Journal of Business Research 124 (2021) 374–388

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Artificial intelligence for human flourishing – Beyond principles for machine learning

B.C. Stahl [a,*], A. Andreou [e], P. Brey [b], T. Hatzakis [c], A. Kirichenko [d], K. Macnish [b], S. Laulhé Shaelou [f], A. Patel [d], M. Ryan [g], D. Wright [c]

[a] De Montfort University, UK
[b] University of Twente, the Netherlands
[c] Trilateral Research, UK
[d] F-Secure, Finland
[e] Aequitas, Cyprus
[f] University of Central Lancashire, Cyprus
[g] Wageningen University & Research, the Netherlands

ARTICLE INFO

Keywords:
Ethics
Artificial intelligence
Big data
Human rights
Governance

ABSTRACT

The technical and economic benefits of artificial intelligence (AI) are counterbalanced by legal, social and ethical issues. It is challenging to conceptually capture and empirically measure both benefits and downsides. We therefore provide an account of the findings and implications of a multi-dimensional study of AI, comprising 10 case studies, five scenarios, an ethical impact analysis of AI, a human rights analysis of AI and a technical analysis of known and potential threats and vulnerabilities. Based on our findings, we separate AI ethics discourse into three streams: (1) specific issues related to the application of machine learning, (2) social and political questions arising in a digitally enabled society and (3) metaphysical questions about the nature of reality and humanity. Human rights principles and legislation have a key role to play in addressing the ethics of AI. This work helps to steer AI to contribute to human flourishing.

1. Introduction

The development of artificial intelligence (AI) is often described in terms of human progress. The recent progress of machine learning, supported by growing amounts of available data combined with rapidly expanding computing capabilities and publicly available tools and libraries, have led to expectations of increased efficiency but also to new and better services for consumers and citizens. This broadly positive discourse is, however, counterbalanced by a discussion of the downsides and risks of AI.

The ethics of AI is a topic of conversation in the disciplines concerned with these technologies including the social sciences, humanities, media and policy. Worries range from discrimination due to biased datasets to the domination of humanity by sentient machines. The social impact of AI-based technologies provides the backdrop and justification for the flurry of activities in public discourse and policy developments about whether and how AI should be regulated or whether other ways should be found to address the downsides of AI.

A growing volume of literature suggests that governance mechanisms need to be devised for these technologies because existing governance structures are not able to address the issues they raise. As a consequence, one can find numerous suggestions on various ways to develop governance structures that range from the informal, such as voluntary industry codes of conduct, to national and international legislation and the creation of regulators.

One weakness of the current discourse is a disconnect between rigorous academic research on the content and implications of these technologies and the development of governance proposals.

In order to move beyond the current discourse, gain a deeper understanding of the nature of ethics in AI, and allow for a critical reflection of the current discourse, we conducted multi-method and interdisciplinary research aimed at contributing to empirical and

* Corresponding author at: De Montfort University, The Gateway, Leicester LE2 9BH, UK.
E-mail address: bstahl@dmu.ac.uk (B.C. Stahl).

https://doi.org/10.1016/j.jbusres.2020.11.030
Received 18 January 2020; Received in revised form 16 November 2020; Accepted 18 November 2020
Available online 16 December 2020

conceptual clarity of the nature of these technologies, the challenges they raise and the potential of new governance structures to address these issues. The aim of this paper is to contribute to the discussion about how to identify, interpret and address ethical issues arising from AI applications[1]. The paper critically reflects on the term AI and explores which aspects of AI raise which types of issues and how these are reflected and addressed in organisational and societal practice. Bringing together conceptual insights and empirical findings, the paper is in a position to propose new ways to think about AI and structure the AI ethics narrative. In order to achieve this aim, the paper first seeks to answer the question of what precisely are the key ethical issues and how best to classify or categorise them. It then explores how existing governance mechanisms may be applied to these issues. This leads to the final question of theoretical and practical next steps.

Our analysis shows that the ethical issues that arise in empirical observations are similar to those that the academic literature discusses, which provides reason for the belief that the discourse on ethics in AI is reasonably expansive. At the same time, however, it becomes clear that the meaning of these issues is largely context-dependent. We use our understanding of the ethical issues to categorise them into three broad categories: (1) issues directly related to machine learning, (2) broader social and political issues arising in modern digitally enabled societies and finally (3) metaphysical questions. These categories allow us to map currently existing and discussed mitigation and governance structures to these issues. This is an important starting point for the practical question of what can and should be done to address these issues. This question is beyond the scope of this paper.

The findings presented in this paper are important in several respects. The paper makes an academic contribution to the quickly spreading discussion of ethics and AI and research around ethics, values, governance and tools of AI. The categorisation of issues suggested here and the mapping of the categories to different governance mechanisms can help streamline the debate. Due to the high practical importance of the underlying technologies, the paper also has practical importance for stakeholders faced with the practical challenge of proactively engaging with the ethics of AI. The paper can help organisations developing, deploying or using AI to identify issues they are likely to face and engage with governance mechanisms that can address these issues.

In order to develop the argument, the paper proceeds as follows. In the next section, we discuss the governance of AI, looking first at definitions, followed by a discussion of ethical issues and currently proposed governance structures. We then describe our multi-dimensional empirical study of AI. The findings and discussion give rise to our categorisation of issues, which we then map to governance structures and stakeholders. Our discussion and conclusions demonstrate the novelty and relevance of our findings while we propose next steps.

2. Governance of AI

This section provides the conceptual basis of tthis article and gives an overview of current discussions regarding AI, its ethical implications and possible governance structures. The term 'governance' as developed in political sciences traditionally refers to alternatives to formal government on a societal or state level. In business research, it frequently refers to structures and processes within organisations, whereas on a higher level the term 'regulation' is used (Braithwaite & Drahos, 2000). However, 'governance' is increasingly used to describe a much broader array of "[…] processes of governing, whether undertaken by a government, market, or network, whether over a family, tribe, formal or informal organization, or territory, and whether through laws, norms, power, or language" (Bevir, 2012, p. 1). The term also refers to specific localised ways of organising (or governing) particular issues, as in data governance (Khatri & Brown, 2010) or information governance (ISO, 2008), rendering it suitable to describe ways of dealing with AI that cover many societal actors and activities.

2.1. AI and big data

The concept of AI, while much discussed, is not well defined. A typical definition of AI is the one provided by the European Commission (2018, p. 1): "Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals." This is consistent with Haenlein and Kaplan (2019) emphasis on the interpretation of external data, learning from such data and using data for the achievement of specific goals. While such definitions are sufficient to give an idea of the scope of AI, they are arguably not specific enough to allow the identification of specific ethical issues or the application of governance structures. Definitions such as the EC's can also be problematic when they seem to imply contentious positions, such as that AI can behave, analyse and act, which can be read as imputing characteristics, notably that of independent agency, that current machine learning technologies do not display. This points to metaphysical assumptions about AI to which we return to below.

One reason for these shortcomings is that the definitions hide the immense breadth and depth of the underlying AI research (Elsevier, 2018). The current prominence of AI is based on long-established principles of machine learning, often implemented through (deep) neural networks. These have recently gained prominence due to the increased availability of computing power and large data sets for training purposes. Ethical discussions of AI therefore need to be sensitive to both the consequences of the application of AI algorithms and techniques as well as the ethical aspects of (big) data analytics (B. D. Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016; Nerurkar, Wadephul, & Wiegerling, 2016; Varley-Winter & Shah, 2016). Public discourse on AI especially focuses on machine learning and the empirical work we have undertaken covers technologies in machine learning and big data analytics. However, as we argue below, the ethics of AI debate is broader than this and refers to other concepts of AI, notably that AI technologies have broader human-like cognitive abilities. The concept of general AI goes back to the beginning of AI research and is sometimes referred to as Good Old Fashioned AI (GOFAI) (Moor & Bynum, 2002). General AI technologies do not currently exist, but they figure strongly in the public discourse. In order to be able to make sense of the broader debate, it is important to be aware of the entire breadth of meaning of the term.

This paper therefore does not attempt to offer a comprehensive definition of AI or of any of its constituent technologies such as machine learning. Instead, it aims to bring greater clarity to the question what people refer to when they talk about AI and, more importantly, when they talk about the ethics of AI or about the ethical issues of AI. These conceptual questions are crucial to dealing with the ethics of AI and questions of governance. They pose the problem of delineating which ethical issues are related to or caused by AI and it complicates questions of governance, where the application area of governance mechanisms is often not clear, as we will show in more detail below.

2.2. Ethics and AI

The concept of ethics is even more contested and open than that of AI. In everyday English, it denotes questions of right or wrong, of good or bad. Following Stahl (2012), we argue that this everyday understanding of ethics constitutes the basis of explicit reasoning and academic reflection, which are the subject matter of philosophical ethics. Answering the question of why a particular action can be seen as good or bad or which processes would allow answering such a question is the

[1] We use the term "issue" in an open sense, accepting as issues whatever our respondents or the literature describes using the term. As a consequence, the issues we cover vary greatly in terms of scope and impact. Some are very precise and focused whereas other are large and fuzzy and cover entire areas where issues arise.

role of philosophical ethical theories. These include classical theories such as virtue ethics (Aristotle, 2007) which determines the ethics quality often action based on the character of the individual undertaking it. Other frequently used ethical theories include deontology, which focuses on the agent's duty (Kant, 1788, 1797), or teleology which looks at the consequences and outcomes of an action to determine its ethical status (Mill, 1861). In addition to these well-established traditional ethical theories, there are more recent ones like the ethics of care (Adam, 2001; Gilligan, 1990) and specific ethical theories aimed at technological applications, such as computer ethics (Terrell Ward Bynum & Rogerson, 2003; D. G. Johnson, 2001), information ethics (Capurro, 2006; L. Floridi, 1999; Luciano Floridi, 2010) or disclosive ethics (Brey, 2000).

Despite a rich history of discussing the relative merits of various ethical positions, the current discourse around ethical issues of AI makes little reference to philosophical ethical theories. Instead, the generally accepted approach to AI ethics seems to define mid-level ethical principles, an approach pioneered by biomedical ethics (Beauchamp & Childress, 2009; The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979), 1979). This approach, sometimes referred to as principlism, is not without criticism (Clouser & Gert, 1990). It has the practical advantage of sidestepping long-standing ethical debates. But it is open to the charge that it fails to solve practical ethical issues, due to the apparent consensus on ethical principles that then fail to guide practical action (B. Mittelstadt, 2019). Nevertheless, the creation and compilation of ethical principles form key aspects of the ethics of AI debate (Anabo, Elexpuru-Albizuri, & Villardón-Gallego, 2019; Asilomar Conference, 2017; Boden et al., 2017). In addition, most of the high-level interventions into the ethics of AI discussion are principle-based, such as the guidelines produced by the European High Level Expert Group on AI (HLEG on AI, 2019). For our purposes, it is sufficient to understand the prevalence of ethical principles in the AI discourse. We agree with Mittelstadt (2019), however, in seeing the focus on principles as limiting and will return to the question of an appropriate ethical theoretical basis for the ethics of AI below.

Discussions of the ethics of AI tend to cover particular ethical issues. These are typically particular features of the technology or consequences of its use that the authors see as problematic. Many of these have long-standing histories in ethics of technology or ethics of computing, such as security, privacy or access. Some of them seem to be particularly linked to the algorithms that drive AI, such as problems of algorithmic biases (CDEI, 2019; K. Johnson, Pasquale, & Chapman, 2019; B. D. Mittelstadt et al., 2016) and many of them are linked to the compilation and manipulation of large data sets that are required for many of the current AI techniques (Metcalf et al., 2016; Nerurkar et al., 2016; Taylor, 2016). Some ethical issues are specific to particular application areas, such as finance or autonomous vehicles, whereas others are seen as broadly relevant to all AI areas.

2.3. Purpose of AI and governance proposals

The number and reach of ethical issues linked to AI is enormous, in particular, when considering the breadth of possible application areas. Addressing them is therefore a challenge that has attracted much attention. One key question that needs to be answered before any mitigation measures can be developed is the role that AI has and is meant to have in society. AI, along with most other information and communications technologies (ICTs), has a particularly high level of interpretive flexibility (Doherty, Coombs, & Loan-Clarke, 2006), which means that it is difficult to predict how it will be used. This has been a key driver for thinking about ethical aspects of ICTs for decades, sometimes discussed under the heading of "logical malleability" (Moor, 1985). What this means is that even in cases where a technology is designed for a particular purpose, it is difficult to foresee whether and to what degree it will be used for this purpose.

We distinguish between different purposes of making use of AI. The first and most prominent purpose is to improve processes and efficiency. For organisations using AI, this translates into lower costs, higher productivity and, eventually, higher profits. The second purpose is the use for social control. AI techniques are the enablers for voice and face recognition and can therefore be used for surveillance and tracking individuals. This is the basis for controlling individuals to ensure they follow specific requirements. This is the underlying idea of the Chinese Social Credit System (Liu, 2019). The third purpose of using AI is to promote human flourishing. Flourishing is an ethical principle typically associated with virtue ethics, which has a well-established history of application to digital technologies (T. W. Bynum, 2006) and which has been used to frame the AI debate more recently (ALLEA & Royal Society, 2019). It is not always trivial to determine what constitutes flourishing or how technology can contribute to it. However, any attempts to use 'AI for Good', as the title of the series of summits organised by the International Telecommunication Union suggests (https://aiforgood.itu.int/), can count in this category.

We realise that this is a strong simplification and that these intentions and purposes of AI are not necessarily mutually exclusive and do not comprehensively cover all possibilities. The need for contact tracing to fight a pandemic, for example, shows that social control can be conducive to human flourishing. Similarly, the optimisation of processes and resulting profit maximisation leads to higher income and welfare, which can (but do not have to) contribute to broader human flourishing. The Venn diagram in Fig. 1 indicates that the three different purposes can intersect and overlap. However, they are recognisably different ways of approaching AI and have different ethical implications and connotations.

Intentions behind promoting AI are important to understand and evaluate perceptions of ethics and possible governance mechanisms employed to address ethical issues. These intentions do not develop in isolation but form part of a larger socio-economic, cultural and political context that influences the way a 'good society' is perceived and the role AI can play in it (Cath, Wachter, Mittelstadt, Taddeo, & Floridi, 2016). We do not wish to overstate differences between regions or underestimate levels of disagreement within political cultures, but we think it is probably safe to say that the European approach to AI aims to promote human flourishing, even where this may lead to trade-offs with efficiency or access, which may result from specific interventions, such as the EU's General Data Protection Regulation (GDPR) (*General Data Protection* Regulation, 2016; see also Kaplan & Haenlein, 2019).

In order to assess whether a response to an ethical issue is appropriate or likely to be successful, we need to not only understand the purpose of AI, but also the range of possible options used to address the issue. This paper does not offer the space to review all governance arrangements or tools that are available to implement them (see Hagendorff, 2019; Morley, Floridi, Kinsey, & Elhalal, 2019). For the purposes of this paper, we seek to understand the types and levels of activity that aim to provide governance mechanisms for AI. Below, we distinguish between measures aimed at the individual, the organisation and society.

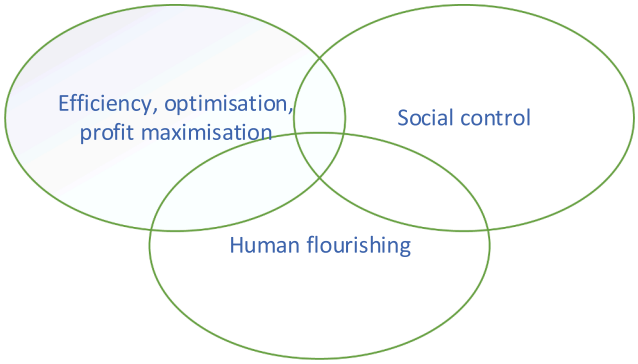

Fig. 1. Possible purposes of AI.

Governance structures that support the identification and mitigation of possible ethical issues of AI cover all levels, notably the individual, organisational and political / societal. Individual researchers and developers can make use of a quickly growing number of AI ethics frameworks originating from companies, governments or other organisations. The EU's High Level Expert Group is a pertinent example (HLEG on AI, 2019) but many others exist. Individual developers can make use of professional guidance, for example, from bodies such as the ACM or BCS (Brinkman et al., 2017). Standardisation initiatives, such as the ISO/IEC JTC 1/SC 42 - Artificial intelligence or the IEEE P7000 family of standards, consider the ethics of AI and, once agreed, can provide guidance. Development methodologies can be created or adapted to pay attention to ethical issues, for example, by integrating specific issues into the design process, such as privacy by design (Cavoukian, 2017; Hansen, 2016; Information Commissioner's Office, 2008) or more broadly by adopting an ethics by design stance (Beard & Longstaff, 2018; Iphofen & Kritikos, 2019; Martin & Makoundou, 2017).

The second level of measures provides guidance for organisations to follow or adopt. According to Clarke (Clarke, 2019b, 2019a), established mechanisms of risk management can go a long way in allowing organisations to address the ethics of AI. Organisations can employ existing impact assessment approaches such as privacy (or data protection) impact assessments (CNIL, 2015), technology assessment (Grunwald, 2009), ethics impact assessment (Wright, 2011), social impact assessment (Becker & Vanclay, 2003) or human rights impact assessment (Latonero, 2018). They can extend existing governance mechanisms, such as those used for quality assurance or data governance (British Academy & Royal Society, 2017; Khatri & Brown, 2010; OECD, 2017), and ensure these cover AI. Similarly, many organisations have established mechanisms for dealing with ethical and broader societal concerns, often discussed under the heading of corporate social responsibility (CSR) (Garriga & Melé, 2004), which can be extended to include AI and emerging technologies.

The final level is the societal and policy level, covering national and international policy and regulation. These drive a lot of individual and organisational activity and therefore play a role in governing AI. It is, therefore, not surprising that policy and regulatory mechanisms play a prominent role concerning the ethics of AI. There are existing statutory instruments, such as the GDPR (*General Data Protection* Regulation, 2016), that clearly address some of the issues that AI raises. Similarly, there are principles of human rights, that are addressed and safeguarded in international agreements, such as the Universal Declaration of Human Rights or the European Convention on Human Rights, that cover relevant rights such as the right not to be discriminated against that have relevant applications to AI. Similarly, legislation in areas such as competition law, product liability or intellectual property can have consequences for AI. One type of regulatory instrument with regard to AI is the creation of a regulator to oversee AI development and use. This can be achieved by extending the remit of existing regulators, such as data protection authorities, or by creating new bodies.

On all three of these levels, the individual, the organisational and the national / international, there can be different focus areas. AI can be looked at in general terms or specific application areas can be emphasised, such as AI in health, finance, politics, public services or

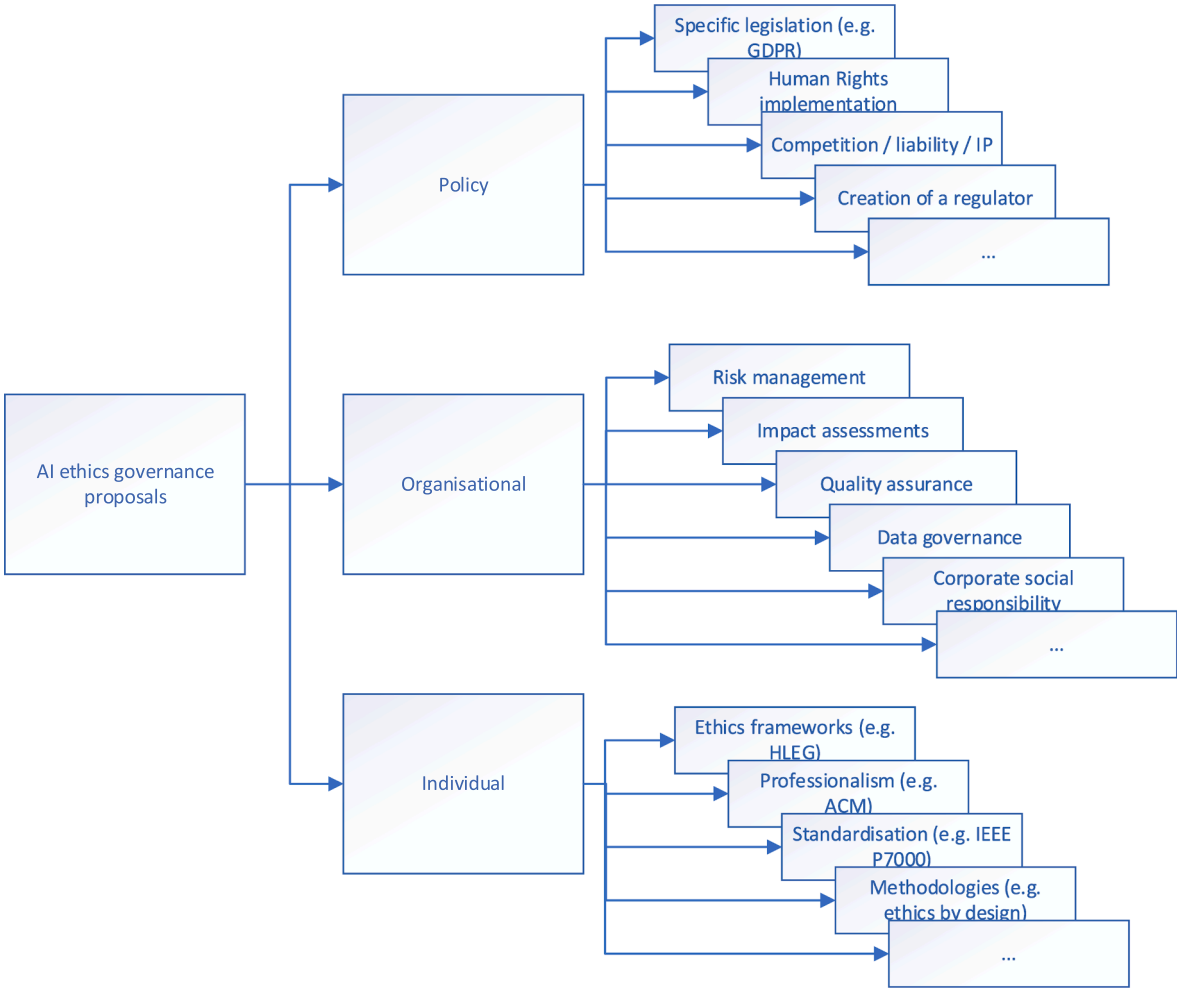

Fig. 2. Overview of proposals for governance mechanisms of AI ethics.

others. In many cases, organisational or technical tools may exist or could be developed to support governance structures.

The aim of this very brief overview of ethics and governance of AI, graphically represented in Fig. 2 (above), was to demonstrate the complexity of the topic area. The discourse of ethics and AI is currently characterised by a cacophony of voices and contributions. There is no lack of ideas or proposals. The challenge is to synthesise a manageable approach from the multitude of activities. This synthesis should start with a useful and manageable categorisation of ethical issues that lends itself to an analysis of suitable governance structures that can be applied to these issues. This paper attempts such a synthesis based on a multi-dimensional research approach, as described in the next section.

3. Methodology: A multi-dimensional approach

While there is a large and quickly growing literature on ethical issues of AI, much of it is anecdotal or speculative. In our research, we therefore aimed to combine academic rigour with detailed insights into the way in which AI is realised in society and a broad and conceptual overview of the field. No one single established methodology can achieve this. We therefore decided to use a multi-dimensional approach that involved using and combining several methods to collect and interpret data and develop an understanding of the field of ethics and AI.

The geographical focus of our study is Europe. We wanted to know whether the influence of AI on human flourishing is noticeable in the way AI is developed and deployed. In order to understand this, to gain an in-depth understanding of the social reality of AI across different application domains at present and in the future and to understand the technical, ethical and human rights implications, we undertook a multi-dimensional study comprising the following:

1 10 interpretive case studies of AI application in particular application domains and organisations
2 Five policy-oriented scenarios exploring near term (<5 years) use of emerging AI applications
3 An ethical impact analysis of AI
4 A human rights analysis of AI
5 A technical analysis of threats and vulnerabilities connected with AI.

The following figure is a graphical representation of the research approach (See Fig. 3).

A single paper such as this one cannot hope to do justice to the complexity of five different major components of a complex multi-dimensional study as presented here. Each of these has been described in detail elsewhere (Andreou et al., 2019; Macnish, Ryan, Gregory, et al., 2019; Macnish & Ryan, 2019; Patel, Hatzakis, Macnish, Ryan, & Kirichenko, 2019; Wright et al., 2019). Instead of a detailed account of all methodological considerations, we focus here on a brief overview of the different methods and why they provided the insights we required for our research objective.

The motivation for undertaking a set of case studies arose from the lack of rigorous empirical academic research of AI across application areas. While there are numerous studies of the impact of AI in particular areas, a broader understanding of AI required a set of comparable insights in different settings. We chose to undertake a set of interpretive case studies (Walsham, 1995) because they allowed us to develop a detailed understanding based on the views of individuals and organisations involved. There are certainly many more application areas than 10, but undertaking 10 studies and doing a comparative analysis (Yin, 2003) gave us the confidence of being able to develop a strong understanding across applications. We developed a case study protocol and pilot tested it during the summer of 2018. The empirical work was undertaken in 2018 and 2019. We interviewed a total of 22 stakeholders across the 10 case studies. For each case, we furthermore undertook background research on the organisation in question as well as the field in which the case study was undertaken (e.g., AI in finance, agriculture). The analysis was undertaken collaboratively using NVivo Server 11. The partners wrote up case studies following an agreed template, cross-reviewed and published on our website (Macnish, Ryan, & Stahl, 2019). All case studies were furthermore developed as stand-alone publications whose references are listed in Table 1.

In order to broaden our understanding further, but also to go beyond the description of current technologies, we decided to develop a set of five scenarios. The social domain in which AI is employed was discussed simultaneously for case studies and scenarios (see table below). This served to broaden the range of insights.

The scenarios were constructed with a specific focus on providing applicable insights that could help decision-makers, notably those working in policy development, to develop and implement governance mechanisms. Based on the rich history of scenario methods (Andersen &

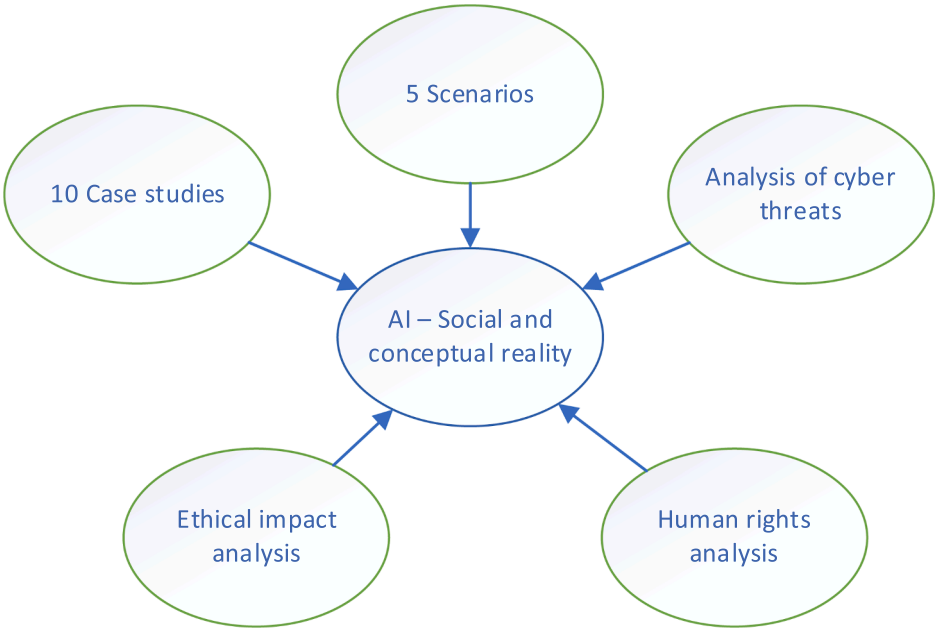

Fig. 3. Components of the multi-dimensional research approach underpinning this research.

Table 1
Social domains of case studies and scenarios (CS:= Case Study; SC:= Scenario).

No.	Social Domain of Case Study / Scenario	Reference
CS01	Employee monitoring and administration	(Antoniou & Andreou, 2019)
CS02	Government	(Ryan, 2019a)
CS03	Agriculture	(Ryan, 2019b)
CS04	Sustainable development	(Ryan & Gregory, 2019)
CS05	Science	(Jiya, 2019b)
CS06	Insurance	(Kancevičienė, 2019)
CS07	Energy and utilities	(Hatzakis, Rodrigues, & Wright, 2019)
CS08	Communications, media and cybersecurity	(Macnish, Inguanzo, & Kirichenko, 2019)
CS09	Retail and wholesale trade	(Macnish & Inguanzo, 2019)
CS10	Manufacturing and natural resources	(Jiya, 2019a)
SC01	Social care	All scenarios are described in detail in (
SC02	Information warfare	Wright et al., 2019)
SC03	Education	
SC04	Transportation	
SC05	Predictive policing	

Jaeger, 1999; Boenink, Swierstra, & Stemerding, 2010; Cairns & Wright, 2017; Ramirez, Mukherjee, Vezzoli, & Kramer, 2015), we applied a modified methodology which we called "policy scenarios" (Wright, Stahl, & Hatzakis, 2020). At its core, a policy scenario aims at a relatively short time horizon that is of relevance to policymakers caught up in election cycles. We aimed to develop scenarios that are relevant in several years. A second important aspect of policy scenarios is close stakeholder engagement. For each scenario, we organised a face-to-face workshop that included stakeholders and involved them in the revision and refinement of the scenarios. All scenarios were peer reviewed and made publicly available in June 2019 (Macnish, Wright, & Jiya, 2020; Wright et al., 2019).

While the work on case studies and scenarios gave us the confidence to be in a position to understand a broad range of ethical aspects of AI in social situations, this knowledge had to be based on and complemented by current academic and other debates about ethical issues and human rights implications. We furthermore realised that it was important to understand the technical aspects of AI, notably questions of security and vulnerabilities of AI systems that have potentially large implications for the use and social impact of these technologies.

The ethical and human rights analyses were undertaken as desk research drawing on appropriate sources of their various disciplines (i.e., philosophy and human rights law). We also conducted a study of security issues, dangers, and implications of the use of data analytics and artificial intelligence. We examined:

- ways in which machine-learning systems are commonly mis-implemented or mis-used (and recommendations on how to prevent this from happening);
- ways in which machine-learning models and algorithms can be attacked (and mitigations against such attacks);
- how artificial intelligence and data analysis methodologies and technologies might be used for malicious purposes.

This initial review of security-related issues and possible sources of harm provided the starting point for further and ongoing studies of specific vulnerabilities of AI. The purpose of this task was to provide a baseline understanding of the current capabilities and applications of machine learning, including examples of potential malicious uses of machine learning techniques, and the implications of attacks against systems powered by machine learning.

4. Findings and Discussion

The purpose of this paper is to contribute to the discussion about how to identify, interpret and address ethical issues arising from AI applications. In order to achieve this, we need a comprehensive overview of both conceptual and empirical insights into AI and its use. We therefore draw on a range of research activities outlined below. The underlying empirical studies were not re-analysed but taken as the starting point for compiling and, more importantly, for categorising these issues. The important contribution to the AI ethics discourse that this paper makes is in the conceptualisation and proposed narrative which is based on the overall findings. It is therefore only possible to provide a high-level overview of the empirical work, all of which is published elsewhere.

Our research showed that the breadth of ethical issues discussed in the literature is reflected to a large extent in organisational practice. A cross-case analysis extracting the moral issues that were raised in the different case studies showed a large number of ethical issues, many of them recurring, as shown in the following table (See Table 2).

This table listing the ethical issues we encountered is the result of our data analysis and represents our interpretation and categorisation of what respondents shared with us. The terms used to denote the individual ethical issues were discussed and agreed during the data analysis. The results of the analysis from the case studies were supported by the first round of our Delphi Study, which targeted experts in AI and big data (Santiago, 2020). Implemented as an online survey, the first set of questions was e-mailed to 231 experts, 50 per cent of whom were women. We received 145 responses.Following review of the data and data cleansing, 41 responses contained sufficient information to warrant analysis. The first (open-ended) question covered the same ground, asking "What do you think are the three most important ethical or human rights issues raised by AI and / or big data?" Fig. 4 shows the most frequently given answers:

One important insight from our empirical work was that the many ethical issues discussed in the literature are reflected in practice. Some issues are almost ubiquitous, such as those related to privacy and data protection. This paper does not offer a detailed analysis of the issues, nor of the exact differences between the case study and Delphi study findings. One key point from our findings is that clearly recognisable issues are well covered are roughly consistent. The Delphi study is understandably broader than the case studies that focused on organisational practice, including issues such as 'awakening' of AI that do not play a role in current implementations of AI. Our conceptual review of the ethics of AI and the analysis of our empirical work agreed to a large extent, showing that the literature covers the same issues of individuals and organisations working with AI (See Fig. 4).

However, while it was possible to analyse our data using widely accepted terms, we note that the local meaning of these terms varied widely. The meaning of the term privacy, for example, in a medical diagnostic context, in the use of social media for logistics prediction or in the case of agricultural optimisation, differs greatly. Privacy is "an inherently heterogenous, fluid and multidimensional concept" (Finn, Wright, & Friedewald, 2013, p. 26) and can be divided up into a number of sub-types (Koops et al., 2017; Solove, 2002). In general, privacy in the AI ethics discourse seems to refer to data privacy or information privacy which touches on various other types of privacy (ibid). Other theoretical positions on privacy (Tavani, 2008) are not widely employed in the AI discourse. While this connection is rarely made explicit, there is an underlying assumption that implementing data protection mechanisms is the way to ensure privacy, although this has been contested (Macnish, 2020). This position is reflected in our findings where the frequent references to privacy as an issue were accompanied by a strong emphasis on often technical data protection measures. However, it is important to note that privacy risks take very different forms in these contexts, requiring different technical and organisational measures to ensure compliance with data protection legislation, but also to ensure that broader ethical issues are covered.

Interestingly, our analysis of human rights concerns, undertaken in parallel with the case studies, found that the human rights issues that can be found in the literature are closely related to and overlap with the

Table 2
Distribution of ethical issues across case studies.

Ethical Issues	CS01 Employee monitoring and administration	CS02 Government	CS03 Agriculture	CS04 Sustainable development	CS05 Science	CS06 Insurance	CS07 Energy and utilities	CS08 Communications, media and cybersecurity	CS09 Retail and wholesale trade	CS10 Manufacturing and natural resources
Access to SIS	●	●	●	●		●	●			●
Accuracy of Data		●	●	●				●	●	●
Accuracy of Recommendations			●	●		●		●	●	
Algorithmic Bias					●	●		●	●	●
Discrimination	●				●	●		●		●
Economic		●	●	●				●		
Employment			●	●		●		●		
Fairness			●	●	●					
Freedom							●			
Human Contact			●							
Human Rights					●			●		●
Individual Autonomy								●	●	
Inequality	●		●	●						
Informed Consent	●		●	●		●	●	●		●
Integrity					●					●
Justice		●	●	●				●		
Ownership of Data		●	●	●		●				●
Military, Criminal, Malicious Use	●			●				●	●	
Power Asymmetries	●	●		●	●		●	●		
Privacy	●	●	●	●	●	●	●	●	●	●
Responsibility	●		●	●		●		●		
Security	●	●	●	●	●	●		●	●	●
Sustainability			●	●						
Transparency	●		●	●	●	●	●	●	●	●
Trust	●	●	●	●		●	●	●		
Use of Personal Data	●	●	●	●	●	●		●	●	

Fig. 4. Delphi survey responses covering the most important issues of AI and big data (Santiago, 2020).

ethical issues described above. The human rights gap analysis was based on a combination of methodological approaches emanating from the interim results of case studies and scenarios, preliminary in-depth interviews on cyberthreats and ethics as well as desktop research. Initial findings were cross-checked and combined to identify the most burning challenges to human rights in the digital world and start formulating solutions. The findings were mindful of the diversity and breadth of data collected and expected results. A list of commonly perceived human rights issues and challenges in the digital world was derived from the process, encompassing general and specific considerations as follows:

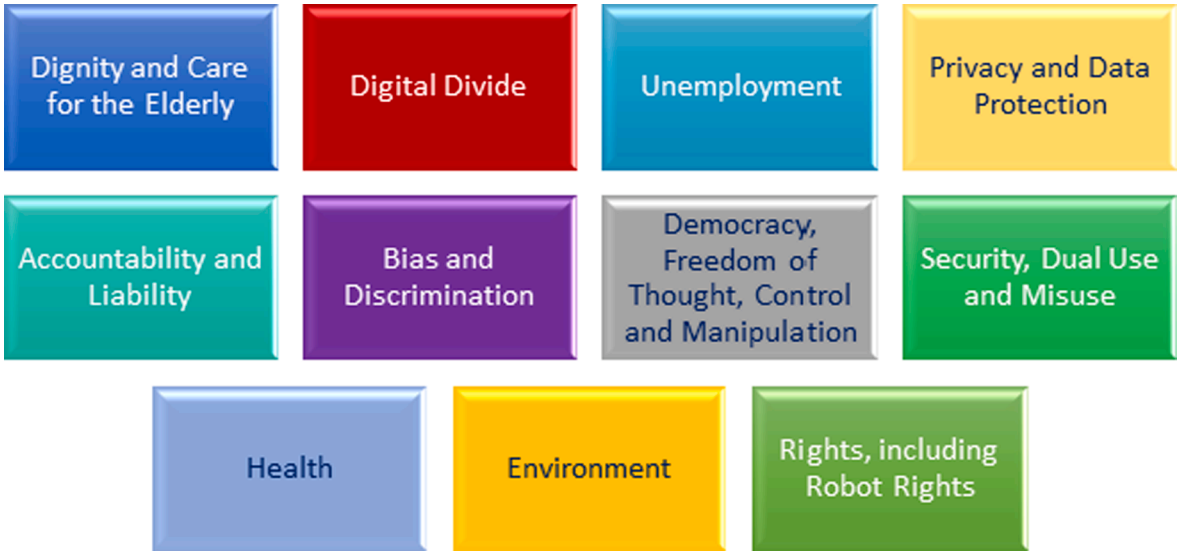

Fig. 5. Key human rights issues in AI.

(See Fig. 5)

In recognition that the relationship between ethics and human rights is complex (van Est & Gerritsen, 2017) and that human rights have an ethical core that must be safeguarded, the human right considerations were analysed on a scale of concepts and notions: (See Fig. 6)

Formalising human rights at different levels is based on the recognition that humans have ethical rights that need to be explicitly safeguarded. Human rights declarations such as the Universal Declaration are political statements, but they have long found their way into positive law. Europe, for example, has the European Convention on Human Rights (ECHR) at the international level, which originates from the Council of Europe but is authoritative in the EU legal order. The EU has its own legally binding human rights instrument, the EU Charter of Fundamental Rights (EUCFR), whose scope can be said to be wider and deeper than the ECHR, as evidenced in the table below. This is relevant because it means that many of the ethical issues identified by our research are not just ethical issues, but, where they relate to human rights questions, may well be open to adjudication in a court of law. The following overview of human rights that may be affected by AI is therefore interesting, as it shows where and how these human rights are enshrined in existing human rights instruments, namely the ECHR and the EUCFR (See Table 3).

All of these rights and freedoms were raised as a concern in at least one of our case studies or scenarios, thus further demonstrating the broad range of concerns. The reason for including this table in the text was that for readers not familiar with the detail of our case studies, it shows a list of human rights that are easy to associate with AI use and demonstrates that there are legal instruments that could deal with these. As several overlaps were identified in terms of, for example, relevant legal instruments and types of solution proposed, these were intertwined in the establishment of a 'blueprint' of current and future practice in relation to the interrelationship of human rights, law, ethics and smart information systems, those systems that have AI and big data analytics at their core (Stahl & Wright, 2018).

Human rights touch on or are directly part of many of the issues related to AI. Our analysis shows the breadth of human rights concerns using the European human rights framework. It seems plausible that a more direct application of human rights legislation to AI can provide some clarity on related issues and point the way to possible solutions. It is therefore not surprising that there are many voices that point to the application of human rights to AI as a key way of addressing these issues (Access Now Policy Team, 2018; BSR, 2018; Committee on Bioethics (DH-BIO), 2019; Council of Europe, 2019; Latonero, 2018; World Economic Forum, 2019).

Table 3
Selection of Rights and Freedoms (EU) with potential relevance to AI.

Rights	ECHR	EUCFR
Right to human dignity		Article 1
Right to the integrity of the person		Article 3
Right to liberty and security	Article 5	Article 6
Right to respect for private and family life	Article 8	Article 7
Protection of personal data		Article 8
Freedom of thought, conscience and religion	Article 9	Article 10
Freedom of expression and information	Article 10	Article 11
Prohibition of discrimination	Article 14; Article 1, Protocol 12	Article 21
Right of property	Article 1, Protocol 1	Article 17
Right to education	Article 1, Protocol 2	Article 14
Right to free election	Article 3, Protocol 1	Articles 39–40
Freedom of movement	Article 2, Protocol 4	Article 45
Freedom to choose an occupation and right to engage in work		Article 15
Freedom to conduct a business		Article 16
Rights of the child		Article 24
Rights of the elderly		Article 25
Integration of persons with disabilities		Article 26
Right to health care		Article 35
Consumer protection		Article 38
Right to good administration		Article 41
Right of access to documents		Article 42
Freedom of movement and residence		Article 45

4.1. Classification of ethical issues

This paper has so far confirmed that there are many potential ethical issues related to AI. We have furthermore shown that there are numerous governance approaches, including human rights legislation, that can address many of these issues. An ongoing problem, however, is the complexity of the landscape, the fact that there are too many ethical issues and ways of addressing them to allow scholars or practitioners to keep an overview.

Fig. 6. Concepts and notions employed for human rights analysis of AI.

In the course of our research and engagement with the ethics of AI, it emerged that there are types of issues that can be clustered in a way conducive to finding appropriate responses. We propose three types of issues that have significant specificities to allow them to be clustered: specific issues of machine learning, general questions about living in a digital world and metaphysical questions.

4.1.1. Specific issues of machine learning

Machine learning and the various techniques used to achieve it have some characteristics at the core of particular ethical issues. Two of these characteristics seem most likely to raise ethical issues: First, many of the current machine-learning techniques are opaque, which means that even experts with relevant equipment cannot determine why and how inputs are transformed into outputs, e.g. how exactly a personal profile leads to a classification in terms of a mortgage application or parole decision. Second, these systems require access to big amounts of data for training and validation purposes.

Resulting ethical issues are, for example, those having to do with bias and discrimination (Johnson et al., 2019; Macnish, 2012), which can arise on the basis of undetected biases included in the training data. Further specific issues are linked to the use of data, which may be personal data and which may allow additional insights into individual personal behaviour through AI and big data analysis. Questions around privacy and data protection therefore arise on this level as do concerns about security and integrity of systems, algorithms and data (Stahl & Wright, 2018).

These issues are common to most specific applications of AI. Algorithmic biases, discrimination, security and transparency are issues that are directly linked to the characteristics of machine learning. Others, such as quality and accuracy of data are closely related. Our case study analysis suggests that these issues are prominent among the AI users in organisations. They materialise and present themselves in specific and context-dependent ways. From the perspective of this paper, an interesting feature is that these issues can become visible and are – at least to some degree – capable of being addressed on the project or organisational level. While these ethical issues represent some of those most frequently mentioned in our empirical research, it is important to see that many of the ethical issues do not seem to be linked to the technical properties of AI as machine learning but point to the broader socio-economic context in which these technologies are used.

4.1.2. General questions about living in a digital world

Our second category of ethical issues includes those that have less to do with the specific capabilities of AI and more with the way societies use technologies many of which incorporate elements of machine learning or other AI techniques. These issues play to a general feeling of unease with the way in which industrialised societies develop and the role that technology plays in promoting certain developments and inhibiting others. These issues are currently discussed in the context of AI because the expectation is that AI will greatly influence them, but they are better understood as questions that relate to how modern societies organise themselves using technologies such as AI. A key characteristic of these technologies seems to be autonomy, i.e., the ability to act without direct human input. This, combined with higher levels ability to detect patterns and act accordingly, can lead to the replacement of humans by machines. Another key feature of digital technologies with high ethical relevance is that they increasingly constitute the environment in which humans live. The malleability of digital technologies means that our realities can easily be changed. Maybe even more importantly, the constitutive element of digital technologies in modern social reality means that the owners and controllers of these technologies become immensely powerful in many different ways.

Key examples of these issues are the influence of technology on economic and political power, the future of warfare or distribution of costs and benefits of AI. The high-profile example of the misuse of social media data for purposes of the manipulation of democratic elections in the case of Cambridge Analytica (Isaak & Hanna, 2018) is the most visible case in point. But while Cambridge Analytica represents a relatively clear-cut case of misuse of AI and big data, there are broader questions about the economic and resulting political power amassed by the tech industry (Macnish & Galliott, 2020). The market capitalisation and therefore economic power of the big tech companies is now such that many observers are increasingly worried about the mere possibility of oversight of these actors. One resulting question is that of justice of distribution of costs and benefits. Big tech companies have the technical infrastructure and know-how to create ever-larger data sets and benefit from these whereas smaller competitors lack the means to catch up. The role of consumers and end users at present is predominantly passive; they produce data and consume services, but have little control over the use of their data.

Other larger-scale societal concerns have to do with the question about what technology can and should do. A high-profile example of this is the use of AI-driven autonomous weaponry that has the potential to change the face of modern warfare (Defense Innovation Board, 2019; Sparrow, 2009). But there are many other examples where it is not clear-cut what machines can and should do, e.g., with regard to autonomous vehicles or care robots (Decker, 2008; Sharkey & Sharkey, 2010; Bernd Carsten Stahl & Coeckelbergh, 2016).

Our empirical findings support the relevance of these issues. In fact, and somewhat surprisingly, ethical issues in this category constitute the majority of issues. Economic consequences, employment, fairness, freedom, the ability of having human contact, individual autonomy, inequality, integrity, justice, ownership, military use, power asymmetry, responsibility and sustainability all fall into the category.

4.1.3. Metaphysical questions

The final set of questions concerns what machines should be allowed to do and points to some of the deeper philosophical and metaphysical questions about the future of AI and autonomous machines and their relationship with humans. There has been a long-standing discussion about the change of human nature due to machines, the emergence of cyborgs (Latimer, 2017) and transhumans (Livingstone, 2015). In parallel, there has been discussion whether machines can ever become sentient or conscious (Carter et al., 2018; Dehaene, Lau, & Kouider, 2017), whether there will be a singularity (Kurzweil, 2006) at which point machines will develop superintelligence (Bostrom, 2016).

These developments are controversially discussed and highly contentious. They are based on the idea of general AI (Baum, 2017). In this paper, we do not take a position on whether these developments are likely or even possible. We also refrain from taking a position on whether current narrow AI can lead to general AI or whether a fundamentally different approach would be needed. In our case studies, we found no evidence of their being considered a current priority, but in our scenarios, there are examples of technologies that come close to this category. In the Delphi study, there was reference to 'awakening of AI', which falls into this category. The reason for including them here is thus less their current practical relevance and more the fact that they are prominent and highly visible in science fiction, the media and increasingly in policy discussions – which, of course, in no way reduces the legitimacy of the concerns they raise.

The following figure aims to summarise these three types of ethical issues of AI (See Fig. 7).

The classification suggested in the above figure serves as a starting point to explore how governance structures are positioned to address ethical issues.

4.2. Governing the three types of issues

The purpose of classifying ethical issues in the previous section was to render the broad array of ethical issues more manageable and impose some order on the chaos of AI ethics. In this section, we now look at how this can help identify suitable governance structures. Before we start

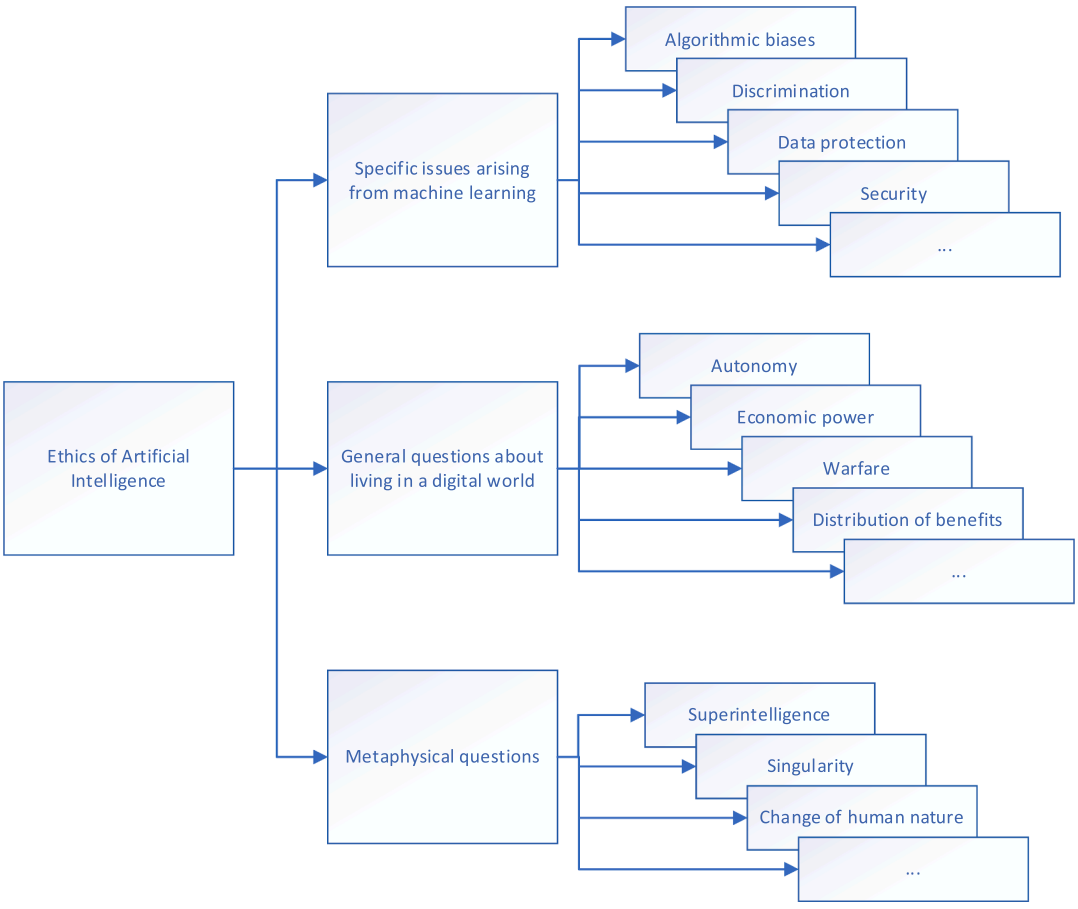

Fig. 7. Classification of ethical issues of AI.

this, we should make it clear that we understand that the above classification is analytic in nature and represents one possible way of thinking about AI ethics among many. We hope that the distinction makes sense and is plausible, but we are happy to concede that it is not exclusive. Privacy, to take an example, materialises on the local and project level, but is seen as a societal question (Roessler & Mokrosinska, 2015), subject to legislation and regulation and arguably based on human needs arising from human nature (Locke, 2010). With this contingent nature of the classification in mind, we now look at how this translates to governance mechanisms.

In our discussion of governance, we take the position that the desirable purpose of AI should be to support human flourishing. As discussed earlier, this does not preclude uses of AI for efficiency maximisation or control but implies that the overall aim of development and deployment of the technology should be to improve human lives. This position has the disadvantage of muddying the waters, raising difficult questions of what counts as flourishing and who determines this. It is also likely to require the balancing of competing goods, values and interests. But it is arguably the purpose of AI research, development, funding and application on which most citizens in democratic states, and maybe most human beings, can agree. It is also aligned with the EU's AI policies. Based on an agreement that AI is to support human flourishing, the question then is how ethical issues can be addressed and which governance structures can help us deal with them.

To return to the suggested classification of ethical issues (Fig. 6) and compare this with the earlier description of governance proposals (Fig. 2), it is easy to see that there is no simple and/or linear relationship. The two sets of classification do not align, but can be better understood as a matrix, where all levels of governance (individual, organisational, policy) can refer to all types of ethical issues (specific issues, general questions, metaphysical questions).

In many cases, the issues that we classified as specific arising from machine learning appear easiest to address. They can be subject to the application of existing governance mechanisms. Data protection or security issues provide a good example. Data protection is furthermore governed by legislation and regulation, even though the details of these legislative requirements differ significantly between jurisdictions. Not only is there ample regulation, there are also structures such as data protection impact assessments (CNIL, 2017a, 2017b) or standards such as the ISO 27000 family that allow structured approaches. Similar observations apply to other examples of the specific issues arising from machine learning, such as algorithmic biases and subsequent discrimination. Unlike data protection, these are less comprehensively defined and regulated. A key problem here is that it may be difficult to understand what exactly the issue is. There are several high-profile examples of these problems, e.g., where a system evaluating job applications leads to discrimination on the basis of gender or where predictive policing or parole decisions lead to discrimination on the basis of race. There is broad consensus that discrimination on the basis of gender or race is immoral and to be avoided. However, it is not clear whether other systems discriminate on the basis of other properties that are less high-profile and of which individuals are unaware. It is conceivable that a medical diagnostic system might lead to discrimination on the basis of blood type or a dating system would prefer individuals on the basis of their height. This raises questions about how we would even know how to check for relevant characteristics and, even if confirmed, whether and why this would constitute an ethical issue.

These issues can be described in terms of transparency and explainability. At the heart of the problem is the fact that humans expect explanations that they can understand, whereas machine learning

algorithms classify individual cases on the basis of complex statistical calculations that defy simple translations into forms that are accessible to humans (Mittelstadt et al., 2019).

This tension between human reasoning and machine-learning data processing is a reason why transparency is among the most visible ethical issue linked to AI. The prominence of transparency as a perceived issue may explain why the same term is used to denote an ethical principle that pervades the AI ethics guidelines (Jobin, Ienca, & Vayena, 2019). The ethical principle of transparency is described as a way to minimise harm, improve AI, foster trust and improve AI (ibid). It is also linked to dialogue, participation and democracy. Elsewhere we have tried to unpack the normative implications of ethical principles, including the principle of transparency (Ryan & Stahl, 2020). Of course, while transparency may be perceived as obligatory, it does not in itself resolve the ethical or privacy risks at issue. Transparency only gets us part of the way to solutions.

The example of transparency indicates that the ethical issues arising from AI are not clearly defined. The perception that something constitutes an ethical issue does not by itself clarify whether this perception is accurate or on which grounds such an evaluation would be possible. It says little about the exact nature of the issue and its status from a theoretical ethical perspective.

In the case of transparency, there is now a quickly growing research community that focuses on such questions of transparency and explainability of AI (USACM, 2017). While these questions remain open, there is at least reason to hope that they can be addressed on the local level by designing, testing and reviewing systems using appropriate methods, relying on professional expertise of developers and users and organisational capacity to develop the relevant expertise. The ISO is developing guidance for developers (e.g., TR 24368). Discussion within ISO has recognised that AI is not neutral, that values embedded in algorithms are intentionally or inadvertently shaped by the programmers', operators', and third-parties' own worldviews and cognitive biases and that ethical violations could also result from AI deployed or developed prematurely, applied without proper consideration of the ways it could negatively impact individuals or society.

The second set of issues, those relating to general questions about living in a digital world, are not capable of being dealt with on the micro or meso level of the individual or organisation. This does not mean that there are no existing governance mechanisms that could be applied to AI. Nation states can make use of policy, legislative and regulatory options and this is happening at large scale with regard to AI. At the same time, there are many existing structures that may well be suited to dealing with at least some of the challenges related to AI, even if such structures can be criticised as having failed to deal adequately with the ethical, privacy and societal issues. Consumer protection legislation, competition law, anti-trust law, data protection law as well as human rights legislation may well be capable of regulating AI and its consequences, at least in part. Similarly, existing regulators may make a claim that they can address the challenges of AI. This suggests a close affinity between this type of ethical issue and the policy level of governance options. The key to the resolution of these issues does indeed seem to be on a policy level, as they touch on constitutive questions of modern societies (e.g., who can own what, how do we distribute wealth and risks, the asymmetries in the power wielded by the big tech companies through their algorithms). Ethical issues of AI may be subject to existing regulation, but it is entirely possible that additional regulation and legislation will be required to facilitate human flourishing. Nevertheless, the policy level needs to be supplemented by actions on the organisational and individual levels.

The most difficult area to govern is the metaphysical. Questions of general AI and its consequences are partly technical and empirical in terms of what technology can achieve in the hands of the big tech companies, intelligence agencies and cyber attackers (for example). They are also partly philosophical insofar as they refer to basic concepts and arguments. To some extent, e.g., where transhumanism displays religious characteristics [overcoming the body, the distinction between (mortal) body and (immortal) essence of the human], they may be most suitably dealt with by theology. While it may be difficult to deal these issues, they did not receive much attention in our empirical findings. But that does not mean that proactive attention to these issues would be misplaced or that it might not be advisable to think about early warning signs or trigger points where more attention should be paid to such issues.

Insights from our case studies suggest that this way of looking at ethical issues and governance clarifies current activities. This paper does not provide the space for a comprehensive analysis, so suffice it to say that the organisations in our case studies focused on the issues arising from machine learning and on established and practical ways of addressing them. The dominant topics were security and data protection, which companies dealt with in ways to ensure they met their legal responsibilities. They focused on technical approaches to achieve these. In addition, companies developed oversight mechanisms and reflective capabilities, e.g., by instituting ethics boards or engaging with stakeholders. Respondents were aware of the policy issues but not actively engaged with them. Similarly, as noted, the metaphysical issues did not play a role in the social reality of the case studies.

One interesting point to observe is that in our case study research, there was little reference even to most of the governance mechanisms that we categorised as aimed at the organisational level. Activities like risk management, impact assessments or human rights integration into company policies were not mentioned by our respondents. This does not mean that the companies did not engage in them or that they were unsuccessful, but that in the responses to our questions, the respondents did not associate them with ethics of AI or big data. It will therefore be important for future studies to evaluate the effectiveness of these governance approaches.

5. Conclusion

In this paper, we have reported some findings and their implications from a multi-dimensional research study into the ethics of AI. Our starting point was the highly complex and often overwhelming AI ethics discourse that motivated our attempt to bring better empirical insights but also conceptual clarity to the discussion.

One outcome of this work was the classification of ethical issues into specific issues arising from machine learning, general questions about living in a digital world and metaphysical questions. We hope that this classification is helpful in identifying issues and finding or developing appropriate governance mechanisms.

Our discussion and mapping of existing proposals for AI governance to the three classes of ethical issues give rise to suggestions for further research as well as organisational practice. One important observation refers to the concept of AI and the implications of the use of a particular concept for subsequent insights. In this study, we started with an inclusive view of AI that covers not only narrow AI and machine learning, but also broader socio-technical systems incorporating AI techniques and artificial general intelligence. Our empirical findings show that the focus of attention in current use of AI is on machine learning and, to some degree, on broader socio-technical systems. This raises the question whether it would be desirable to limit the scope of reflection, e.g., by focusing exclusively on machine learning. We suggest that the term AI is broader than machine learning and a rich conceptualisation of AI and its ethical consequences needs to take this into consideration.

This paper has shown that a key next step in the AI ethics discourse needs to be a more detailed and thorough mapping exercise that not only lists and clearly defines ethical issues but explores to what degree these are novel or manifest inadequate regulatory structures and in need of novel solutions. As we have suggested above, the broad array of extant governance structures *may* be able to cover many of the issues raised by AI. Whether they do so adequately is another question. Where no remedies currently exist, or existing ones are insufficient, further and novel

solutions may well be required. The work presented here suggests that the AI ethics debate can benefit from a different perspective, which we have offered above. This new perspective is based on a novel categorisation of ethical issues that point towards next steps and the bigger question about how the different types of issues can be addressed, which can be used to generate recommendations for organisations. Many companies are interested in exploiting the advantages of AI but are unsure about how to deal with societal and ethical issues. Our research shows that there are numerous existing activities, many of which are related and interlinked, that will go a long way towards showing that a company is serious about engaging with these questions. The integration of CSR structures into organisations in a way that takes into account the organisation's research and development is one such step. An explicit commitment to human rights and the adoption of processes designed to integrate human rights into corporate processes would be another such step. These suggestions are, of course, no guarantees that nothing will go wrong, but they are serious steps in ensuring that the corporate culture supports human flourishing, through AI and otherwise.

A similar conclusion offers itself to policymakers and decision-makers. Just as companies can look through a portfolio of existing governance mechanisms, policymakers should take stock of the adequacy of existing policy and regulatory options while developing specific steps to address AI. Some of the issues discussed in the context of AI and the general questions arising in our increasingly digital world have little to do with AI in the narrow sense. Questions of justice, distribution and power may be exacerbated by particular technologies but exist independent of them. As a consequence, it may well be that existing legislation or regulatory bodies are in a good position to deal with at least some of these questions and the task for policymakers is to ensure that existing policy and the adequacy of its application are taken into account in the rush to develop new policy.

The nature of integrating human rights into organisations and the creation of policy provides another pointer to lessons learned from our work. The way we deal with the ethics of AI will need to be sensitive to the conceptually challenging and changing nature of the technologies in question and the social perceptions they engender. We should simply not assume that we can provide a permanently stable definition of AI or of the ethical issues related to it. Instead, we need to embrace a world where concepts are changing and contested, where moral preferences change over time, where scientific, media and political discourses dynamically interact and where impacts of new technologies such as AI need to be adequately assessed. Ethical positions based on process and exchange, such as discourse ethics (Mingers & Walsham, 2010; Rehg, 2014), are well suited to reflect the need for the ongoing negotiations of facts and values needed to make new technologies work in society.

Our work immediately suggests further areas of study. The case studies, for example, while covering a breadth of activities, could be expanded to be even more comprehensive. Our approach was open and exploratory. Future empirical work could also test the adequacy of existing governance mechanisms as well as proposals for new ones, their impacts and possible side effects (e.g., a loss of social trust in existing mechanisms). A direct mapping of ethical issues and governance proposals would be helpful in assessing the effectiveness of the governance proposals, an important factor in developing these further. Further research should also be undertaken to broaden the geographical scope of the study. Our European focus means that we simply took for granted certain aspects of the AI ecosystem that may not be applicable elsewhere. This includes the strong legal data protection regime of the European Union or the existing regulatory, industrial and professional settings that shape the way in which AI is developed, deployed and used. The international nature of AI renders necessary such further work.

The high-level view of the AI ethics debate presented in this paper should make an important contribution to theory and practice. We believe that our proposed categorisation is helpful for both scholars and practitioners. It offers a way to navigate the complexities of the AI debate and helps individuals, organisations and policymakers to find the most suitable ways of moving on. It also shows that AI ethics is not a problem to be 'solved' in the sense that there are clear solutions that will make problems go away. Instead, uncertainties around definitions, ethics and values can form the basis for a creative but unpredictable journey towards a desirable future where new technologies including AI are conducive to human flourishing.

This paper provides a clearer, more structured and inclusive narrative of ethics of AI. It draws on a set of different research activities to develop a way of thinking about AI and ethics that covers the broad range of technologies that fall under the heading of AI, that covers the manifold ethical issues associated with AI and that provides a theoretical ethical position that can be used to reflect on all of these. It is clear, however, that this is only the basis for further work. The difficult questions about what stakeholders need to do, about which laws and regulations need to be updated or developed, about definitions of professional or organisational responsibilities etc. still need to be addressed. Undertaking these next steps is important and urgent. It will also be greatly helped by an empirically based conceptual view of ethics and AI as we offer in this paper.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 786641 (SHERPA).

References

Access Now Policy Team. (2018). The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems. Access No. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.

Adam, A. (2001). Computer ethics in a different voice. *Information and Organization, 11*(4), 235–261.

ALLEA & Royal Society. (2019). Flourishing in a Data-enabled Society (ALLEA Discussion Paper No. 4). https://www.allea.org/wp-content/uploads/2019/06/DiscussionPaper_DataGov_Digital.pdf.

Anabo, I. F., Elexpuru-Albizuri, I., & Villardón-Gallego, L. (2019). Revisiting the Belmont Report's ethical principles in internet-mediated research: Perspectives from disciplinary associations in the social sciences. *Ethics and Information Technology, 21*(2), 137–149. https://doi.org/10.1007/s10676-018-9495-z.

Andersen, I.-E., & Jaeger, B. (1999). Scenario workshops and consensus conferences: Towards more democratic decision-making. *Science and Public Policy, 26*(5), 331–340.

Andreou, A., Shaelou, S. L., & Stahl, B. (2019). D1.5 Current Human Rights Frameworks. https://doi.org/10.21253/DMU.8181827.v1.

Antoniou, J., & Andreou, A. (2019). Case Study: The Internet of Things and Ethics. *ORBIT Journal, 2*(2), Article 2.

Aristotle. (2007). The Nicomachean Ethics. Filiquarian Publishing, LLC.

Asilomar Conference. (2017). Asilomar AI Principles. Future of Life Institute. https://futureoflife.org/ai-principles/.

Baum, S. (2017). *A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy* (SSRN Scholarly Paper ID 3070741). *Social Science Research Network*. https://papers.ssrn.com/abstract=3070741.

Beard, M., & Longstaff, S. (2018). Ethical By Design: Principles For Good Technology. *THE ETHICS CENTRE*.

Beauchamp, T. L., & Childress, J. F. (2009). Principles of Biomedical Ethics (6th ed.). OUP USA.

Becker, H. A., & Vanclay, F. (2003). *The International Handbook of Social Impact Assessment: Conceptual and Methodological Advances*. Edward Elgar Publishing.

Bevir, M. (2012). Governance: A Very Short Introduction. OUP Oxford.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., … Winfield, A. (2017). Principles of robotics: Regulating robots in the real world. *Connection Science, 29*(2), 124–129. https://doi.org/10.1080/09540091.2016.1271400.

Boenink, M., Swierstra, T., & Stemerding, D. (2010). Anticipating the Interaction between Technology and Morality: A Scenario Study of Experimenting with Humans in Bionanotechnology. *Studies in Ethics, Law, and Technology, 4*(2). https://doi.org/10.2202/1941-6008.1098.

Bostrom, N. (2016). *Superintelligence: Paths*. Dangers, Strategies (Reprint edition): OUP Oxford.

Braithwaite, J., & Drahos, P. (2000). *Global Business Regulation*. Cambridge University Press.

Brey, P. (2000). Disclosive Computer Ethics: Exposure and Evaluation of Embedded Normativity in Computer Technology. CEPE2000 Computer Ethics: Philosophical Enquiry. CEPE2000 Computer Ethics: Philosophical Enquiry, Dartmouth College. http://ethics.sandiego.edu/video/CEPE2000/Responsibility/Index.html.

Brinkman, B., Flick, C., Gotterbarn, D., Miller, K., Vazansky, K., & Wolf, M. J. (2017). Listening to Professional Voices: Draft 2 of the ACM Code of Ethics and Professional Conduct. *Commun. ACM, 60*(5), 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. https://royalsociety.org/~/media/policy/projects/data-governance/data-management-governance.pdf.

BSR. (2018). *Artificial Intelligence: A Rights-Based Blueprint for Business Paper 3: Implementing Human Rights Due Diligence* [Working paper]. *BSR.*

Bynum, T. W. (2006). Flourishing Ethics. *Ethics and Information Technology, 8*(4), 157–173.

Bynum, Terrell Ward, & Rogerson, S. (2003). *Computer Ethics and Professional Responsibility: Introductory Text and Readings*. WileyBlackwell.

Cairns, G., & Wright, G. (2017). *Scenario Thinking: Preparing Your Organization for the Future in an Unpredictable World*. Springer.

Capurro, R. (2006). Towards an ontological foundation of information ethics. *Ethics and Information Technology, 8*(4), 175–186. https://doi.org/10.1007/s10676-006-9108-0.

Carter, O., Hohwy, J., van Boxtel, J., Lamme, V., Block, N., Koch, C., & Tsuchiya, N. (2018). Conscious machines: Defining questions, 400 400 *Science, 359*(6374). https://doi.org/10.1126/science.aar4163.

Cath, C. J. N., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2016). *Artificial Intelligence and the "Good Society": The US, EU, and UK Approach* (SSRN Scholarly Paper ID 2906249). *Social Science Research Network.* https://papers.ssrn.com/abstract=2906249.

Cavoukian, A. (2017). Global privacy and security, by design: Turning the 'privacy vs. security' paradigm on its head. *Health and Technology, 1–5*. https://doi.org/10.1007/s12553-017-0207-1.

CDEI. (2019). Interim report: Review into bias in algorithmic decision-making. Centre for Data Ethics and Innovation. https://www.gov.uk/government/publications/interim-reports-from-the-centre-for-data-ethics-and-innovation/interim-report-review-into-bias-in-algorithmic-decision-making.

Clarke, R. (2019a). Principles and Business Processes for Responsible AI. *Computer Law & Security Review, 35*(4), 410–422.

Clarke, R. (2019b). Regulatory Alternatives for AI. *Computer Law & Security Review, 35*(4), 398–409.

Clouser, K. D., & Gert, B. (1990). A Critique of Principlism. *Journal of Medicine and Philosophy, 15*(2), 219–236. https://doi.org/10.1093/jmp/15.2.219.

CNIL. (2015). Privacy Impact Assessment (PIA) Good Practice. CNIL. http://www.cnil.fr/fileadmin/documents/en/CNIL-PIA-3-GoodPractices.pdf.

CNIL. (2017a). How Can Humans Keep the Upper Hand? The Ethical Matters Raised by Algorithms and Artificial Intelligence.

CNIL. (2017b). Algorithms and artificial intelligence: CNIL's report on the ethical issues. (The Ethical Matters Raised by Algorithms and Artificial Intelligence). CNIL. https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf.

Committee on Bioethics (DH-BIO). (2019). Strategic Action Plan on Human Rights and Technologies in Biomedicine (2020-2025) (CM(2019)198). Council of Europe. https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680994df7.

Council of Europe. (2019). Unboxing artificial intelligence: 10 steps to protect human rights. https://www.coe.int/en/web/commissioner/view/-/asset_publisher/ugj3i6qSEkhZ/content/unboxing-artificial-intelligence-10-steps-to-protect-human-rights.

Decker, M. (2008). Caregiving robots and ethical reflection: The perspective of interdisciplinary technology assessment. *Ai & Society, 22*(3), 315–330.

Defense Innovation Board. (2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. US Department of Defense. https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*(6362), 486–492.

Doherty, N. F., Coombs, C. R., & Loan-Clarke, J. (2006). A re-conceptualization of the interpretive flexibility of information technologies: Redressing the balance between the social and the technical. *European Journal of Information Systems, 15*(6), 569–582.

Elsevier. (2018). ArtificiaI Intelligence: How knowledge is created, transferred, and used—Trends in China, Europe, and the United States. Elsevier. https://www.elsevier.com/?a=827872.

Commission, European (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* (COM(2018) 237 final). *European Commission.* http://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF.

General Data Protection Regulation, (2016) (testimony of European Parliament & European Council). http://www.dhealth.co.uk/resources/Documents/Event%20Content/CONSIL-ST_5419_2016_INIT-EN-TXT.pdf.

Finn, R. L., Wright, D., & Friedewald, M. (2013). Seven Types of Privacy. In S. Gutwirth, R. Leenes, P. de Hert, & Y. Poullet (Eds.), *European Data Protection: Coming of Age* (pp. 3–32). Netherlands: Springer. https://doi.org/10.1007/978-94-007-5170-5_1.

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology, 1*(1), 33–52.

Floridi, Luciano (Ed.). (2010). *The Cambridge Handbook of Information and Computer Ethics*. Cambridge University Press.

Garriga, E., & Melé, D. (2004). Corporate Social Responsibility Theories: Mapping the Territory. *Journal of Business Ethics, 53*(1–2), 51–71. https://doi.org/10.1023/B:BUSI.0000039399.90587.34.

Gilligan, C. (1990). *In a Different Voice: Psychological Theory and Women's Development (Reissue)*. Harvard University Press.

Grunwald, A. (2009). Technology Assessment: Concept and Methods. In D. M. Gabbay, A. W. M. Meijers, J. Woods, & P. Thagard (Eds.), *Philosophy of Technology and Engineering Sciences*. North Holland, 9 (pp. 1103–1146).

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review, 61*(4), 5–14.

Hagendorff, T. (2019). The Ethics of AI Ethics—An Evaluation of Guidelines. *ArXiv: 1903.03425 [Cs, Stat]*.

Hansen, M. (2016). Data Protection by Design and by Default à la European General Data Protection Regulation. In A. Lehmann, D. Whitehouse, S. Fischer-Hübner, L. Fritsch, & C. Raab (Eds.), *Privacy and Identity Management. Facing up to Next Steps* (pp. 27–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-55783-0_3.

Hatzakis, T., Rodrigues, R., & Wright, D. (2019). Smart Grids and Ethics. *ORBIT Journal, 2*(2), Article 2.

HLEG on AI, H. E. G. on A. I. (2019). Ethics Guidelines for Trustworthy AI. European Commission - Directorate-General for Communication. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Information Commissioner's Office. (2008). Privacy by design. http://www.ico.gov.uk/upload/documents/pdb_report_html/privacy_by_design_report_v2.pdf.

Iphofen, R., & Kritikos, M. (2019). Regulating artificial intelligence and robotics: Ethics by design in a digital society. *Contemporary Social Science*, 1–15.

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer, 51*(8), 56–59.

ISO. (2008). BS ISO/IEC 38500:2008—Corporate governance of information technology. https://bsol.bsigroup.com/Bibliographic/BibliographicInfoData/000000000030162049.

Jiya, T. (2019a). Ethical Implications of Predictive Risk Intelligence. *ORBIT Journal, 2*(2), Article 2.

Jiya, T. (2019b). Ethical Reflections of Human Brain Research and Smart Information Systems. *ORBIT Journal, 2*(2), Article 2.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Johnson, D. G. (2001). *Computer Ethics* ((3rd ed.).). Prentice Hall.

Johnson, K., Pasquale, F., & Chapman, J. (2019). Artificial Intelligence, Machine Learning, and Bias in Finance: Toward Responsible Innovation. *Fordham Law Review, 88*(2), 499.

Kancevičienė, N. (2019). Insurance, Smart Information Systems and Ethics. *ORBIT Journal, 2*(2), Article 2.

Kant, I. (1788). Kritik der praktischen Vernunft. Reclam, Ditzingen.

Kant, I. (1797). Grundlegung zur Metaphysik der Sitten. Reclam, Ditzingen.

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. *Business Horizons, 62*(1), 15–25.

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM, 53*(1), 148–152.

Koops, B.-J., Newell, B., Timan, T., Škorvánek, I., Chokrevski, T., & Galič, M. (2017). A Typology of Privacy. *University of Pennsylvania Journal of International Law, 38*(2), 483.

Kurzweil, R. (2006). The Singularity is Near. Gerald Duckworth & Co Ltd.

Latimer, J. (2017). Donna J Haraway, Manifestly Haraway: The Cyborg Manifesto, The Companion Species Manifesto, Companions in Conversation (with Cary Wolfe). *Theory, Culture & Society, 34*(7–8), 245–252.

Latonero, M. (2018). Governing artificial intelligence: Upholding human rights & dignity. Data&Society. https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

Liu, C. (2019). Multiple Social Credit Systems in China (SSRN Scholarly Paper ID 3423057). Social Science Research Network. https://papers.ssrn.com/abstract=3423057.

Livingstone, D. (2015). Transhumanism: The History of a Dangerous Idea. *CreateSpace Independent Publishing Platform.*

Locke, J. L. (2010). Eavesdropping: An Intimate History (Illustrated Edition). OUP Oxford.

Macnish, K. (2012). Unblinking eyes: The ethics of automating surveillance. *Ethics and Information Technology, 14*(2), 151–167. https://doi.org/10.1007/s10676-012-9291-0.

Macnish, K. (2020). Mass Surveillance: A Private Affair? *Moral Philosophy and Politics, 7*(1), 9–27. https://doi.org/10.1515/mopp-2019-0025.

Macnish, K., & Galliott, J. (2020). *Big Data and Democracy*. Edinburgh University Press.

Macnish, K., & Inguanzo, A. F. (2019). Customer Relation Management, Smart Information Systems and Ethics. *ORBIT Journal, 2*(2), Article 2.

Macnish, K., Inguanzo, A. F., & Kirichenko, A. (2019). Smart Information Systems in Cybersecurity. *ORBIT Journal, 2*(2), Article 2.

Macnish, K., & Ryan, M. (2019). SHERPA Deliverable 1.4 Report on Ethical Tensions and Social Impacts (Online Resource Project deliverable). SHERPA project. https://doi.org/10.21253/DMU.8181827.v2.

Macnish, K., Ryan, M., Gregory, A., Jiya, T., Antoniou, J., Hatzakis, T., Andreou, A., Rodrigues, R., Kirichenko, A., & Stahl, B. C. (2019). SHERPA Deliverable 1.1 Case studies (Online Resource Project deliverable). SHERPA project. https://doi.org/10.21253/DMU.8181827.v2.

Macnish, K., Ryan, M., & Stahl, B. (2019). Understanding Ethics and Human Rights in Smart Information Systems. *ORBIT Journal, 2*(2), Article 2.

Macnish, K., Wright, D., & Jiya, T. (2020). Predictive Policing in 2025: A Scenario. In H. Jahankhani, B. Akhgar, P. Cochrane, & M. Dastbaz (Eds.), *Policing in the Era of AI and Smart Societies* (pp. 199–215). Springer International Publishing. https://doi.org/10.1007/978-3-030-50613-1_9.

Martin, C. D., & Makoundou, T. T. (2017). Taking the high road ethics by design in AI. *ACM Inroads, 8*(4), 35–37.

Metcalf, J., Keller, E. F., & boyd, danah. (2016). Perspectives on Big Data, Ethics, and Society. Council for Big Data, Ethics, and Society. http://bdes.datasociety.net/wp-content/uploads/2016/05/Perspectives-on-Big-Data.pdf.

Mill, J. S. (1861). *Utilitarianism* ((2nd Revised edition).). Hackett Publishing Co Inc.

Mingers, J., & Walsham, G. (2010). Towards ethical information systems: The contribution of discourse ethics. *MIS Quarterly, 34*(4), 833–854.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 2019. https://doi.org/doi:10.1038/s42256-019-0114-4.

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

Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining Explanations in AI. FAT*'19, Atlanta, Georgia. https://www.academia.edu/37701175/Explaining_Explanations_in_AI.

Moor, J. H. (1985). What is computer ethics. *Metaphilosophy, 16*(4), 266–275.

Moor, J. H., & Bynum, T. W. (2002). Introduction to cyberphilosophy. *Metaphilosophy, 33*(1/2), 4–10.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). *From What to How- An Overview of AI Ethics Tools*. ArXive: Methods and Research to Translate Principles into Practices. https://arxiv.org/abs/1905.06876.

Nerurkar, M., Wadephul, C., & Wiegerling, K. (2016). Ethics of Big Data: Introduction. *International Review of Information. Ethics, 24*.

OECD. (2017). Recommendation of the OECD Council on Health Data Governance. http://www.oecd.org/health/health-systems/Recommendation-of-OECD-Council-on-Health-Data-Governance-Booklet.pdf.

Patel, A., Hatzakis, T., Macnish, K., Ryan, M., & Kirichenko, A. (2019). SHERPA Deliverable D1.3: Security Issues, Dangers and Implications of Smart Information Systems. SHERPA project. https://dmu.figshare.com/articles/D1_3_Cyberthreats_and_countermeasures/7951292.

Ramirez, R., Mukherjee, M., Vezzoli, S., & Kramer, A. M. (2015). Scenarios as a scholarly methodology to produce "interesting research". *Futures, 71*, 70–87. https://doi.org/10.1016/j.futures.2015.06.006.

Rehg, W. (2014). Discourse ethics for computer ethics: A heuristic for engaged dialogical reflection. *Ethics and Information Technology, 17*(1), 27–39. https://doi.org/10.1007/s10676-014-9359-0.

Roessler, B., & Mokrosinska, D. (Eds.). (2015). *Social Dimensions of Privacy: Interdisciplinary Perspectives*. Cambridge University Press.

Ryan, M. (2019a). Ethics of Public Use of AI and Big Data. *ORBIT Journal, 2*(2), Article 2.

Ryan, M. (2019b). Ethics of Using AI and Big Data in Agriculture: The Case of a Large Agriculture Multinational. *ORBIT Journal, 2*(2), Article 2.

Ryan, M., & Gregory, A. (2019). Ethics of Using Smart City AI and Big Data: The Case of Four Large European Cities. *ORBIT Journal, 2*(2), Article 2.

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society, ahead-of-print(ahead-of-print)*. https://doi.org/10.1108/JICES-12-2019-0138.

Santiago, N. (2020). SHERPA Delphi Study—Round 1 Results [Project Deliverable]. *SHERPA project*. https://www.project-sherpa.eu/wp-content/uploads/2020/03/sherpa-delphi-study-round-1-summary-17.03.2020.docx.pdf.

Sharkey, A., & Sharkey, N. (2010). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-010-9234-6.

Solove, D. J. (2002). Conceptualizing Privacy. *California Law Review, 90*(4), 1087–1156.

Sparrow, R. (2009). Predators or plowshares?: Arms control of robotic weapons. *IEEE Technology and Society Magazine, 28*(1). https://doi.org/10.1109/MTS.2009.931862.

Stahl, B. C., & Wright, D. (2018). Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security & Privacy, 16*(3), 26–33. https://doi.org/10.1109/MSP.2018.2701164.

Stahl, Bernd Carsten (2012). Morality, Ethics, and Reflection: A Categorization of Normative IS Research. *Journal of the Association for Information Systems, 13*(8), 636–656. https://doi.org/10.17705/1jais.00304.

Stahl, Bernd Carsten, & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*. https://doi.org/10.1016/j.robot.2016.08.018.

Tavani, H. (2008). Informational Privacy: Concepts, Theorie and Controversies. In J. V. D. Hoven, & J. Weckert (Eds.), *Information Technology and Moral Philosophy* (pp. 131–164). Cambridge University Press.

Taylor, L. (2016). The ethics of big data as a public good: Which public? Whose good? *Phil. Trans. R. Soc. A, 374*(2083), 20160126. https://doi.org/10.1098/rsta.2016.0126.

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont Report—Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Department of Health, Education, and Welfare. http://videocast.nih.gov/pdf/ohrp_belmont_report.pdf.

USACM. (2017). Statement on Algorithmic Transparency and Accountability. *ACM US Public Policy Council*.

van Est, R., & Gerritsen, J. (2017). Human rights in the robot age—Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality [Report to the Parliamentary Assembly of the Council of Europe (PACE)]. *Rathenau Instituut*. https://www.rathenau.nl/en/file/9605/download?token=OQgFllZS.

Varley-Winter, O., & Shah, H. (2016). The opportunities and ethics of big data: Practical priorities for a national Council of Data Ethics. *Phil. Trans. R. Soc. A, 374*(2083), 20160116. https://doi.org/10.1098/rsta.2016.0116.

Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *European Journal of Information Systems, 4*(2), 74–81. https://doi.org/10.1057/ejis.1995.9.

World Economic Forum. (2019). Responsible Use of Technology [White paper]. *WEB*. http://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology.pdf.

Wright, D. (2011). A framework for the ethical impact assessment of information technology. *Ethics and Information Technology, 13*(3), 199–226. https://doi.org/10.1007/s10676-010-9242-6.

Wright, D., Rodrigues, R., Hatzakis, T., Pannofino, C., Macnish, K., Ryan, M., & Antoniou, J. (2019). SHERPA Deliverable 1.2: SIS Scenarios. *De Montfort University*.

Wright, D., Stahl, B., & Hatzakis, T. (2020). Policy scenarios as an instrument for policymakers. *Technological Forecasting and Social Change, 154*, Article 119972. https://doi.org/10.1016/j.techfore.2020.119972.

Yin, R. K. (2003). *Case Study Research: Design and Methods (Third Edition)*. Sage Publications Inc.

Bernd Carsten Stahl is Professor of Critical Research in Technology and Director of the Centre for Computing and Social Responsibility at De Montfort University, Leicester, UK. His interests cover philosophical issues arising from the intersections of business, technology, and information. This includes ethical questions of current and emerging of ICTs, critical approaches to information systems and issues related to responsible research and innovation.

Andreas Andreou works as a Researcher and Project Manager at the human rights non governmental organisation Aequitas which is based in Cyprus. He holds a BA Humanities from the University of Essex and an LLM Master of Laws from the University of Central Lancashire. He has a research interest and experience in artificial intelligence and human rights, LGBTQI rights and hate speech/hate crime. He published academic articles and a research book.

Philip Brey (PhD, University of California, San Diego, 1995) is professor of philosophy of technology at the Department of Philosophy, University of Twente, the Netherlands. He is on the editorial board of eleven leading journals and book series in his field, including Ethics and Information Technology, Nanoethics, Philosophy and Technology, Techné, Studies in Ethics, Law and Technology and Theoria. He is also former president of the International Society for Ethics and Information Technology (INSEIT), and former president of the Society for Philosophy and Technology (SPT). His research focuses on ethics and philosophy of emerging technologies, in particular digital technologies, AI, robotics, biomedical technology and sustainable technology. He currently coordinates the EU H2020 SIENNA project on the ethical and human rights dimensions of emerging technologies, and the NWO-funded 10-year research programme Ethics of Socially Disruptive Technologies.

Tally Hatzakis is a senior research analyst at Trilateral Research with more than 15 years of research experience in academia as well as in the public and private sectors. Her domains of interest include smart cities, future tansport, big data and smart technologies, AI and robotics, personal wearables and privacy-by-design business models, as well as open innovation, green technologies and circular economy.

Alexey Kirichenko is research collaboration manager at F-Secure, coordinating F-Secure's participation in European and Finnish national research projects. He also represents the company in the Steering Board of WG6 (SRIA) of European Cyber Security Organisation (ECSO). His interests are mainly in applications of Machine Learning to cybersecurity.

Kevin Macnish is Assistant Professor in Ethics and Technology at the University of Twente. His interests include ethical questions regarding privacy, surveillance, security, and AI.

Andrew Patel is a researcher in F-Secure Corporation's Artificial Intelligence Center of Excellence. His main research areas include natural language processing, disinformation hunting and analysis, graph analysis and reinforcement learning. He is also a frequent contributor to F-Secure's blog.

Mark Ryan works as a Digital Ethics Researcher at Wageningen Economic Research, Wageningen University & Research. His interests are around topics in the ethics of technology, with a particular focus on ethical issues pertaining to the use of artificial intelligence and Big Data. He has published on topics, such as the ethics of smart cities, self-driving vehicles, agricultural data analytics, social robotics, and trusting AI.

David Wright is Director of Trilateral Research (London and Waterford), a company he founded in 2004. The company has partnered in more than 60 EU-funded projects. He has published four books on ambient intelligence, privacy and surveillance and more than 60 articles in peer-reviewed journals. He coined the term ethical impact assessment and published the first article on EIA. The ISO privacy impact assessment standard is based on the methodology he developed as is the CEN Workshop Agreement on ethical impact assessment. His interests include scenario construction, horizon scanning, impact assessments, cybersecurity, artificial intelligence, privacy, ethics and surveillance.