80,81] Philosophical views on responsibility and control have also already been developed in the context of AI, specifically linked to the concept of meaningful human control over autonomous technology.[82] More attention has also

[75] Genus, A. and Stirling, A. "Collingridge and the dilemma of control: Towards responsible and accountable innovation" (2018) *Research Policy*, 47(1): 61–69.

[76] Rawls, 2001.

[77] See also the discussion of Rawls in Chapter 5 of this book.

[78] Kuppler, M., Kern, C., Bach, R. L., and Kreuter, F. "Distributive justice and fairness metrics in automated decision-making: How much overlap is there?" (2021) arXiv preprint arXiv:2105.01441.

[79] Barsotti, F. and Koçer, R. G. "MinMax fairness: From Rawlsian Theory of Justice to solution for algorithmic bias" (2022) *AI & SOCIETY*, 1–14.

[80] Beisbart, C. and Räz, T. "Philosophy of science at sea: Clarifying the interpretability of machine learning" (2022) *Philosophy Compass*, 17(6): e12830.

[81] Buijsman, 2022.

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

been paid to the ethics of influence, notably the nature and ethics of manipulation, which can inform the design and deployment of AI-mediated influence, such as (hyper-)nudges.[83,84] None of these are general theories of ethics, but the more detailed understanding of important (ethical) values that they provide are nevertheless useful when trying to responsibly design and use AI systems. Even then, however, we need an idea of how we go from the philosophical, conceptual, analysis to the design of a specific AI system. For that, the (relatively recent) design approaches to (AI) ethics are crucial. They require input from all the different parts of philosophy mentioned in this section, but add to that a methodology to make these ethical reflections actionable in the design and use of AI.

3.4 DESIGN-APPROACHES TO AI ETHICS

In response to these challenges the ethics of technology has switched, since the 1980s[85] to a constructive approach of integrating ethical aspects already in the design stage of technology. Frameworks such as value-sensitive design[86] and design for values,[87] coupled with methods such as participatory design[88] have led the way in doing precisely this. Here we will highlight the design for values approach, but note that there are close ties with other design approaches to ethics of technology and design for values is not privileged among these. It shares with other frameworks the starting point that technologies are not value neutral, but instead embed or embody particular values.[89] For example, biases can be (intentionally or unintentionally) replicated in technologies, whether it is in the design of park benches with middle armrests to make sleeping on them impossible or in biased AI systems. The same holds for other values, as the design of an engine will strike a balance between cost-effectiveness and sustainability or content moderation at a social media platform realizes values of the decision-makers. The challenge is to ensure that the relevant values are embedded in AI systems and the socio-technical systems of which they are a part. This entails three different challenges: identifying the relevant values, embedding them in systems, and assessing whether these efforts were successful.

When identifying values, it is commonly held important to consider values of all stakeholders, both those directly interacting with the AI system and those indirectly

[83] Klenk and Jongepier, 2022.
[84] Yeung, K. "'Hypernudge': Big Data as a mode of regulation by design" (2017) *Information, Communication & Society*, 20(1): 118–136.
[85] Friedman, B., Kahn, P., and Borning, A. "Value sensitive design: Theory and methods" (2002) *University of Washington Technical Report*, 2: 12.
[86] Umbrello and De Bellis, 2018.
[87] van den Hoven et al., 2015.
[88] Spinuzzi, C. "The methodology of participatory design" (2005) *Technical Communication*, 52(2): 163–174.
[89] Van de Poel, 2020.

affected by its use.[90] This requires the active involvement of (representatives of) different stakeholder groups, to elicit the different values that are important to them. At the same time, it comes with a challenge. Design approaches to AI ethics require that values of a technology's stakeholders (bottom-up) are weighed up against values derived from theoretical and normative frameworks (top-down). Just because people think that, for example, autonomy is valuable does not imply that it is valuable. To go from the empirical work identifying values of stakeholders to a normative take on technologies requires a justification that will likely make recourse to one of the normative ethical approaches discussed earlier. Engaging stakeholders is thus important, because it often highlights aspects of technologies that one would otherwise miss, but not sufficient. The fact that a solution or application would be *de facto accepted* by stakeholders, does not imply that it would be (therefore) also *morally acceptable*. Moral acceptability needs to be independently established, a good understanding of the arguments and reasons that all directly and indirectly affected parties bring to the table is a good starting point, but not the end of the story. We should aim at a situation where technology is accepted, because it is morally acceptable, and that if technologies are not accepted, that is because they are not acceptable.

Here the ethical and more broadly philosophical theories touched upon in the previous section can help. They are needed for two reasons: first, to justify and ground the elicited values in a normative framework, the way, for example, accounts of fairness, responsibility, and even normative takes on the value of explainability[91] can justify the relevance of certain values. Here, it also helps to consider the main ethical theories as championing specific values (per Nagel), be they agent-relative, focused on social relations or universalistic. For these sets of values, these theories help to justify their relevance. Second, they help in the follow-up from the identification of values to their implementation. Saying that an AI system should respect autonomy is not enough, as we need to know what that entails for the concrete system at issue.

As different conceptualizations of these values often lead to different designs of technologies, it is necessary to both assess different conceptions and develop new conceptions. This work can be fruitfully linked to the methods of conceptual engineering[92] and can often draw on the existing conceptions in extant philosophical accounts. Whether those are used or new conceptions are developed, one needs to make the steps from values to norms, and then from norms to design requirements.[93]

[90] Friedman, B., Hendry, D. G., and Borning, A. "A survey of value sensitive design methods" (2017) *Foundations and Trends® in Human–Computer Interaction*, 11(2): 63–125.

[91] Cortese et al., 2022.

[92] Veluwenkamp, H. and van den Hoven, J. "Design for values and conceptual engineering" (2023) *Ethics and Information Technology*, 25(1): 1–12.

[93] Van de Poel, I. "Translating values into design requirements" (2013) *Philosophy and Engineering: Reflections on Practice, Principles and Process*, 253–66.

To give a concrete example, one may start from the value of privacy. There are various aspects to privacy, which can be captured in the conceptual engineering step to norms. Here things such as mitigating risks of personal harm, preventing biased decision-making, and protecting people's freedom to choose are all aspects that emerge from a philosophical analysis of privacy[94] and can act as norms in the current framework. For they, in turn, can be linked to specific design requirements. When mitigating risks, one can look at specific technologies such as coarse graining[95] or differential privacy[96] that aim to minimize how identifiable individuals are, thus reducing their risks for personal harm. Likewise, socio-technical measures against mass surveillance can support the norm for protecting people's freedom to choose, by preventing a situation where their choices are impacted by the knowledge that every action is stored somewhere.

For the actual implementation of values there are a number of additional challenges to consider. Most prominently is the fact that conflicts can occur between different design requirements, which is more often referred to as value conflicts or trade-offs.[97] These already came up in passing in the cases discussed in Section 3.2, such as conflicts between accuracy and fairness or between privacy and fairness. If we want to use statistical fairness measures to promote equal treatment of, for example, men and women, then they need datasets labeled with gender, thus reducing privacy. Likewise, it turns out that when optimizing an AI system for conformity with a statistical fairness measure its accuracy is (greatly) reduced.[98] Such conflicts can be approached in a number of ways[99]: (1) maximizing the score among alternative solutions to the conflict, assuming that there is a way to rank them; (2) satisficing among alternatives, finding one that is good enough on all the different values; (3) respecifying design requirements to ones that still fit the relevant norms but no longer conflict; and (4) innovating, as with synthetic data and the privacy/fairness conflict, to allow for a way to meet all the original design requirements. All of these are easier said than done, but highlight different strategies for dealing with the fact that often we have to balance competing *prima facie* (ethical) requirements on AI systems.

Another problem is that recent work has drawn attention to the possibility of changing values. Perceptions of values certainly change over time. That is, people's

[94] Moore, A. D. "Privacy: its meaning and value" (2003) *American Philosophical Quarterly*, 40(3): 215–27.

[95] Gedik, B. and Liu, L. "Protecting location privacy with personalized k-anonymity: Architecture and algorithms" (2007) *IEEE Transactions on Mobile Computing*, 7(1): 1–18.

[96] Dwork, C. "Differential privacy" in *Automata, Languages and Programming: 33rd International Colloquium*, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II 33 (pp. 1–12). Springer Berlin Heidelberg.

[97] Van de Poel, I. "Conflicting values in design for values," *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (2015), 89–116.

[98] Kozodoi, N., Jacob, J. and Lessmann, S. "Fairness in credit scoring: Assessment, implementation and profit implications" (2022) *European Journal of Operational Research*, 297(3): 1083–1094.

[99] Van de Poel, 2015.

interpretation of what it means for a technology to be sustainable (to adhere to or embody that value) may change over time and people may begin to value things that they did not value before: sustainability is a case in point. That means that, even if people's perceptions of values are correctly identified at the beginning of a design project, they may change, and insofar as people's perceptions of values matter (see above), the possibility of value change represents another methodological challenge for design for value approaches. Actively designing for this scenario, by including adaptability, flexibility and robustness[100] is thus a good practice. We may not be able to anticipate value changes, just as it is hard to predict more generally the impact of an AI system before it is used, but that is no reason not to try to do everything in our power to realize systems that are as responsible as possible.

Because we cannot predict everything, and because values may change over time, it is also important to assess the AI systems once they are in use – and to keep doing so over time. Did the envisaged design requirements indeed manage to realize the identified values? Were values missed during the design phase that now emerge as relevant – the way Uber found out that surge pricing during emergencies is ethically wrong (because it privileges the rich who can then still afford to flee the site of an attack) only after this first happened in 2014.[101] And are there no unintended effects that we failed to predict? All of these questions are important, and first attempts to systematically raise them can be found in the emerging frameworks for ethics-based auditing[102] as well as in the EU AI Act's call for continuous monitoring of AI systems. In these cases, too, the translation from values to design requirements can help. Design requirements should be sufficiently concrete to be both implementable and verifiable, specifying for example a degree of privacy in terms of k-anonymity (how many people have the same attributes in an anonymized dataset) or fairness in terms of a statistical measure. These can then guide the assessment afterward, though we have to be careful that the initial specification of the values may be wrong. Optimizing for the wrong fairness measure can, for example, have serious negative long-term consequences for vulnerable groups[103] and these should not be missed due to an exclusive focus on the earlier chosen fairness measure during the assessment.

In all three stages (identification, implementation, and assessment), we should not forget the observations from Section 3.2: we should design more than just the technical AI systems and what implications values have will differ from context to context. The problem of Uber's algorithm raising prices whenever demand

[100] Van de Poel, I. "Design for value change" (2021) *Ethics and Information Technology*, 23(1): 27–31.

[101] Sullivan, W. "Uber backtracks after jacking up prices during Sydney hostage crisis" (2014) *Washington Post*, December 15, 2014, www.washingtonpost.com/news/morning-mix/wp/2014/12/15/uber-backtracks-after-jacking-up-prices-during-syndey-hostage-crisis/

[102] Mökander, J. and Floridi, L. "Ethics-based auditing to develop trustworthy AI" (2021) *Minds and Machines*, 31(2): 323–27.

[103] Liu et al., 2018.

increases regardless of the cause for that demand was ultimately not solved in the AI system, but by adding on a control room where a human operator can turn off the algorithm in emergencies. Response times were an issue initially,[104] but it shows that solutions need not be purely technical. Likewise, an insurance company in New Zealand automated its claims processing and massively improved efficiency while maintaining explainability when it counts, by automatically paying out every claim that the AI approved but sending any potential rejections to humans for a full manual review.[105] In this case, almost 95% of applications get accepted almost instantaneously, while every rejected application still comes with a clear motivation and an easily identifiable person who is accountable should a mistake have been made. A combination that would be hard to achieve using AI alone is instead managed through the design of the wider socio-technical system. Of course, this will not work in every context. Crucial to this case is that the organization knew that fraudulent claims are relatively rare and that the costs of false positives are thus manageable compared to the saving in manpower and evaluation time. In other situations, or in other sectors such as healthcare (imagine automatically giving someone a diagnosis and only manually checking when the AI system indicates that you do not have a certain illness) different designs will be needed.

To sum up, design approaches to AI ethics focus on the identification of values, the translation of these values into design requirements, and the assessment of technologies in the light of values. This leads to a proactive approach to ethics, ideally engaging designers of these systems in the ethical deliberation and guiding important choices underlying the resulting systems. It is an approach that aims to fill in the oft-noted gap between ethical principles and practical development.[106] With the increasing adoption of AI, it becomes ever more pressing to fill this gap, and thus to work on the translation from ethical values to design requirements. Principles are not enough[107] and ethics should find its way into design. Not only are designs value laden as we discussed earlier, but values are design consequential. In times where everything is designed, commitment to particular values implies that one is bent on exploring opportunities to realize these values – when and where appropriate – in technology and design that can make a difference. We therefore think that we can only tackle the challenges of AI ethics by combining normative ethical theories, and

[104] Cox, J. "London terror attack: Uber slammed for being slow to turn off 'surge pricing' after rampage" (2017) *Independent*, June 4, 2017, www.independent.co.uk/news/uk/home-news/london-terror-attack-uber-criticised-surge-pricing-after-london-bridge-black-cab-a7772246.html

[105] Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. "Algorithmic decision-making and the control problem" (2019) *Minds and Machines*, 29: 555–78.

[106] Georgieva, I., Lazo, C., Timan, T., and van Veenstra, A. F. "From AI ethics principles to data science practice: A reflection and a gap analysis based on recent frameworks and practical experience" (2022) *AI and Ethics*, 2(4): 697–711.

[107] Mittelstadt, B. "Principles alone cannot guarantee ethical AI" (2019). *Nature Machine Intelligence*, 1(11): 501–07.

detailed philosophical accounts of different values, with a design approach.[108] Such an approach additionally requires a range of interdisciplinary, and often transdisciplinary, collaborations. Philosophy alone cannot solve the problems of AI ethics, but it has an important role to play.

3.5 CONCLUSION

Artificial intelligence poses a host of ethical challenges. These come from the use of AI systems to take actions and support decision-making and are exacerbated by our limited ability to steer and predict the outputs of AI systems (at least the machine learning kind). AI thus raises familiar problems, of bias, privacy, autonomy, accountability, and more, in a new setting. This can be both a challenge, as we have to find new ways of ensuring the ethical design of decision-making procedures, and an opportunity to create even more responsible (socio-technical) systems. Thanks to the developments of AI we now have fairness metrics that can be used just as easily outside of the AI context, though we have to be careful in light of their limitations (see also Chapter 4 of this Handbook).[109] Ethics can be made more actionable, but this requires renewed efforts in philosophy as well as strong interdisciplinary collaborations.

Existing philosophical theories, starting with the main ethical theories of virtue ethics, consequentialism and deontology, are a good starting point. They can provide the normative framework needed to determine which values are relevant and what requirements are normatively justified. More detailed accounts, such as those of privacy, responsibility, distributive justice, and explanation, are also needed to take the first step from values that have been identified to conceptualizations of them in terms of norms and policies or business rules. Often, we cannot get started on bridging the gap from values and norms to (concrete) design requirements before we have done the conceptual engineering work that yields a first specification of these values. After that design approaches to AI ethics kick in, helping guide us through the process of identifying values for a specific case, and then specifying them in requirements that can finally be used to assess AI systems and the broader socio-technical system in which they have been embedded.

While we have highlighted these steps here from a philosophical perspective, they require strong interdisciplinary collaborations. Identifying values in practical contexts is best done in collaboration with empirical sciences, determining not only people's preferences but also potential impacts of AI systems. Formulating design requirements requires a close interaction with the actual designers of these systems

[108] van den Hoven, J., Miller, S., and Pogge, T. (eds.). *Designing in Ethics* (Cambridge University Press, 2017). doi:10.1017/9780511844317.

[109] Carey, A. N. and Wu, X. "The statistical fairness field guide: Perspectives from social and formal sciences" (2023) *AI and Ethics*, 3(1): 1–23.

(both technical and socio-technical), relating the conceptions of values to technological, legal, and institutional possibilities and innovations. Finally, assessment again relies heavily on an empirical understanding of the actual effects of socio-technical (and AI) systems. To responsibly develop and use AI, we have to be proactive in integrating ethics into the design of these systems.