AI & SOCIETY (2022) 37:23–37
https://doi.org/10.1007/s00146-021-01148-6

ORIGINAL ARTICLE

Organisational responses to the ethical issues of artificial intelligence

Bernd Carsten Stahl[1] · Josephina Antoniou[2] · Mark Ryan[3] · Kevin Macnish[4] · Tilimbe Jiya[5]

Received: 23 May 2020 / Accepted: 20 January 2021 / Published online: 16 February 2021

Abstract
The ethics of artificial intelligence (AI) is a widely discussed topic. There are numerous initiatives that aim to develop the principles and guidance to ensure that the development, deployment and use of AI are ethically acceptable. What is generally unclear is how organisations that make use of AI understand and address these ethical issues in practice. While there is an abundance of conceptual work on AI ethics, empirical insights are rare and often anecdotal. This paper fills the gap in our current understanding of how organisations deal with AI ethics by presenting empirical findings collected using a set of ten case studies and providing an account of the cross-case analysis. The paper reviews the discussion of ethical issues of AI as well as mitigation strategies that have been proposed in the literature. Using this background, the cross-case analysis categorises the organisational responses that were observed in practice. The discussion shows that organisations are highly aware of the AI ethics debate and keen to engage with ethical issues proactively. However, they make use of only a relatively small subsection of the mitigation strategies proposed in the literature. These insights are of importance to organisations deploying or using AI, to the academic AI ethics debate, but maybe most valuable to policymakers involved in the current debate about suitable policy developments to address the ethical issues raised by AI.

Keywords Artificial intelligence · Ethics · Organisational response · Case study · AI policy

1 Introduction

The discussion of the ethical aspects of artificial intelligence (AI) remains lively. Following the increasing success of AI, in particular AI technologies based on machine learning and involving big data analytics, ethical concerns have been a high-profile issue. While the concept of AI stems from the 1950s and ethical questions related to AI have been discussed for decades, this discussion has only reached the attention of policymakers, civil society and the media in recent years. The twin reports to the US President on the topic (2016a; b) may be a good marker for the growth in public attention.

The reason for this growing attention is rooted in the capabilities of these technologies which can increasingly fulfil tasks that used to be reserved for humans. Not only can they beat the best humans in games that were previously thought to require human intelligence, but they are approaching or surpassing human-level achievements in a broad range of activities, from facial recognition and diagnosis of cancer cells to the optimisation of organisational or societal processes. The recognition of the capabilities of AI provides the basis for the expectation that further technical development will lead to fundamental changes to the way we live our lives.

This fundamental change thrust upon society by the current and future implementation of AI across all social spheres has led to a range of activities aimed at better understanding and responding to the ethical concerns associated with AI. One can observe a proliferation of ethical principles, statements, guidelines and other documents, from academia, governmental bodies and industry that purport to provide guidance on the ethics of AI (Ryan and Stahl 2020).

✉ Bernd Carsten Stahl
bstahl@dmu.ac.uk

1 Centre for Computing and Social Responsibility, De Montfort University, Leicester, UK

2 School of Sciences, UCLan Cyprus, Larnaca, Cyprus

3 Wageningen Economic Research, Wageningen University and Research, Wageningen, The Netherlands

4 Department of Philosophy, University of Twente, Enschede, The Netherlands

5 Business Systems and Operations, University of Northampton, Northampton, UK

One aspect that is not well understood is how this discourse on the ethics of AI translates into the practice of organisations that use these technologies. The question of how organisations using AI perceive and address the ethical concerns linked to these technologies needs to be understood if various ethical principles and policy guidelines are to be effective (Stahl et al. 2021). The vast majority of ethical concerns arise because of organisational uses, be they in the private sector, such as insurance or automotive, or in the public or third sector, such as local councils or universities. This paper, therefore, presents the findings of a cross-case analysis of ten case studies of organisations using AI and focuses on the findings concerning the organisational responses to the perceived issues of AI.

The paper makes an important contribution to the discussion of the ethics of AI, which is generally weak in terms of empirical support and insights, in particular with regards to the actual practices employed in responding to these issues. Such insight is required to inform the academic debate but also, and maybe even more crucially, to provide input in the ongoing policy discussion and to establish good practice that organisations can draw on.

The paper is organised as follows: The next section describes the concept of AI, ethical issues related to it and the landscape of proposals on how to best address these issues. This is followed by an account of the methodology employed in our empirical study. We then describe our findings of organisational practice and contrast it with the proposals from the literature review. The discussion highlights our main findings while the conclusion points to some potential next steps.

2 Possible responses to the ethics of AI

To understand and contextualise the organisational responses to the ethics of AI, we first need to clarify the terminology, give an overview of what these ethical issues might be and present an overview of the mitigation strategies proposed in academic literature.

2.1 Artificial Intelligence

The current discourse on AI ethics abounds with technical definitions. These often refer to the capabilities of the technology, such as an "AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" (OECD 2019, p. 7). An alternative approach is to look at the underlying technologies (2020, p. 2): "AI is a collection of technologies that combine data, algorithms and computing power" (European Commission (2020, p. 2). An important aspect to note when relying on the technology-centred view of AI is that it is not a single technology, but rather a "set of techniques and sub-disciplines" (Gasser and Almeida 2017, p. 59) (Floridi and Cowls 2019). A third alternative is capability-based views such as having an "ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan and Haenlein 2019, p. 17).

One of the reasons why the definition of AI remains open and contested is that it implicitly or explicitly refers to humans (or other animals) as a point of reference and the fact that AI is meant to replicate some aspect of human or animal ("natural") intelligence. This raises the question of what counts as natural, as opposed to artificial intelligence. It also leads to the philosophical and ethical problem of what, if anything, fundamentally distinguishes machines from human beings (Brooks 2002; Haraway 2010).

This paper is not focused on the definition of AI, but it is plain to see that the definition influences or even determines what can count as ethical issues arising from AI. As our main interest is in how organisations perceive and deal with ethics of AI, it is appropriate to use a broad definition that accommodates the wide variety of views on AI that can be encountered. We thus include the capabitility-based approach alongside (Kaplan and Haenlein 2019, p. 17) technology-based views. This broad definition of AI is consistent with the term smart information systems that is sometimes used to denote the same technical systems (Stahl and Wright 2018).

In the following sections, we provide a brief introduction to the ethical issues of AI. As an initial orientation to the reader, we offer this graphical overview in Fig. 1 which distinguishes ethical benefits and three categories of ethical issues that are explained in more detail below.

2.2 Benefits of AI

Most texts focusing on ethics of AI are primarily concerned with negative consequences. However, it is also important to be aware of the benefits that AI can bring. As a general-purpose technology, it is impossible to predict where exactly AI will be used, but in most of the potential areas of application, there will be beneficial consequences.

The fundamental capacity of AI to process large amounts of data at speed and the resulting ability to optimise choices brings numerous advantages. The ability to analyse large quantities of data facilitates the generation of insights that humans would not be able to achieve (IRGC 2018), thus promising progress in science and knowledge generation in most areas, often with unpredictable consequences. Another area of benefit is the possibility of optimising processes and thereby increasing efficiency. This leads to economic benefits, increasing productivity and creating wealth in society,

Fig. 1 Ethical issues of AI

as well as supporting areas of the economy other than just the AI sector (European Commission 2020). Such efficiency gains can have immediate effects by reducing environmental damage arising from inefficient production.

Beyond this, there are broader vistas of positive AI-enabled futures. In addition to environmental sustainability, AI can contribute to other aspects of sustainability as summarised in the United Nation's (2015) Sustainable Development Goals (SDGs) (Ryan et al. 2019,2020). These SDGs, interpreted as the current consensus of humanity concerning morally desirable outcomes, are therefore, often taken as the yardstick to measure whether AI can achieve morally desirable goals, whether AI can be "good" (Berendt 2019). The link between AI and the SDGs is stressed by the EU's HLEG (2019) and is discussed in detail in (Ryan et al. 2020).

Much of what follows focuses on the ethically problematic consequences of AI. However, to understand the context in which mitigation measures for such ethical issues are developed and implemented, it is important to bear in mind that any approach to governing AI will not only need to address ethical issues, but also maintain the balance between positive and negative consequences of these technologies.

2.3 Ethical issues of AI

Ethical issues of AI are the subject of different interlinking discourses. There is a long-standing academic discussion that explores ethical issues of computers and digital technologies (Floridi 1999; Bynum 2008), going back to seminal figures such as Norbert Wiener (1954, 1964) or Josef Weizenbaum (1977). A second stream of work is taking place in the policy area. This is represented by policy documents that are developed by governmental or parliamentary bodies (House of Commons Science and Technology Committee 2016; Executive Office of the President 2016a; European Parliament 2017; House of Lords 2018). These publications often focus on AI policy more broadly, but tend to cover ethical questions as one category of issues that needs to be addressed. Public bodies can also focus specifically on ethical issues or commission ethics-related reports where these have been flagged up as being of particular urgency (HLEG on AI 2019). The third stream of discussion takes place in the media (Ouchchy et al. 2020), normally informed by both the academic and the policy discourse.

All three approaches to ethics of AI tend to start by establishing ethical values. Despite the large number of publications, there appears to be a relatively stable set of values that are shared across regional and cultural boundaries (Jobin et al. 2019). These include values like justice, freedom, transparency, trust, dignity or solidarity. The list of these values is long and difficult to delineate. This paper focuses on the organisational responses to the ethics of AI, and therefore, needs to be based on a sound understanding of the current discourse.[1] We believe that it is useful to distinguish between three types of ethical issues: those that arise specifically from machine learning, those that relate to the way we use technology to organise life in the digital world, and those referring to broader metaphysical questions. We briefly discuss each of these categories in a separate subsection below.

2.3.1 Issues arising from machine learning

The first group, issues arising from machine learning, are specific to one set of AI techniques, namely those that have led to the recent successes of AI and thus to the prominence of the debate. Machine learning has traditionally been seen as one component of AI, but the progress in compute power and the availability of large amounts of data has enabled machine learning techniques to lead to exciting breakthroughs, for example in areas of facial recognition, natural language processing, and autonomous driving (Horvitz 2017; Ryan 2020a). Machine learning based on neural networks and big data analytics has some characteristics that give rise to ethical concerns. These arise from the need for large training datasets, which are particularly problematic when they contain personal or unrepresentative data (O'Neil 2016; Criado Perez 2019). The other problematic characteristic of these techniques is their opacity, their character as a black box which means that the exact functioning of the AI is difficult to ascertain.

One can observe three main groups of concerns that are directly related to machine learning: control of data, lack of transparency, and reliability. The first group of concerns, related to control of data, covers issues related to the protection and confidentiality of data, in particular personal data, which also covers questions of data security and integrity. While none of these are new or of exclusive relevance to AI, they gain new prominence in the context of AI, which not only relies on access to large datasets which may include personal data, but which may also lead to new vulnerabilities and new ways of linking data and identifying individuals (Stahl and Wright 2018).

The second group of concerns related to machine learning can be characterised as relating to the reliability of the system and its outputs. Underlying worries are the quality and accuracy of data. As current machine learning techniques allow drawing conclusions from large data sets, the quality of these underlying datasets determines the quality of outputs. A prominent example of such systems would be AI systems applied in health, for example for the purpose of diagnosis of radiological data. If such systems are to be used successfully in clinical practice, they need to be highly reliable, a feature that currently few systems exhibit (Topol 2019). Also, this is related to gender-biased machine learning where AI is trained using a majority of information or data that do not equally represent males and females within a given dataset. There is a risk that algorithms trained on male-dominated datasets may result in inaccurate or unreliable outputs from AI systems. The author identifies as part of the chapter entitled "The myth of meritocracy" the multilevel bias that has existed historically in the field of computer science (in terms of students, practitioners, publishing authors, perceptions in the media, etc.). Machine learning based on such historical data may consequently result in biased models (Criado Perez 2019).

The third and final group of concerns related to machine learning has to do with the lack of transparency (Hagendorff 2019). As these systems currently tend to work as black-box systems whose inner workings not even their developers understand, there are worries that this lack of transparency, in itself arguably ethically problematic, can also cause or

[1] Following standard practice for applied ethics papers, we do not explicitly draw on any one ethical tradition (Himma 2004, p. 3). The values we discuss are generally consistent with most deontological, rule utilitarian, and intutionist frameworks. However, we operate from a pluralist Rossean perspective of highlighting *prima facie* duties which may at times conflict (Ross 2002).

at least hide other issues. These could be biases, a much-discussed topic in the current discourse, which could arise due to hidden biases in the datasets, thus linking back to the earlier concern about data quality, or they could arise from the functioning of the algorithms (CDEI 2019). The lack of transparency also leads to worries about accountability and liability for the consequences of use of AI systems.

2.3.2 Ethics of living in a digital world

The second set of concerns discussed in the literature on ethics and AI has to do with the ways in which modern societies use technology to organise themselves. Many of these are not exclusively caused by AI and apply similarly to other technologies. They form part of the discussion, however, because it is expected that AI will influence or exacerbate these issues. These types of issues can be categorised in many ways. For the purposes of this paper, we suggest the following groups of issues: economic issues, questions of justice, issues related to human freedoms, broader societal issues and unknown issues.

Discussions of economic issues include the high-profile issue of changes to employment caused by AI which may lead to loss of employment or changes in the nature of work, with a particular emphasis on more skilled employment (Haenlein and Kaplan 2019). Other economic concerns relate to the ways in which AI can support the concentration of economic power, which is closely related to matters of intellectual property. The growing worldwide dominance of big internet companies is a frequently cited concern (Nemitz 2018).

The second set of concerns, those related to justice, cover the impact on individual groups, and sections of society (for example in cases where AI is used for purposes of law enforcement). There are worries expressed here about access to public services, in particular for vulnerable groups. Fairness is also discussed, linking back to issues of bias and discrimination, as well as to fairness of distribution and economic participation.

The category of human freedoms represents a set of issues related to the ways in which AI can change how humans act or perceive available options. There are worries that the increasing autonomy of machines may limit human autonomy. More specific issues deal with harm to physical integrity (e.g. through injuries suffered because of autonomous vehicles) and impact on health due to health-related AI. Human freedoms may also be affected by lack of access to information or to systems. There are worries about the loss of human contact, for example in care scenarios (Stahl and Coeckelbergh 2016). Very generally, there are further concerns about human rights of both the end-users of AI and individuals in the AI supply chain (WEF 2018).

The category of broader social issues includes those questions that refer to larger societal developments caused by AI that are perceived as problematic. Examples are the impact of AI on the environment, the consequences for democracy, and the use of AI for military purposes.

Finally, there is the category of 'unknown issues' covering concerns that arise from a lack of knowledge. These relate to unintended and unforeseen adverse impacts which by definition are not yet known. A more specific worry is that of the costs to innovation arising from intervention into the AI development processes. Criminal and malicious future uses can be counted in this category, as well as the concern that focusing on particular issues at this point may deflect resources from focusing on other problems that may be more important.

2.3.3 Metaphysical issues

The final set of issues are those that we call metaphysical issues. These have long been discussed in contexts of philosophy of AI as well as broader discourses, including science fiction. The core of these issues is that AI will fundamentally change its own nature or the nature of humans in ways that are ethically problematic. These concerns are typically not linked to current techniques of AI such as machine learning, neural networks etc., but are expected to arise from what is often called 'general AI', as opposed to the narrow or specific AI that we observe in practice (Baum 2017). It is an open question whether and how narrow AI can give rise to or lead to the development of general AI.

Metaphysical issues in this sense often refer to the possibility of machines becoming conscious, at least to some degree (Dehaene et al. 2017; Carter et al. 2018), and thereby acquiring a different moral status. If this were the case it might not only change the moral status of the AI itself as an 'autonomous moral agent' (Stahl 2004; Floridi and Sanders 2004; Wallach et al. 2011), but it could have further consequences, for example if AI could improve itself, thereby leading to an exponential growth in AI capability, sometimes referred to as super-intelligence (Torrance 2012; Bostrom 2016; Kurzweil 2006; Tipler 2012). While it is unclear what would happen at this stage, whether machines would be benevolent, malevolent or neutral towards humans, it is plausible that it would fundamentally change human societies and thus be ethically relevant.

More immediate metaphysical issues might be those where existing and emerging technologies change the way humans can act and interact. The close coupling of humans and AI, for example through wearable devices or implants, but also through the general pervasiveness of machines, such as the ubiquitous mobile phones most of us carry. There are interesting questions about how this close integration

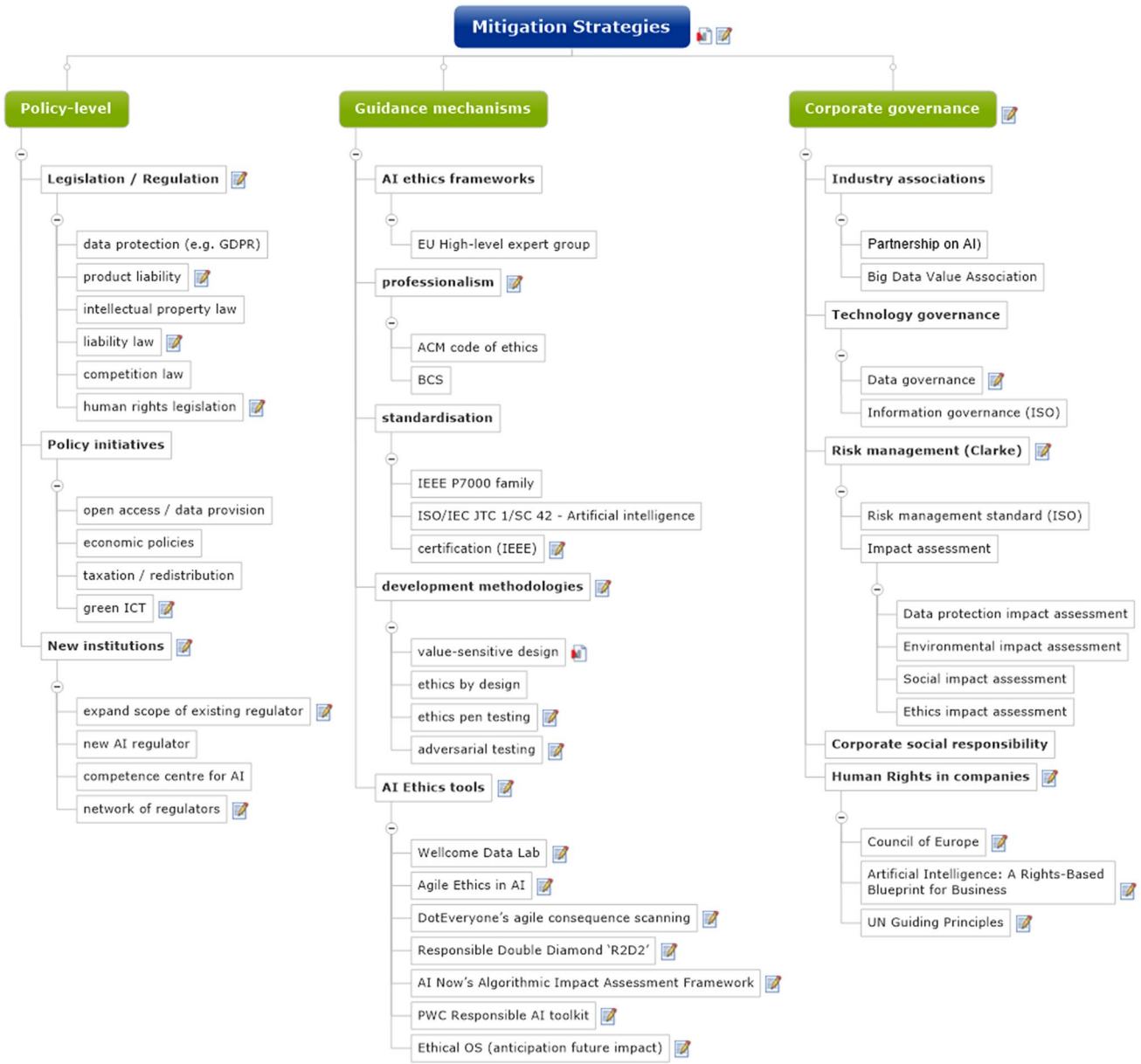

Fig. 2 Key mitigation strategies for ethical issues of AI

of humans and machines changes the nature of humans and how this is to be evaluated from an ethical perspective.

2.4 Mitigation strategies

The previous section provided an overview of the ethical issues that are typically discussed in relation to AI. It is important to be aware of these, to understand how they can be addressed. This paper focuses on organisational ways of addressing the ethics of AI. However, organisations work in a political and societal environment and are made up of individuals, so it is important to understand the breadth of interventions and mitigation strategies that are currently discussed. This section does not provide a comprehensive overview of all possible strategies, of which Ouchchiy et al. (2020) identified 106, but instead offers examples of the most widely discussed topics. There are again many ways in which these could be categorised and organised. In this paper, we start with approaching mitigation strategies at the policy level, then look at guidance mechanisms and other supporting activities before looking at suggestions for corporate governance.

Similar to the previous section covering the ethical issues of AI, we start this section with a graphical overview in Fig. 2 which is then elaborated and explained in the subsequent sections. It shows the distinction of three

levels of mitigation strategies that we applied: policy-level strategies, guidance mechanisms and corporate governance mechanisms.

2.4.1 Policy-level mitigation

There is broad agreement at policy level (regional (e.g. EU) and international (e.g. UN)) as to how the values determining the development, deployment and use of technologies should be safeguarded. On the policy level, one can distinguish between regulation/legislation, the creation of institutions and wider policy activities.

The first significant body of legislation with direct applicability to at least some areas of AI is data protection legislation. In the EU the General Data Protection Regulation (GDPR 2016) is the most visible, but there are other legal instruments protecting personal data that are applicable to AI. In addition, AI, due to the perceived or real autonomy of systems, raises questions about accountability, responsibility and liability (Expert Group on Liability and New Technologies 2019). A third area of relevance is that of intellectual property law, which to a large degree determines who can own data, what data can be used for and what count as legitimate business models for organisations using AI. Human rights legislation has also been identified as crucial to the way ethical issues can be addressed (WEF 2018; Committee on Bioethics (DH-BIO) 2019).

Legislation and regulation are only one tool that policymakers can use to instigate or promote wider policy goals. Broader policy agendas of relevance include open access policies to allow potential users and developers access to data required for the development of AI systems, and policies around green ICT to counter environmental damage arising from the increased power consumption of AI (European Commission 2020). Questions of justice and distribution also play a central role in economic policies, including tax policy and social welfare.

Finally, one way of implementing policies and enforcing legislation is to use regulatory bodies. It is therefore, not surprising that debate has begun to determine whether existing regulators can adequately deal with AI (2017), whether sectoral regulators need to assume some new regulatory tasks to cover AI, or whether new regulators are required. The EC's White Paper (2020) suggests the development of a network of regulators.

2.4.2 Guidance mechanisms

Policy level initiatives are crucial for setting the agenda, providing incentives and focusing stakeholders' minds on dealing with the ethics of AI. However, they tend to be broad, lack detail, and do not normally provide specific guidance at an individual or organisational level. This is the role of guidance mechanisms, methods that provide practical suggestions, steps to follow, and methodological approaches.

At the highest level of abstraction, guidance mechanisms tend to resemble general policies. We consider AI ethics frameworks as an example of such a mechanism (e.g. the EU HLEG (2019) ethics guidelines for trustworthy AI). Jobin et al.'s recent (2019) review of AI ethics guidelines identified 84 examples of such guidelines with widely varying lengths, levels of abstraction and audiences. The proliferation of guidelines has been criticised because the multiplicity of guidelines has the potential to confuse users (Floridi and Cowls 2019) and may be difficult to apply.

One way in which guidelines could find traction is through their adoption and implementation by professional bodies. The advantage of this approach would be that they could build on established ethics processes, such as codes of conduct and disciplinary mechanisms (Gotterbarn et al. 1999; Brinkman et al. 2017). However, there is a longstanding debate as to whether the professions in ICT, and by extension AI, are sufficiently developed to have an effective guidance function (Mittelstadt 2019).

A further mechanism is standardisation and certification. At present, there are several standardisation initiatives in AI, notably the IS SC42 group of standards and the IEEE, which is developing a family of standards touching on ethics and AI (IEEE 2017). Standards, once established, can be used for implementation, for example via certification. The IEEE (2019) has started an ethics certification program for autonomous and intelligent systems, even prior to the publication of their standards. Notably, certification is seen by the European Commission as a mechanism to promote trustworthy AI (European Commission 2020).

A further set of guidance mechanisms is based on existing design methodologies, such as value-sensitive design (Friedman et al. 2008; Simon 2017), privacy by design (Information Commissioner's Office 2008; van Rest et al. 2014; European Commission 2018) or ethics by design (Shilton 2013; Martin and Makoundou 2017; d'Aquin et al. 2018; EDPS 2020) (SHERPA 3.2) (Clarke 2019). Further suggestions have been put forward to test ethical aspects as part of the overall systems design and review process, for example through adversarial testing (WEF 2018) or ethics penetration testing (Berendt 2019).

At the most specific level of guidance mechanisms one can find tools that help identify, reflect on or address ethical issues. There is an array of tools published by research funders, such as the Wellcome Data Lab's method for ethical data science (Mikhailov 2019), civil society organisations, such as doteveryone's (2019) consequence scanning kit, and commercial organisations, such PWC's (2019) practical guide to responsible AI. Initial attempts to categorise the available tools show the breadth of issues they cover and how they can be put into practice (Morley et al. 2019).

2.4.3 Corporate governance of AI ethics

The organisational response to the ethics of AI on which this paper focuses is influenced by policy and legislative initiatives as well as specific guidance mechanisms. Companies are not restricted to being passive recipients of policies or users of guidelines. They can take the initiative and help steer policy development or establish good practice. They can do so as individual organisations or through collective means, such as the Partnership on AI (https://www.partnershiponai.org/) or the Big Data Value Association (http://www.bdva.eu/).

While companies can actively shape the environment in which they operate, they can also individually address ethics of AI. For example, they can implement the principles of corporate governance of IT (ISO 2008). It is true that information governance and data governance (Khatri and Brown 2010; British Academy and Royal Society 2017) are not always geared towards ethical issues (Khatri and Brown 2010; British Academy and Royal Society 2017), but they can be constructed in ways that embrace ethical awareness (Fothergill et al. 2019).

There are various other organisational processes that are relevant. The ethics of AI can be seen as a possible risk factor to organisational success. It has been suggested that risk management (ISO 2010) strategies incorporate ethics of AI as an appropriate way of dealing with such issues (Clarke 2019). As part of risk management, organisations undertake impact assessments. The GDPR calls for data protection impact assessments (Clarke 2009; CNIL 2015) to be integrated into data protection measures. Other types of impact assessment, such as environmental impact assessments (Hennen 2002), social impact assessments (Becker 2001) or ethics impact assessments (CEN-CENELEC 2017) may be sensitive to particular consequences of AI use.

Attention to ethical and social concerns is something many companies commit themselves to in various ways. One well-established approach is to develop a strategy to for corporate social responsibility (CSR) (Carroll 1991; Garriga and Melé 2004; Porter and Kramer 2006). While CSR traditionally does not focus on technology and innovation activities, it has been suggested that including these activities into CSR can contribute to responsible innovation (Martinuzzi et al. 2018; Brand and Blok 2019).

A final approach involves integrating and emphasising human rights in organisations. (United Nations 2011)Due to the potential human rights impacts of AI, it has been suggested that strengthening and integrating human rights in organisations may be an appropriate way to address ethical issues. The World Economic Forum (2019) promotes this approach as does the Council of Europe (2019a, b). The global non-profit organisation BSR has developed implementation guidelines for this purpose (BSR 2018).

Figure 2 does not claim to be comprehensive but demonstrates the breadth of mitigation strategies that have been suggested. It shows that larger policy and legislative initiatives set the tone and shape the environment in which companies operate. It also shows that organisations have a large set of options they can pursue.

This raises the question of what organisations do in practice. So far there has been very little empirical research that tries to answer this question. Where empirical observations inform publications on the ethics of AI, they often focus on particular, often high-profile cases or they are illustrative examples of particular points (O'Neil 2016). While such work is important and has contributed to the quality and visibility of the debate, it leaves a gap in our understanding of how ethics of AI is perceived and addressed by average organisations.

3 Methodology

To answer the question of how organisations address ethical issues of AI, we chose a multiple case study research strategy. Case study research has been recommended as a suitable methodology where new topic areas are investigated (Eisenhardt 1989). Case studies provide answers to "how" or "why" questions in contemporary events over which the investigator has little or no control (Yin 2003a; Cavaye 1996; van der Blonk 2003; Keutel et al. 2013). In this paper, we adopt the interpretive approach to case study research in information systems (Walsham 1995) and its focus on sense-making in organisational contexts.

One drawback of the case study approach is that it can only provide temporally-limited insights. While it is possible to generalise from interpretive case studies, for example by generalising to theory (Walsham 1996), case studies do not allow drawing conclusions about populations on the basis of statistical samples. To address this issue, and to allow the generating of insights into a broader range of organisational applications of AI, we structured the research as multiple case studies. While this does not guarantee statistically reproducible results, it can strengthen the robustness of findings (Darke et al. 1998; Yin 2003b).

To determine the focus of the case studies, we started with a brainstorming exercise of the research team. The purpose was to identify areas of application of AI that are likely to raise ethical issues or that have already been highlighted in the literature as ethically relevant. This exercise was informed by a parallel review of the literature on ethics and AI. We started by identifying relevant social domains where AI is likely to be employed and have ethical relevance. Once a set of social application domains was identified, we engaged in an iterative process to locate possible case study sites that would allow us to undertake the research. The

Table 1 Overview of case study domains

Case study No	Case study social domain	Country	Organisations
CS01	Employee monitoring and administration	Cyprus	A company using the Internet of Things (IoT) for employee monitoring and administration
CS02	Government	The Netherlands	A division within government, a municipality, using AI
CS03	Agriculture	Germany	Large agribusiness using AI
CS04	Sustainable development	The Netherlands; Denmark; Germany; and Finland	1. Large municipality; 2.Public organisation; 3. Telecommunications company; 4. Large municipality
CS05	Science	UK	A large scientific research project
CS06	Insurance	Germany	National health insurance companies
CS07	Energy and utilities	The Netherlands	National energy and utilities company
CS08	Communications, media and entertainment	Finland	Cybersecurity department within a multinational telecommunications company
CS09	Retail and wholesale trade	Finland	A national telecommunications company developing AI for retail customer-relation management
CS10	Manufacturing and natural resources	Austria	A company developing AI for risk prediction in supply-chain management

result of this was a set of ten domains in which we could perform case study research (Macnish et al. 2019b).

Table 1 gives an overview of the case study domains. In accordance with ethics requirements, we cannot reveal the individuals or organisations in question and thus refer to the case studies by number and, where relevant, by social domain.

Following agreement on the case studies to be investigated, we developed a protocol that determined the details of the research, including interview questions and reporting principles. A responsible approach was employed for the case study protocol, in which gender equality was one of the aspects considered for designing the corresponding interview protocol. Questions specific to responsibility approaches (including gender equality) within the organisations were included, and participants interviewed were selected to aim for a gender-balanced input. In fact, the overall set of interviewees included both male and female interviewees from the organisations where that was possible. This was used to obtain ethics approval, which included participant information and consent forms. The case studies were undertaken in a distributed way, with each case relying on one partner of the research team for data collection and analysis. However, they were closely coordinated, starting with the shared protocol, central reporting and review processes, and a shared approach to data analysis based on established principles of qualitative and interpretive research (Miles and Huberman 1994; Aronson 1995; Braun and Clarke 2006). We used the qualitative data analysis software NVivo (server version 10). During an initial workshop the detailed data collection was agreed as well as the data analysis process. We defined a set of top-level codes for the data analysis, but allowed all researchers to add new nodes to the coding tree. A weekly teleconference of all participants ensured a shared view of the codes and constant exchange on the case studies. All cases were reported using a collectively developed case study template.

The empirical data collection was undertaken between June and December 2018. For each case, a minimum of two organisational members were interviewed, with at least one having technical expertise on the AI in use, the other having managerial/organisational expertise. For the 10 case studies, we interviewed a total of 42 individuals. Following the initial report of each case, we went through a period of peer review among the research team, to ensure that the cases were consistent and comparable. Further and more detailed description of the case studies, the protocol, the cross-case analysis and the methodology employed are described in (Macnish et al. 2019b).

4 Findings: organisational responses to the ethics of AI

The findings in this paper focus on the way in which organisations in our case studies dealt with and addressed the ethical issues arising from their use of AI. We have reported elsewhere (Stahl et al. 2021) the details of the ethical issues that the case studies encountered. Suffice it to say that these correspond closely to the categorisation of ethics of AI as shown in Fig. 1, with the notable absence of metaphysical issues. This is not surprising, as the case study organisations

used existing AI and big data techniques, none of which are currently at a stage where they display general AI capabilities, still less machine consciousness or superintelligence. In the cross-case analysis that followed the completion of the individual case studies (Macnish et al. 2019b) we identified five groups of methods used in the case studies to ensure the ethical development and use of AI: organisational awareness; technical approaches; human oversight; ethical training and education; and balancing competing goods. In the following subsections, we highlight key insights. Due to space constraints in this paper, we do not provide detail or many original quotes, which can be found in (Macnish et al. 2019a). We do, however, provide references to the case studies that gave rise to the relevant insights, e.g. CS02 for Case Study 02 (Government), as per Table 1.

In the following sections we describe five groups of mitigation strategies that were identified in the process of analysing the case study data and then clustered during the cross-case analysis. These groups of mitigation strategies are the result of a team effort which was based on a collective review of data, shared insights and discussion. As is normally the case for interpretive qualitative data analysis, we do not claim that this is the only way in which the data could be classified. In this case all of the individuals involved in the data analysis were also involved in the data collection of initial analysis of individual case studies and thus close to the data, thereby ensuring a high level of plausibility of the categorisation.

4.1 Organisational awareness and reflection

An initial insight arising from our work is that the respondents were aware of the ethics of AI debate and its relevance to their work. The organisations involved were already engaged in finding ways to mitigate these issues. Some of the approaches taken included responsible data science, stakeholder engagement, ethics review boards, and following codes of ethics and standards of practice. Respondents showed an awareness that responsible development and use of AI could create positive relations between corporations and people by reducing inequality, improving well-being, and ensuring data protection and privacy. For example, several interviewees stressed that they did not want personal data (CS02, CS04) or that they sought to minimise its collection (CS04, CS09).

The organisations were concerned about the implementation of ethical and human rights-focused approaches in the development, deployment, and use of AI. However, they were often conflicted by the legal, economic, technical or practical ability to follow through with many of their goals. For example, CS10 explicitly attempted to preserve human rights and view their use of AI as a way to protect these rights. However, they stated that predictive risk intelligence companies are often challenged by the most profitable way to use AI. Many of the interviewees stated that the technological robustness of their AI was one of the primary ways to ensure their technologies were used ethically and safely.

4.2 Technical approaches

The organisations implemented a wide range of technical methods to ensure the protection of privacy during the use of AI, such as *k*-anonymity (CS02, CS04); encryption (CS01, CS03, CS05); government-supported secure storage (CS02); and anonymisation or pseudonymisation of data (CS01, CS04). Some companies employed third-party penetration testers to examine their systems for weaknesses (CS03), while others held regular hackathons and sent test phishing emails to staff (CS08). While some companies relied on technical solutions to privacy concerns, those with greater technical expertise, especially in computer security, were more cautious (CS05 and CS08).

4.3 Human oversight

Trust in AI is often affected by the lack of human involvement in decision-making, as highlighted in CS03, CS04 and CS09, and more recently in the AI HLEG's (2019) promotion of 'trustworthy AI'. Despite the promises made on behalf of AI, technical systems still have some inadequacies and so continue to require human oversight and intervention. For instance, in CS03 it was mentioned that '*AI cannot replace agronomists but can support them and there is still a need for a knowledgeable person to provide further support'*. Thus, greater trust in people and their expertise remains necessary, compared with trust in information systems (Ryan 2020b). In the literature, the issue of trust also arises around the uncertainty of new AI capabilities that may adversely affect people. This was echoed in CS09 where a question was raised about whom does one trust, and whether one can trust a machine or an algorithm as opposed to a human being. While a machine may be trusted for its reliability, for instance, this is distinct from it being trusted to make the "right" decision in cases of moral uncertainty (Simpson 2011; Lord 2019). Respondents stated that mistrust could result from AI or humans making unfair decisions or a lack of transparency in how those decisions are made. They proposed adequate human oversight of AI and implementing adequate accountability procedures for when issues do occur.

4.4 Ethics training and education

Often the weight for the transformation of technology falls on the technology experts, designers and developers of software and databases (CS01). They must decide issues

Fig. 3 Mitigation strategies employed in case studies

about data collection, data manipulation, and computational aspects of AI applications (CS10). Software designers must incorporate the necessary reliability aspects and consequently redesign the software to ensure their inclusion (CS01). However, people with such skill sets are rarely trained in ethical analysis. This was especially apparent in CS01, CS08 and CS09, where all of the interviewees were technology experts and had no background in the study of ethics but were acutely aware of privacy concerns in AI use.

4.5 Balancing competing goods

The case study respondents understood the need to balance competing goods and claims. One example is the control of data. Some interviewees aimed to place more control in the hands of those to whom the data pertain. In CS04, an explicit link was made between citizens having control over their data as a means to ensure privacy. The issue here is transparency: whether citizens know what happens to their data and why (CS02 and CS04). However, private companies may not want to be entirely transparent about their algorithms for reasons of intellectual property and fears were expressed that some might 'game the system' (CS09) However, some interviewees stated that while the details of specific processes might not be transparent, codes of conduct and general principles should be made publicly available (CS10), which suggests that transparency can play a role in finding acceptable solutions for trade-offs.

The possibility of AI use resulting in stigmatisation and discrimination was mentioned. This calls for mechanisms to test the fairness and accuracy of algorithmic scoring systems and to allow citizens to challenge algorithms that cause them harm. One of the interviewees (CS04) stated that public–private relationships on AI projects have the potential to enhance and improve the lives of citizens, but that they also hold the possibility of increasing costs, harming sustainability efforts, and creating power asymmetries. The interviewee stated that there is a need for careful, explicit, and collaborative efforts between public–private organisations to ensure mutually beneficial partnerships. If this is not possible, public bodies need to develop in-house expertise (CS02) to ensure they reap ethically-sensitive economic, employment, and sustainability benefits from AI. Finally, it is important to reinforce the importance of being aware and be able to reflect on existing hurdles towards such benefits, for instance, the historical gender-bias in society and the workplace especially in the tech world (Criado Perez 2019).

Figure 3 shows a summary of the mitigation strategies employed across all ten case studies.

5 Discussion

Comparing the discussion of the mitigation strategies based on the literature with the findings of the activities undertaken by the case study companies, the empirical findings reflect only a fraction of the possible mitigation

strategies. While generally companies do not work on a policy level, there was little reference to the use of existing corporate governance mechanisms, such as data governance or risk management, to address ethical issues of AI. This may be because the respondents did not make an explicit connection between corporate governance, which at least some of the organisations have in place, or because technical approaches (e.g. anonymisation, penetration testing) form part of their corporate data governance regime. Either way, the underlying governance structure was not mentioned by the respondents.

A second observation is that the interviewees emphasised organisational awareness and reflection. This includes engagement with external stakeholders and internal processes such as ethics review boards. Each can help the internal reflection of the organisation and prepare it to make ethical decisions. This is an important part of engaging with ethical issues of AI, but it did not figure strongly in the literature review of mitigation measures. The case studies demonstrated that the respondents were aware of the broader societal discourse regarding ethical AI and were willing to engage with it (Macnish et al. 2019a). The case studies may suffer from a self-selection bias in that we may only have been able to talk to organisations and individuals who had an interest in the topic and were therefore ahead of the curve. However, while participants showed a willingness to engage with, understand and address ethics of AI, there were some notable absences. For private companies, there was little reference to corporate social responsibility or other organisational commitments to the public good. Also, while respondents were acutely aware of privacy threats arising from AI and used a number of technical measures to strengthen data protection in response, there was almost no reference to privacy as a human right or to other human rights. The approach to AI ethics based on strengthening human rights in businesses as proposed by the Council of Europe and others does not seem to have arrived in organisations.

The case study respondents were aware of the problems arising from living in a digital world, our second category of ethical issues. They tried to address some, for example by finding sustainable ways of dealing with data ownership. Organisational awareness-raising and reflection may well cover some of these issues, but it is less clear how they were to be addressed on an organisational level. Many of these issues are cumulative and only become visible on a societal level. They are thus frequently seen to be beyond the reach and remit of most individual organisations.

The final category of ethical issues, the metaphysical issues, was occasionally mentioned in passing, but did not influence the mitigation strategies that organisations put in place. This can be explained by the fact that our case study organisations all worked on AI or big data technologies that can be described as narrow AI and the metaphysical issues are expected to arise from general AI, which currently does not exist.

Overall, the organisational mitigation strategies that we observed covered an important sub-section of the possible strategies that could be expected from the literature but by no means all. A similar picture emerges when looking at the ethical issues that these strategies are meant to address. Attention was given to issues that arise from machine learning, in particular those that are regulated or clearly recognisable. Control of data and data protection, regulated by GDPR, hence played the most prominent role and all of the technical approaches directly relate to them (see also Macnish et al. 2019b).

6 Conclusion

We believe that these insights are important and contribute to the literature on the ethics of AI. Our research plugs an obvious gap that stems from a lack of broadly-based empirical research across individual application domains of AI. The findings of the study provide some important insights: They confirm that organisations making use of AI were not only aware of the ethical issues that these technologies can cause, but that they were willing to engage with and accept responsibility for doing something about them (Macnish et al. 2019a). At the same time, the organisations made use of only a limited subset of mitigation measures and focused on only a limited set of issues. Many of the ethical issues are seen to be either beyond the organisations' expertise or lie outside their remit. This confirms that organisations can—and already do—play an important role addressing ethics of AI, but that they are not (nor do they see themselves as) the only stakeholder. A broader framework is hence required that covers other stakeholders to ensure a more comprehensive coverage of AI ethics. The question of what exactly lies within the remit of organisations and which issues and measures are the responsibilities of policymakers, professional bodies, and regulators needs also to be addressed.

This study, while empirically rich and rigorous, does not hold all the answers. Using the described methodology, one could add further cases to cover more application domains, including multiple cases from one domain, for contrasting purposes. It should also be extended beyond Europe to cover perceptions and activities in other parts of the world. It could be complemented by other types of data and research approaches, including larger scale quantitative social studies and societal impact studies, which could make use of publicly available data sets. Similar studies could be undertaken looking at other types of stakeholders or in more detail at the dynamics within the organisation. A more detailed analysis of different types of organisation (companies, charities, public sector bodies etc.) would also be helpful.

Despite these various ways of improving the study, the insights presented here contribute to a better understanding of how AI is perceived and how it influences the way modern societies are run. Ethical and human rights issues are important factors in assessing AI's impact. AI raises ethical concerns, some of which can be straightforwardly addressed, but in many cases, these involve fundamental trade-offs. The question about the nature of these trade-offs, the way in which they are perceived and dealt with is at the core of the ethics of AI debate, but it closely involves political, legal and professional discourses. All of these need to be informed by sound concepts, and empirical insights. We therefore, hope that this paper will contribute to a better understanding of the role of AI in organisations and thus to an overall societally acceptable use of these technologies.

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786641 (SHERPA; www.project-sherpa.eu). The authors would like to thank the respondents involved in the case studies. The authors acknowledge the contributions of all project consortium members to various aspects of the planning and implementing of the case study research.

References

Aronson J (1995) A pragmatic view of thematic analysis. Qualit Rep 2:1–3

Baum S (2017) A survey of artificial general intelligence projects for ethics, risk, and policy. Social Science Research Network, Rochester

Becker HA (2001) Social impact assessment. Eur J Oper Res 128:311–321. https://doi.org/10.1016/S0377-2217(00)00074-6

Berendt B (2019) AI for the common good?! Pitfalls, challenges, and ethics pen-testing. Paladyn J Behav Robot 10:44–65. https://doi.org/10.1515/pjbr-2019-0004

Bostrom N (2016) Superintelligence: paths, dangers, strategies, reprint edition. OUP Oxford, Oxford

Brand T, Blok V (2019) Responsible innovation in business: a critical reflection on deliberative engagement as a central governance mechanism. J Respons Innov 6:4–24. https://doi.org/10.1080/23299460.2019.1575681

Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psychol 3:77–101

Brinkman B, Flick C, Gotterbarn D et al (2017) Listening to professional voices: draft 2 of the ACM code of ethics and professional conduct. Commun ACM 60:105–111. https://doi.org/10.1145/3072528

British Academy, Royal Society (2017) Data management and use: Governance in the 21st century A joint report by the British Academy and the Royal Society. London

Brooks RA (2002) Flesh and machines: how robots will change us. Pantheon Books, New York

BSR (2018) Artificial intelligence: a rights-based blueprint for business paper 3: implementing human rights due diligence. BSR

Bynum T (2008) Computer and information ethics. Stanford Encyclopedia of Philosophy

Carroll AB (1991) The pyramid of corporate social responsibility: toward the moral management of organizational stakeholders. Bus Horiz 34:39–48

Carter O, Hohwy J, van Boxtel J et al (2018) Conscious machines: defining questions. Science 359:400–400. https://doi.org/10.1126/science.aar4163

Cavaye ALM (1996) Case study research: a multi-faceted research approach for IS. Inf Syst J 6:227–242. https://doi.org/10.1111/j.1365-2575.1996.tb00015.x

CDEI (2019) Interim report: review into bias in algorithmic decision-making. Centre for Data Ethics and Innovation

CEN-CENELEC (2017) Ethics assessment for research and innovation—Part 2: ethical impact assessment framework. CEN-CENELEC, Brussels

European Parliament (2017) Civil law rules on robotics—European parliament resolution of 16 February 2017 with recommendations to the commission on civil law rules on robotics (2015/2103(INL))

Clarke R (2009) Privacy impact assessment: Its origins and development. Comput Law Secur Rev 25:123–135. https://doi.org/10.1016/j.clsr.2009.02.002

Clarke R (2019) Principles and business processes for responsible AI. Comput Law Secur Rev 35:410–422

CNIL (2015) Privacy impact assessment (PIA) good practice. CNIL

European Commission (2018) Communication from the commission to the European Parliament, the European council, the Council, the European Economic and Social Committee and the Committee of the Regions Artificial Intelligence for Europe. European Commission

European Commission (2020) White Paper on Artificial Intelligence: a European approach to excellence and trust. Brussels

Council of Europe (2019) Unboxing artificial intelligence: 10 steps to protect human rights

Committee on Bioethics (DH-BIO) (2019) Strategic action plan on human rights and technologies in biomedicine (2020–2025). Council of Europe

Criado Perez C (2019) Invisible women: exposing data bias in a world designed for men, 01 Edition. Chatto & Windus

d'Aquin M, Troullinou P, O'Connor NE, et al (2018) Towards an "Ethics in Design" methodology for AI research projects

Darke P, Shanks G, Broadbent M (1998) Successfully completing case study research: combining rigour, relevance and pragmatism. Inf Syst J 8:273–289. https://doi.org/10.1046/j.1365-2575.1998.00040.x

Dehaene S, Lau H, Kouider S (2017) What is consciousness, and could machines have it? Science 358:486–492

Doteveryone (2019) Consequence scanning—an agile practice for responsible innovators | doteveryone. https://www.doteveryone.org.uk/project/consequence-scanning/. Accessed 10 Apr 2020

EDPS (2020) A preliminary opinion on data protection and scientific research

Eisenhardt KM (1989) Building theories from case study research. Acad Manag Rev 14:532–550. https://doi.org/10.2307/258557

Executive Office of the President (2016a) Artificial intelligence, automation, and the economy. Executive Office of the President National Science and Technology Council Committee on Technology
Executive Office of the President (2016b) Preparing for the future of artificial intelligence. Executive Office of the President National Science and Technology Council Committee on Technology
Expert Group on Liability and New Technologies (2019) Liability for artificial intelligence and other emerging digital technologies. European Commission, Luxembourg
Floridi L (1999) Information ethics: on the philosophical foundation of computer ethics. Ethics Inf Technol 1:33–52
Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Sci Rev. https://doi.org/10.1162/99608f92.8cd550d1
Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14:349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d
Fothergill BT, Knight W, Stahl BC, Ulnicane I (2019) Responsible data governance of neuroscience big data. Front Neuroinform. https://doi.org/10.3389/fninf.2019.00028
Friedman B, Kahn P, Borning A (2008) Value sensitive design and information systems. In: Himma K, Tavani H (eds) The handbook of information and computer ethics. Wiley Blackwell, New York, pp 69–102
Garriga E, Melé D (2004) Corporate social responsibility theories: mapping the territory. J Bus Ethics 53:51–71. https://doi.org/10.1023/B:BUSI.0000039399.90587.34
Gasser U, Almeida VAF (2017) A layered Model for AI governance. IEEE Internet Comput 21:58–62. https://doi.org/10.1109/MIC.2017.4180835
GDPR (2016) REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union L119/1
Gotterbarn D, Miller K, Rogerson S (1999) Software engineering code of ethics is approved. Commun ACM 42:102–107
Haenlein M, Kaplan A (2019) A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. Calif Manag Rev 61:5–14
Hagendorff T (2019) The ethics of AI ethics—an evaluation of guidelines. arXiv: 190303425 [cs, stat]
Haraway D (2010) A cyborg manifesto. In: Szeman I, Kaposy T (eds) Cultural theory: an anthology. Wiley Blackwell, Chichester, pp 454–475
Hennen L (2002) Impacts of participatory technology assessment on its societal environment. In: Joss S, Belluci S (eds) Participatory technology assessment: European perspectives. University of Westminster, Centre for the Study of Democracy, London, pp 257–275
Himma KE (2004) The ethics of tracing hacker attacks through the machines of innocent persons. Int J Inf Ethics 2:1–13
HLEG on AI HEG on AI (2019) Ethics guidelines for trustworthy AI. European Commission—Directorate-General for Communication, Brussels
Horvitz E (2017) AI, people, and society. Science 357:7–7. https://doi.org/10.1126/science.aao2466
House of Lords H of L (2018) AI in the UK: ready, willing and able? Select Committee on Artificial Intelligence, London
ICO (2017) Big data, artificial intelligence, machine learning and data protection. Information Commissioner's Office
IEEE (2017) The IEEE global initiative on ethics of autonomous and intelligent systems. https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html. Accessed 10 Feb 2018
IEEE (2019) IEEE SA—the ethics certification program for autonomous and intelligent systems (ECPAIS). https://standards.ieee.org/industry-connections/ecpais.html. Accessed 10 Apr 2020
Information Commissioner's Office (2008) Privacy by design
IRGC (2018) The governance of decision-making algorithms
ISO (2008) BS ISO/IEC 38500:2008—Corporate governance of information technology
ISO (2010) ISO 31000:2009(E)—Risk management. Principles and guidelines
Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2
Kaplan A, Haenlein M (2019) Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz 62:15–25
Keutel M, Michalik B, Richter J (2013) Towards mindful case study research in IS: a critical analysis of the past ten years. Eur J Inf Syst. https://doi.org/10.1057/ejis.2013.26
Khatri V, Brown CV (2010) Designing data governance. Commun ACM 53:148–152
Kurzweil R (2006) The singularity is near. Gerald Duckworth & Co Ltd, London
Lord C (2019) Objections to Simpson's argument in 'Robots, Trust and War.' Ethics Inf Technol 21:241–251. https://doi.org/10.1007/s10676-019-09505-2
Macnish K, Ryan M, Gregory A, et al (2019a) SHERPA Deliverable 1.1 Case studies. SHERPA project
Macnish K, Ryan M, Stahl B (2019b) Understanding ethics and human rights in smart information systems. ORBIT J. https://doi.org/https://doi.org/10.29297/orbit.v2i1.102
Martin CD, Makoundou TT (2017) Taking the high road ethics by design in AI. ACM Inroads 8:35–37
Martinuzzi A, Blok V, Brem A et al (2018) Responsible research and innovation in industry—challenges. Insights Perspect Sustain 10:702. https://doi.org/10.3390/su10030702
Mikhailov D (2019) A new method for ethical data science. https://wellcome.ac.uk/news/new-method-ethical-data-science. Accessed 10 Apr 2020
Miles MB, Huberman AM (1994) Qualitative data analysis: an expanded sourcebook. SAGE, Thousand oaks
Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell. https://doi.org/10.1038/s42256-019-0114-4
Morley J, Floridi L, Kinsey L, Elhalal A (2019) From what to how—an overview of ai ethics tools, methods and research to translate principles into practices. arXive
Nemitz P (2018) Constitutional democracy and technology in the age of artificial intelligence. Phil Trans R Soc A 376:20180089. https://doi.org/10.1098/rsta.2018.0089
O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Penguin UK
OECD (2019) Recommendation of the council on artificial intelligence. OECD
Ouchchy L, Coin A, Dubljević V (2020) AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. AI Soc. https://doi.org/10.1007/s00146-020-00965-5
Porter ME, Kramer MR (2006) The link between competitive advantage and corporate social responsibility. Harvard Bus Rev 84:78–92
PWC (2019) A practical guide to responsible artificial intelligence (AI)
House of Commons Science and Technology Committee (2016) Robotics and artificial intelligence
Ross D (2002) The right and the good. Clarendon Press, Oxford
Ryan M (2020) In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 26:2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Ryan M (2020) The future of transportation: ethical, legal, social and economic impacts of self-driving vehicles in the year 2025. Sci Eng Ethics 26:1185–1208

Ryan M, Stahl BC (2020) Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc. https://doi.org/10.1108/JICES-12-2019-0138

Ryan M, Antoniou J, Brooks L et al (2020) The ethical balance of using smart information systems for promoting the United Nations' sustainable development goals. Sustainability 12:4826. https://doi.org/10.3390/su12124826

Ryan M, Antoniou J, Brooks L, et al (2019) Technofixing the future: ethical side effects of using AI and big data to meet the SDGs. In: Proceeding of IEEE Smart World Congress 2019. IEEE, De Montford University, Leicester, UK

Shilton K (2013) Value levers: building ethics into design. Sci Technol Human Values 38:374–397. https://doi.org/10.1177/0162243912436985

Simon J (2017) Value-sensitive design and responsible research and innovation. In: Hansson SO (ed) The ethics of technology: methods and approaches, 1st edn. Rowman & Littlefield International, London, pp 219–236

Simpson TW (2011) Robots, trust and war. Philos Technol 24:325–337. https://doi.org/10.1007/s13347-011-0030-y

Stahl BC (2004) Information, ethics, and computers: the problem of autonomous moral agents. Mind Mach 14:67–83. https://doi.org/10.1023/B:MIND.0000005136.61217.93

Stahl BC, Coeckelbergh M (2016) Ethics of healthcare robotics: towards responsible research and innovation. Robot Auton Syst. https://doi.org/10.1016/j.robot.2016.08.018

Stahl BC, Wright D (2018) Ethics and privacy in AI and big data: implementing responsible research and innovation. IEEE Secur Priv 16:26–33. https://doi.org/10.1109/MSP.2018.2701164

Stahl BC, Andreou A, Brey P et al (2021) Artificial intelligence for human flourishing—beyond principles for machine learning. J Bus Res 124:374–388. https://doi.org/10.1016/j.jbusres.2020.11.030

Tipler FJ (2012) Inevitable existence and inevitable goodness of the singularity. J Conscious Stud 19:183–193

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56. https://doi.org/10.1038/s41591-018-0300-7

Torrance S (2012) Super-intelligence and (super-)consciousness. Int J Mach Conscious 4:483–501. https://doi.org/10.1142/S1793843012400288

United Nations (2011) Guiding principles on business and human rights—implementing the United Nations "protect, respect and remedy" framework. United Nations Human Rights, New York

United Nations (2015) Sustainable development goals—United Nations. In: United Nations Sustainable Development. https://www.un.org/sustainabledevelopment/sustainable-development-goals/. Accessed 9 Jun 2018

van der Blonk H (2003) Writing case studies in information systems research. J Inf Technol 18:45–52. https://doi.org/10.1080/0268396031000077440

van Rest J, Boonstra D, Evert M, et al (2014) Designing privacy-by-design. Brussels

Wallach WA, Allen CB, Franklin SC (2011) Consciousness and ethics: artificially conscious moral agents. Int J Mach Conscious 3:177–192

Walsham G (1995) Interpretive case studies in IS research: nature and method. Eur J Inf Syst 4:74–81. https://doi.org/10.1057/ejis.1995.9

Walsham G (1996) Ethical theory, codes of ethics and IS practice. Inf Syst J 6:69–81. https://doi.org/10.1111/j.1365-2575.1996.tb00005.x

WEF (2018) White paper: how to prevent discriminatory outcomes in machine learning

Weizenbaum J (1977) Computer power and human reason: from judgement to calculation. W.H.Freeman & Co Ltd, London

Wiener N (1954) The human use of human beings. Doubleday, New York

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

World Economic Forum (2019) Responsible use of technology. WEB, Geneva

Yin RK (2003) Applications of case study research, 2nd edn. Sage Publications Inc, Thousand Oaks

Yin RK (2003) Case study research: design and methods, 3rd edn. Sage Publications Inc, Thousand Oaks