Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

# Exploring ethics and human rights in artificial intelligence – A Delphi study

Bernd Carsten Stahl [a,d,*], Laurence Brooks [b,d], Tally Hatzakis [c], Nicole Santiago [c], David Wright [c]

[a] School of Computer Science, University of Nottingham, UK
[b] Information School, University of Sheffield, Sheffield, UK
[c] Trilateral Research, London, UK
[d] Centre for Computing and Social Responsibility, De Montfort University, Leicester, UK

ARTICLE INFO

*Keywords:*
Delphi
Ethics
Human rights
Artificial intelligence (AI)
Expert stakeholders
Policy issues

ABSTRACT

Ethical and human rights issues of artificial intelligence (AI) are a prominent topic of research and innovation policy as well as societal and scientific debate. It is broadly recognised that AI-related technologies have properties that can give rise to ethical and human rights concerns, such as privacy, bias and discrimination, safety and security, economic distribution, political participation or the changing nature of warfare. Numerous ways of addressing these issues have been suggested. In light of the complexity of this discussion, we undertook a Delphi study with experts in the field to determine the most pressing issues and prioritise appropriate mitigation strategies. The results of the study demonstrate the difficulty of defining clear priorities. Our findings suggest that the debate around ethics and human rights of AI would benefit from being reframed and more strongly emphasising the systems nature of AI ecosystems.

## 1. Introduction

Ethical and human rights concerns arising from technologies related to artificial intelligence (AI) remain a prominent topic of public and scientific debate (Eliot, 2022; Hallamaa and Kalliokoski, 2022).The significant economic and broader societal benefits that AI promises are counterbalanced by worries that these technologies can disadvantage some individuals and communities and have detrimental effects on people's rights and legitimate expectations (Mantelero and Esposito, 2021). Worries range from data protection, information security and algorithmic biases to unemployment, the exacerbation of economic inequality and manipulation of democratic processes. Scientific researchers look for technical solutions, professional associations provide guidance, standards bodies promote good practice and policymakers seek appropriate regulatory mechanisms (European Commission, 2021). All of this happens under close scrutiny of the media and the public at large.

Key challenges of the AI ethics discourse are its scope and complexity (Russell and Norvig, 2016). AI is not an easily defined term and both the use of the term and the underlying technologies are mushrooming (Murdick et al., 2020). The potential impact of AI is ubiquitous (Makridakis, 2017). The number and variety of affected stakeholder groups grows quickly, and most individuals are, at least potentially, subject to the consequences of AI use. The scope of the discussion is thus almost universal. At the same time, the sheer number of technologies, applications, issues and possible mitigations is too large to allow for simple classifications. This is exacerbated by the idea that there are multiple interactions and feedback loops between all of these aspects, rendering the overall topic area difficult to understand and even more difficult to govern.

In light of these challenges, this paper seeks to answer the following research question: *"What are the most pressing ethical and human rights issues of AI and which mitigation measures should be prioritised to address them?"* This is not a straightforward and scientific research question to which one could expect to find a simple answer, but a question that calls for complex and multi-dimensional judgements on the basis of detailed understanding of the subject matter. The methodology chosen to answer it is well-suited to this type of complex future-oriented question as we will show in the methodology section. We undertook a three-stage Delphi study with the aim of finding an expert agreement on these questions.

The most striking finding arising from this research is that the Delphi study did not lead to a clear expert consensus. This was despite the rich structure of the debate on ethics and human rights of AI across the various stages of the study and the detailed understanding on both the issues and possible mitigation strategies of the expert panel. While there

[*] Corresponding author at: School of Computer Science, University of Nottingham, UK.
*E-mail address:* bernd.stahl@nottingham.ac.uk (B.C. Stahl).

https://doi.org/10.1016/j.techfore.2023.122502
Received 18 September 2021; Received in revised form 10 March 2023; Accepted 11 March 2023
Available online 23 March 2023

was some agreement on key ethical and human rights concerns, there was little agreement on the ways in which these should be addressed. Even some of the most prominently discussed mitigation measures, notably regulatory interventions and legislation, did not figure among the most promising solutions. The most highly rated proposals included education, investigative journalism and exchange of good practice which are characterised by their broad nature and lack of specificity to AI.

The paper makes several important contributions to knowledge. The findings are of importance for the discussion of ethics and AI in that they suggest that the current framing of AI ethics may need to be reconsidered. The difficulty of the Delphi panel to converge on a set of clear priorities highlights a fundamental ambiguity around AI technologies and thus disparity about the framing of relevant ethical and human issues and mitigations. Based on our study, we suggest that the AI ethics discourse needs to employ a different conceptual basis. This is an important insight for scholars in the disciplines contributing to research and innovation policy and those working on ethics and human rights in AI, but it is equally important for policymakers and practitioners working on AI regulation and implementation. Failure to rethink the view of ethics, human rights and AI is likely to lead to a scattergun approach, unlikely to resolve the underlying issues it is meant to address.

In order to develop the argument, the paper proceeds as follows. It begins with a review of the discourse on ethical and human rights issues of AI, covering conceptual bases, issues and key mitigation measures. This is followed by a description of the methodology of the study. We then present and discuss findings from the Delphi study. The conclusion spells out key insights and points to further work to be undertaken.

## 2. Ethics and human rights in AI

The Delphi study and the responses we received need to be interpreted in the context of the broader discourse around AI. We therefore briefly introduce the concept of AI, ethical and human rights issues currently discussed and mitigation measures that have been traditionally proposed to address these issues.

### 2.1. Artificial intelligence

AI is not easily defined. While there is an abundance of proposed definitions, none fully captures all aspects of interest. The definition that coined the term as part of a proposal for the 1955 *Dartmouth summer research project on artificial intelligence* (McCarthy et al., 2006) suggested that every aspect of human intelligence could be simulated by a machine. This raises questions about what counts as intelligence. Possible answers include cognitive functions such as perceiving, reasoning, learning, problem solving, decision-making and creativity (Rai et al., 2019). This raises questions whether and to which degree any of these are special to human beings, whether AI has to replicate human intelligence, other animal intelligence or can surpass either (Brooks, 2002).

Many definitions refer not so much to the underlying technology but to the scientific discipline that aims to achieve it. Stone et al. (2016, p. 13), for example, describe AI as "a branch of computer science that studies the properties of intelligence by synthesizing intelligence". Bundage et al. (2018, p. 9) focus on the technical artefacts and suggest that "AI refers to the use of digital technology to create systems that are capable of performing tasks commonly thought to require intelligence." This type of definition has proven to be appealing to policymakers, probably because of its simplicity and can be found in policy-related documents, such as the European Commission's (2020a, p. 2) White Paper on AI that states that "Simply put, AI is a collection of technologies that combine data, algorithms and computing power."

The problem with this range of definitions is that they do not offer a unified view of AI that would allow for an unambiguous identification of the ethical and human rights issues raised by it. Kaplan and Haenlein (Kaplan and Haenlein, 2019) liken the quest for a definition of AI to the attempt to define beauty. The quest for a unified definition may be misleading because there is not one converging phenomenon that can be called AI. This suspicion is supported by a comprehensive analysis of the literature on AI undertaken by the publisher Elsevier (2018) that found a number of clusters of discourses in the AI literature but no universally agreed core.

This uncertainty of the term AI poses a problem for the academic discourse as well as practical and policy interventions. It constitutes an obstacle for a detailed understanding of which issues are associated with AI as well as ways of addressing them, which will be discussed in the following sections.

### 2.2. Ethical and human rights issues of AI

For the purpose of this paper, we are primarily interested in the non-technical issues that arise from AI. These are often described as ethical issues, which raises the question of the definition of ethics. Ethics is a philosophical discipline that explores what counts as right or wrong, good or bad and on what grounds such judgements are made. Prominent ethical theories point to the importance of consequences of actions in judging their ethical quality (Bentham, 1789; Mill, 1861), others underline the importance of duty and logic (Kant, 1797, 1788), whereas others focus on personal character and virtue (Aristotle, 2007). Ethical discourses have been applied to many areas, creating sub-disciplines of applied ethics, for example, in biomedical ethics (Beauchamp and Childress, 2009) or neuroethics (Farah, 2005). Technical developments have led to bodies of applied ethical work. Of relevance to AI are topics such as the ethics of technology (Royakkers and Poel, 2011), computer ethics (Bynum, 2008; Johnson, 2001) and information ethics (Floridi, 2010, 1999).

Ethical theories have an important role in helping us understand why something is considered ethically problematic and how we can find ways to address the issue. In the current discourse on ethics and AI, however, there is relatively little emphasis on foundational ethical theories. The ethical positions that are used in AI ethics tend to be based on principles, which are mid-level applicable concepts that are meant to guide action. This principle-based approach has been developed and broadly accepted by the biomedical ethics community which has been influential in shaping AI ethics. While there is a large and fast-growing body of work promoting AI ethics principles, several comparative studies have shown that the number of these principles is relatively limited (Jobin et al., 2019; Ryan and Stahl, 2020). Analysis has furthermore shown that AI ethics principles are often directly linked to established human rights (Fjeld et al., 2020), which is why we cover both ethics and human rights issues of AI simultaneously.

This paper is less concerned with a philosophical ethical analysis of issues and more with an understanding of what constitutes such issues with a view to finding solutions. We therefore adopt a pluralist ethical perspective (Ess, 2006) which means that we accept that something is an ethical or human rights issue if a source (e.g., the literature, our Delphi respondents) describes it as an issue. This position is consistent with current approaches to research and guide research and innovation policy and practice, such as technology assessment (Coenen and Simakova, 2013; Roessner and Frey, 1974; Schot and Rip, 1996) or responsible innovation (Owen et al., 2021; Pandza and Ellwood, 2013; Stilgoe et al., 2013; Wiarda et al., 2021) which promote reflection on normative issues without positing a particular ethical standard.

The number of ethical and human rights issues discussed in the literature is substantial. We offer a brief overview to provide context for the Delphi study described below, but we do not claim that this overview is complete or exhaustive. A good starting point for such an overview of ethical issues of AI is to highlight that AI not only raises concerns but also promises ethical benefits. The most widely cited advantage of AI is its ability to optimise processes, improve use of data and thereby help organisations improve their bottom line. The European Union Agency for Fundamental Rights stated, based on significant empirical research stated that “The single most important reason for using AI is increased efficiency” (FRA, 2020). While this is not an ethical benefit per se, it has been pointed out that trustworthy AI “can improve individual flourishing and collective wellbeing by generating prosperity, value creation and wealth maximization” (AI HLEG, 2019, p. 3). In addition, AI in various applications is expected to help humans avoid hazardous or heavy and unpleasant work (Muller, 2020), improve scientific research with benefits in health (Haque et al., 2020; Topol, 2019), improve logistics and transportation and many more. From a human rights perspective, some AI applications may help guarantee rights (e.g., improving access to healthcare for visually impaired, optimising agricultural practices to improve access to food as part of the right to an adequate standard of living) (Access Now, 2018). Overall, AI can contribute to the achievement of ethical aims such as those represented in the UN's sustainable development goals (SDGs) or various forms of grand challenges (Murray et al., 2012) and be a “force for good” (Taddeo and Floridi, 2018).

However, these benefits are counterbalanced by numerous actual harms and potential risks, many of which are ethically problematic. Some of these issues arise from or are related to the properties of particular AI techniques. At the core of the current AI debate are advances of machine learning (Babuta et al., 2020). Machine learning is not a new idea, but it has only become a high-profile success relatively recently due to the availability of computing power and large datasets required for training and validating models using neural networks (Dignum, 2019). There are different types of machine learning (Boden, 2018). Two main categories of machine learning models refer to their transparency: transparent (also called interpretable) and opaque (also called black-box). Transparent models include linear regression, decision trees, and k-nearest neighbours. They provide insights into how they make predictions. Opaque models include random forests and gradient boosting machines. These are more difficult to interpret, but they can achieve higher accuracy on complex tasks. Opacity may contribute to concerns about data protection, especially in those cases where the AI requires large data sets. Where those data sets include personal data, AI can raise questions around privacy and data protection (Buttarelli, 2018). The opacity of the system leads to concerns about hidden biases (CDEI, 2019) and resulting unfair discrimination (Access Now Policy Team, 2018; Latonero, 2018). Systems furthermore need to be reliable, safe and secure (Brundage et al., 2018) which can be difficult to ascertain.

In addition to these concerns that arise from the nature of AI technologies, there are numerous worries about their role in larger sociotechnical systems and the impact this can have on individual and collective lives. Systems containing AI have a certain level of autonomy in the sense that they can act on their environment without immediate human intervention. Such automated decision-making may have advantages over human decision-making (e.g., if an autonomous vehicle has a better safety record than (some) human drivers), but it raises concerns when automatic decisions affect humans (Council of Europe, 2019). In addition, this raises the issue of responsibility and accountability, for example, is the AI responsible for an action or is the person who designed and/or built the AI the responsible party? AI systems are expected to have significant economic impacts, raising questions about unemployment, worker surveillance and the justice of economic distribution (Foster-McGregor et al., 2021; Zuboff, 2019). AI systems can affect political processes, lead to power concentration (Nemitz, 2018) and damage democracy. AI can have a problematic impact on the environment (European Commission, 2020b), can change the nature of warfare (Defense Innovation Board, 2019) and more broadly structure the scope for human action in undesirable ways (Coeckelbergh, 2019).

In addition to these current concerns, there are worries about future developments that might lead to artificial general intelligence that has truly human capabilities. While such technology does not exist at present and it is unclear whether current technologies can facilitate such capabilities, the worries about them exist. This raises questions about what the social and ethical consequences of such “super intelligent” machines might be (Müller, 2020). In addition to questions whether such machines could be subjects of responsibility or rights, it raises the broader question of the future relationship between humans and machines (Vallor, 2016).

### 2.3. *Mitigation measures*

The preceding section has demonstrated the complexity of the ethical issues that AI is expected to raise. This complexity is matched by the complexity of suggestions on how to address these issues. It is important to quickly review the main themes of these proposed mitigation strategies to appreciate the value and contribution of our Delphi study.

There are at least three ways of organising the discourse on mitigation strategies.

- The first way is to look at them by organisational level. This can start at the national, regional and international policy levels, and include strategies at the corporate, community and individual level.
- A second way is to look at what type of organisation the mitigation strategy targets (e.g., government v. industry; public v. private).
- A third way to organise mitigation strategies is by type of strategy (e.g., legislation, policy, standards, codes of conduct, guiding principles).

At just the policy level, there are numerous initiatives where attempts to address ethical issues often form part of broader policy activities aimed at promoting AI research, development and uptake (e.g. OECD, 2019). AI and the ethical issues it raises touch on many policy areas such as research and innovation, data protection, taxation, competition and intellectual property (Borrás and Edler, 2020). There are, thus, numerous policy and legislative proposals that are discussed in parliaments, governments or other policymaking bodies that may influence how ethical issues of AI can be realised (Rodrigues et al., 2020). The EU, for example, has proposed a Regulation for AI (European Commission, 2021, 2020a). In addition, there are suggestions for the creation of institutions such as regulators to accompany and implement regulations (Miller and Ohrvik-Stott, 2018) as well as proposals for international coordination of these policy-oriented activities, for example, on the level of the G20 (Jelinek et al., 2020).

Much of AI development, deployment and use are facilitated by companies and other types of organisations. Industry is clearly aware of these challenges and the creation of industry bodies such as the Partnership on AI demonstrate an intention to contribute to and shape the discourse. On the organisational governance level, there are numerous activities and processes that organisations typically already have in place that can help them deal with AI-related issues. These include risk management (Clarke, 2019), data governance (British Academy and Royal Society, 2017), data protection processes (EDPS, 2020) and various types of impact assessment. Such activities can be integrated into strategic initiatives of organisations that aim to position them in an ethically desirable way, such as corporate social responsibility agendas (Garriga and Melé, 2004) and stakeholder engagement. Organisations taking these ideas seriously can make use of processes aimed at integrating attention to human rights into organisational structures (BSR, 2018; Council of Europe, 2019).

For individuals, whether they work for governments, organisations or in a personal capacity, there are numerous mechanisms that provide guidance on how to deal with ethical issues of AI. These start with ethical frameworks and principles such as the one produced by the EU High Level Expert Group (HLEG) on AI (2019). There is a growing number of standards that provide guidance, led by the IEEE (2017a), in particular their family of P7000 family of standards (IEEE, 2017b; Peters et al., 2020). Established principles of professional ethics in computing are available and in some cases being updated to reflect the need of recent technological developments such as AI (Brinkman et al., 2017). The Special Committee 42 Working Group 3 of the International Organization for Standardization has developed a set of ethical and societal principles for AI in its draft ISO 24368 document, which are similar to those of the HLEG.

There is also a growing number of technical tools and approaches that aim to address specific problems, such as security vulnerabilities (Brundage et al., 2018) or transparency and explainability. There is a growing number of development methodologies that aim to help developers integrate ethical concerns into the early stages of AI development, such as ethics pen testing (Berendt, 2019), the VCIO (Values, Criteria, Indicators, Observables) model (AIEI Group, 2020) or the ART (accountability, responsibility, transparency) principles (Dignum, 2019). Many of these models are based on the principles of value-sensitive design (Friedman et al., 2008; Winkler and Spiekermann, 2018). In addition, there are numerous tools meant to support individuals in understanding and dealing with these issues (Morley et al., 2019).

The broad range of mitigation options and strategies illustrates that there is no consensus on the prioritisation of approaches to the governance of AI. The earlier overview of the concept of AI, ethical issues caused by AI and ways in which these may be addressed point to a key challenge of the ethics and AI debate: the problem of prioritisation. The broad range of definitions and meanings of the term AI, the large number and complexity of the ethical and human rights issues in combination with rapidly changing landscape of possible mitigation strategies make it difficult to determine who should undertake which actions in which order. The review of the literature presented here mirrors what we believe to be the general structure of the discourse. This can be summarised as follows: There are clearly identifiable technologies that constitute AI. These technologies have characteristics that either on their own or in particular application context raise ethical and human rights concerns. Such concerns can be addressed using well-defined mitigation mechanisms and governance structures. This summary is of course overly simplified and the discourse recognises that simple linear relationships between AI, ethics and solutions are the exception rather than the rule. The overview of the AI ethics discourse nevertheless demonstrates the relevance of our research question which we formulate as follows: What are the most pressing ethical and human rights issues of AI and which mitigation measures should be prioritised in order to address them?

To answer this question, we chose a methodology suitable for finding responses to complex future-oriented questions as described in the next section.

## 3. Methodology: Delphi study design

A Delphi study (Dalkey et al., 1969; Linstone and Turoff, 2011, 2002) is typically described as a future-oriented methodology, one example of future and foresight research (Martin, 2010; Sardar, 2010) which includes numerous other methodologies, such as scenario development, horizon scanning, citizen panels and simulations (Georghiou et al., 2008; Rowe et al., 2017; Wright et al., 2020). Delphi studies have been identified as a particularly useful tool to support policy development (Adler and Ziglio, 1996; Rowe and Wright, 2011).

According to Ziglio (1996), there are three key considerations for the application of Delphi studies to a policy problem:

1. "the problem does not lend itself to precise analytical techniques but can benefit from subjective judgements on a collective basis […];
2. the problem at hand has no monitored history nor adequate information on its present and future development […];
3. addressing the problem requires the exploration and assessment of numerous issues connected with various policy options where the need for pooled judgement can be facilitated by judgmental techniques".

These considerations relate well to the questions of ethics and human rights in AI, which call for collective judgements, rather than research questions that require exact scientific findings. The methodology is thus suitable to provide a response to our research question. A Delphi study is not expected to create new knowledge in a traditional scientific sense, but rather to make best use of existing knowledge and the collective wisdom of the participants (Sandrey and Bulger, 2008). Delphi studies typically involve a number of experts, who are unaware of who else is participating in the study, to avoid undue influence and biases (Paré et al., 2013). They can collect qualitative and quantitative data (Tapio et al., 2011). Participant responses are anonymised and are communicated so that individuals are freed from concerns about repercussions for their attitudes and convictions. Consensus, or at least a clarification of the existing positions, can be reached over time as opinions are swayed. However, what is seen as 'consensus' remains open to interpretation and does vary across studies (Powell, 2003). The Delphi study presented here consisted of three rounds of online surveys, starting with an open qualitative round that was used to identify options that were then narrowed down and quantified in the two subsequent rounds. The following sections describe the main steps of the study and other key considerations.

### 3.1. *Preparation and pilot testing and ethical approval (July – September 2019)*

This phase consisted of developing, pilot testing and agreeing a Delphi study protocol internally to ensure that questions are suitable and understandable to external participants. In accordance with standard practice of social science, each of the Delphi rounds was pilot tested. This means that the Delphi survey instrument was checked for comprehensibility and usability. This was done by first circulating the survey in the consortium and asking for feedback. Following this, the survey was tested by selected experts. The project supporting the Delphi study focused on AI and big data analytics using the term "smart information systems" (Stahl and Markus, 2021). For the purposes of this paper, we continue to use the term AI, which is currently more prominent in the literature.

The study was awarded ethical approval in October 2019, by the Faculty Research Ethics Committee in the Faculty of Computing, Engineering and Media (CEM), De Montfort University, UK. A condition of ethics approval was that no personal data was to be collected and responses were to be analysed anonymously, an aspect that had consequences for the study analysis to which we will return in the discussion section.

### 3.2. *Delphi Round 1 (October 2019 – January 2020)*

The first round (R1) asked respondents to brainstorm on the issues raised by AI and ways to address these issues. The survey consisted of a set of five open questions asking respondents (I) to list the three most important ethical or human rights issues, (ii) to name current approaches, methods or tools for addressing these issues, (iii) and the advantages and disadvantages of these responses, (iv) their proposals for addressing such issues better and (v) their top three criteria to select and prioritise the most appropriate measures.

The responses to R1 were first analysed by one of the authors and the results reviewed by members of the project team, including the remaining authors. Coding terms were developed after reviewing like words and phrases in the collected responses and grouping like responses. The results were synthesised into a 14-page summary and a link to this summary was shared with respondents as part of the invitation to round 2.

### *3.3. Delphi Round 2 (March – August 2020)*

The purpose of the second round (R2) of the Delphi study was to narrow down the issues and approaches identified in the R1 brainstorming. R2 consisted of four sets of questions, asking participants to rate (on a scale of 1–5) issues and potential measures across three criteria. Question 1 asked respondents to rate a list of ethical and human rights issues in terms of reach, significance and attention. Questions 2, 3 and 4 asked respondents to rate potential mitigation measures in terms of desirability, feasibility and probability. These measures were divided into regulatory measures (question 2), technical measures (question 3) and other measures (question 4). The list of issues and approaches was based on the responses in R1, supplemented with additional issues and measures identified in parallel research activities (Stahl et al., 2022a; Stahl et al., 2022b; Stahl et al., 2021). The full list of questions and the options they were meant to evaluate is provided in appendix A.

The responses to R2 were also first analysed by one of the authors and reviewed by members of the project team. The results were synthesised into a summary that was shared with respondents as part of the invitation to round 3.

### *3.4. Delphi Round 3 (September – October 2020)*

The final round of the Delphi study (R3) was designed to determine consensus on the prioritisation of potential governance measures. Respondents were asked to select the three most important potential governance measures for immediate action, from the list of 15 highest scoring measures in R2. For each selection, respondents were prompted to explain (a) why the measure is important, (b) how the measure should be implemented and by whom, and (c) what indicators would show the successful implementation of the measure. Respondents were also given the option to identify any potential governance measures that should not be prioritised, as well as to provide any additional comments.

The responses to R3 were first analysed by one of the authors and reviewed by members of the project team.

### *3.5. Selection of participants*

A Delphi study is an expert-based method that requires input from qualified experts with a detailed understanding of the subject matter. Delphi studies explicitly do not aim to draw a representative sample to represent a population. The selection of participants therefore focused on their expertise. In preparation for this study, we identified 1200 experts on various aspects of AI ethics and human rights. Members of the project consortium identified these experts from their own knowledge, web searches and using a snowball system. From this long list, we selected 231 candidates who were approached with the request to participate in the Delphi study. While the Delphi study does not claim statistical representativeness, the problem of bias in AI is widely discussed. In order to counter this, we aimed to ensure that there was a spread of geographic diversity and representation of different stakeholder groups (e.g., policymakers, technologists, businesspeople, academics, civil society organisations). We aimed for a 50 % representation of non-male participants.

### *3.6. Delivery and responses*

In all three rounds, we sent out an initial invitation to the selected experts and several reminders. While the number of individuals who clicked on the initial invitation was around 50 %, the number who engaged with the survey to the point that they provided sufficient information to warrant analysis was much lower and is shown in the following table (Table 1).

**Table 1**
Survey delivery dates and number of responses.

| | Survey open | Survey close | Usable responses |
|---|---|---|---|
| Round 1 | 24.10.2019 | 15.01.2020 | 41 |
| Round 2 | 18.03.2020 | 30.06.2020 | 26 |
| Round 3 | 18.09.2020 | 07.10.2020 | 43 |

### *3.7. Statistical significance of findings*

This Delphi study includes some quantitative data, notably the responses to the 5-point Likert scale type questions in round 2. We calculated the confidence intervals for the responses to all questions in round 2 of the survey. These were between 0.14 and 0.2 at a significance level of 0.05 or 5 %. As the responses we received ranged from 2.3 to 4.7, we are confident that the findings are relevant and statistically meaningful. However, in many cases, the differences between individual items are below 0.15, rendering statements about exact ranking statistically unsupported. We therefore present groups of items in the following section and report on trends.

## 4. Findings

This section provides a brief overview of key findings from the Delphi study. Across the three stages, the study produced large amounts of data, not all of which can be presented here. Instead, we structure the presentation of the findings according to the question: What are the most pressing ethical and human rights issues of AI and which mitigation measures should be prioritised in order to address them? This calls for a description of these issues, an identification of the mitigation measures and an evaluation of their relevance. For clarity, we refer to clusters of responses by noting the question to which they responded, e.g., R1-Q2 would stand for responses given in round 1 of the Delphi study to question 2.

### *4.1. Ethical and human rights issues of AI*

The content analysis of the qualitative responses to R1-Q1 (see Fig. 1) shows that there is a significant overlap between responses and that these reflect the broader literature on the topic.

When asked to rank these issues according to their reach, i.e., the number of people who are likely to be affected, their significance, i.e., their overall societal importance, and attention, i.e., the likelihood of stimulating public debate, a more nuanced picture emerged. The top-ranking issue in terms of reach was the misuse of data, the highest score in terms of significance was bias and discrimination whereas the highest level of attention was attributed to the disappearance of jobs. The following radar diagram (Fig. 2) shows the average scores that the issues received in R2-Q1.

This diagram shows that in most cases significance is assessed as higher than reach and in most cases, these are clearly higher than attention. This suggests that, overall, our expert respondents felt that the issues were receiving less attention than they deserve, with the notable exceptions of disappearance of jobs and “awakening” of AI, two topics that are prominently discussed in the literature.

### *4.2. Mitigation measures and their evaluation*

We categorised the responses to R1-Q2 (current mitigation strategies) into three main groups: regulatory, technical and other. We defined ‘Regulatory measures’ as those requiring action by a governmental entity for implementation; these measures could be legally binding as a matter of law. ‘Technical measures’ as those developed and implemented by technical industry actors. And ‘Other measures’ was a catch-all category for non-technical measures developed by industry, civil society organisations, academia, independent expert bodies and

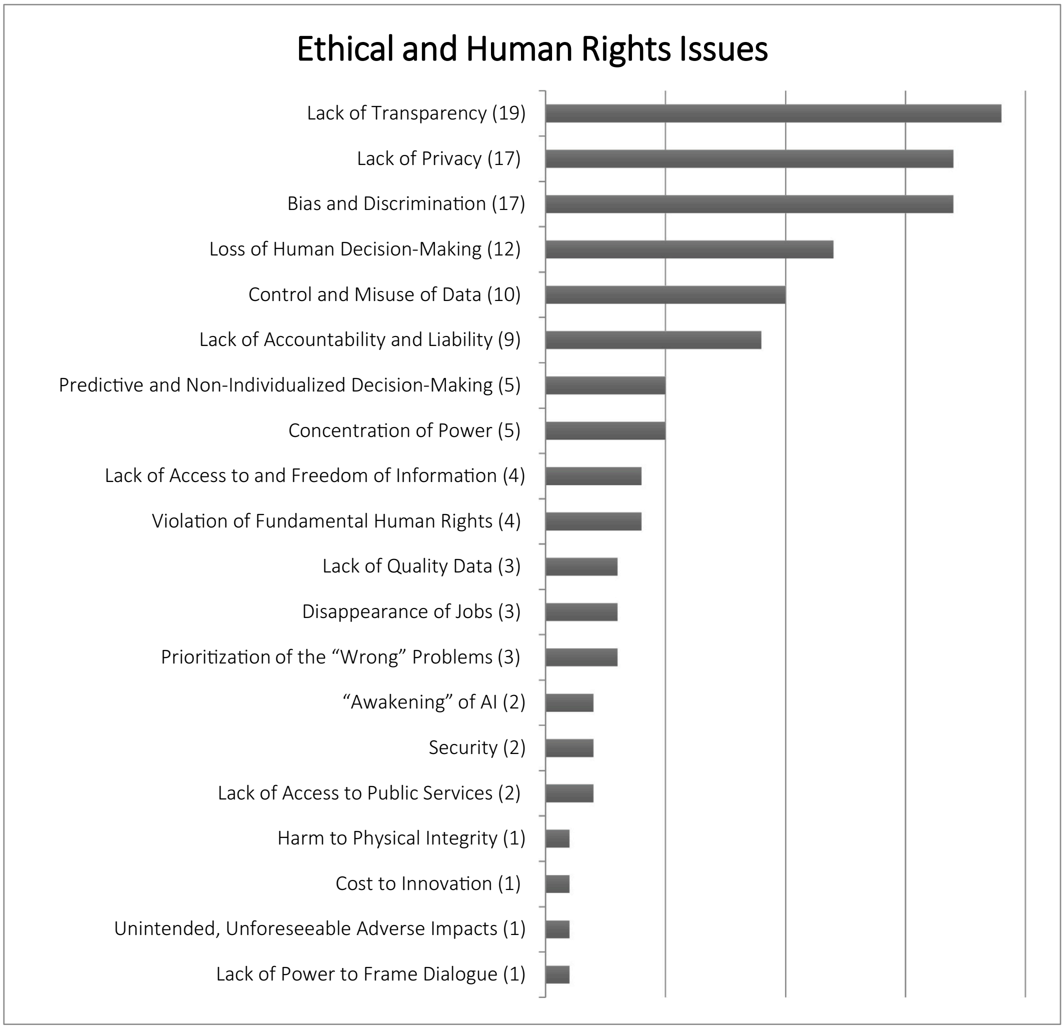


**Fig. 1.** Overview of ethical issues in response to R1-Q1.

individuals. We developed and used this categorisation because it reflects key aspects of current debate and we felt that the content of the regulatory and technical measures could be identified easily. The "other" category thus contained all other options that fit neither of these categories. The overview of the findings is displayed in Table 2. The number in parentheses represents the number of times that a particular item was named.

A later question (R1-Q4) asked respondents which measures they would propose. The response to this question differed in the frequency in which some of the items were mentioned but was remarkably consistent with the list of mitigation strategies currently used as represented in Table 2.

During the analysis of round 1, we identified 52 mitigation options (see Appendix, Round 2) across the three main categories, which respondents then evaluated in terms of their desirability, feasibility and probability in round 2. When looking at the average of the scores across these three measures, we can identify the top and bottom 15 options as shown in the following figure.

A more detailed analysis using the individual scores of desirability, feasibility and probability shows that these properties were not judged to be consistent. The overall average score across all options for desirability was 3.99, for feasibility 3.62 and for probability was 3.21. The difference between these scores differed greatly for various options. Desirability was regarded as higher (more important) than feasibility for options that are technically challenging or require a high degree of international collaboration, such as the creation of tools capable of identifying synthetically created content (e.g., deepfakes), the establishment of rules on how decisions in systems that have the capability to cause physical harm should be made in difficult situations (e.g., self-driving vehicles) or the enforcement of human rights laws. On the other end of the scale, probability was unsurprisingly seen as higher (more important) than desirability for those options that already exist and are highly visible, such as high-level expert groups or ethical codes of conduct. A similar relationship could be observed between desirability and probability. It is not surprising that feasibility and probability are related. An option that is not perceived to be feasible is not likely to be judged as probable to be implemented.

However, there are still different evaluations of feasibility and desirability of various options. Feasibility was generally perceived to be higher than probability. In some cases, this was strikingly so, notably for the two options of "retaining 'unsmart' products and services by keeping them available to purchase and use" and "Grievance mechanisms for complaints on SIS [smart information systems]".

When asked in R1-Q3 "What do you think are the pros and cons of these current approaches, methods or tools?", the respondents provided the following answers (Table 3).

These responses demonstrate that the strengths and weaknesses are perceived to differ between different mitigation options. Looking at the strengths and weaknesses in isolation is unlikely to allow for the prioritisation of options.

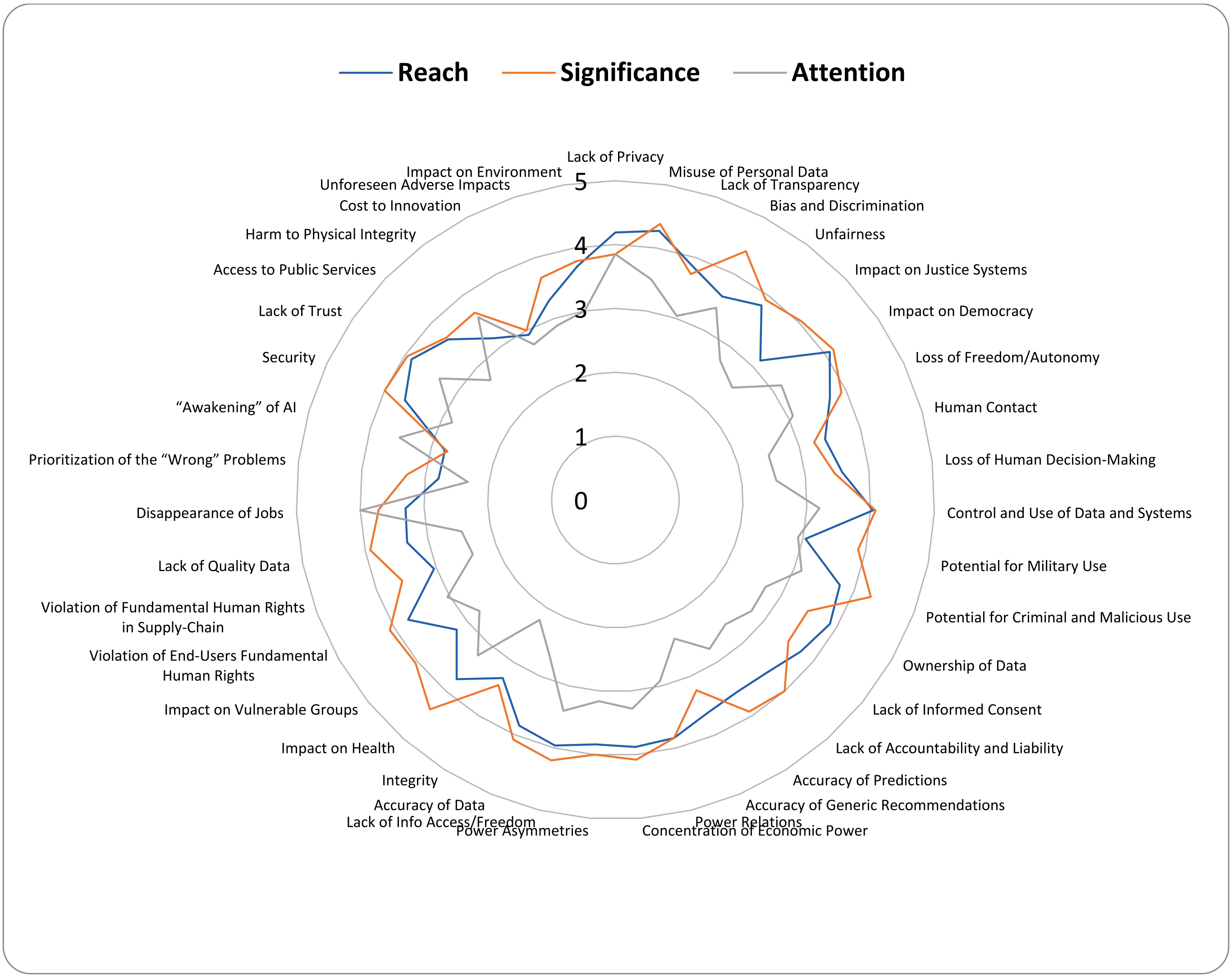


**Fig. 2.** Ranking of ethical issues, showing average scores per issue on a five-point Likert scale.

### 4.3. *Priorities and implementation*

The overall purpose of the Delphi study was to develop practical and applicable insights into ways of effectively addressing ethical and human rights issues of AI and big data. The third aspect of the Delphi study to be presented here refers to the question of how one could prioritise between the different mitigation measures.

It is therefore important to think about criteria according to which different measures and options can be prioritised. Respondents were asked to suggest up to three criteria for prioritisation (Fig. 4). The analysis of the answers to this question (R1-Q5) provided the following answers (Fig. 1).

Respondents applied these criteria to the mitigation strategies during R2 (see Fig. 3) selected up to three of these options to prioritise in R3. Table 4 provides an overview of the number of times the various options were chosen in R3.

Respondents' use of the ability to provide in-depth comments on the options varied. In some cases, for example, open letters and NGO coalitions, no qualitative data was provided. For most of the more popular choices, respondents did provide their rationale for choosing them, including details on implementation and success measures. However, it is important to note that even among the more popular choices, no consensus on these aspects emerged. With regard to the methodologies for systematic and comprehensive testing, respondents suggested that these should be implemented by "global experts", "Europe", "industries" and "AI coalition". Measures of success suggested for this option included "number of applications", "reports" and "standardisation".

## 5. Discussion

An initial observation concerning the findings of the Delphi study is that, contrary to our expectations, it did not lead to the formation of a consensus or at least of some areas of general agreement among the experts. Our hope that this expert-oriented approach to the ethics of AI would help overcome the complexity of the broader discussion and highlight clearly identifiable ways to address the issues was not realised.

The study did show, maybe not surprisingly, that the experts were aware of the issues discussed in the literature and confirmed them as relevant. The findings suggest that there may be some issues that receive more attention and media coverage than the level of concern by the experts would warrant, notably the loss of employment and the possibility of "awakening of AI". These are extensively discussed topics. Employment is a key concern of policy-related work on AI and has figured strongly in the debate ever since the twin report to the US president (Executive Office of the President, 2016a,b). There is an ongoing and open debate about the net effect of AI on employment (Kaplan and Haenlein, 2019; Willcocks, 2020). Awakening of AI is a topic that has interested certain parts of the academic debate and is

**Table 2**
Current measures used to address ethical issues.

| Current measures | |
|---|---|
| Regulatory measures | • Regulations (18)<br>• Public register of permissions to use data (1)<br>• Reporting Guidelines (1)<br>• Monitoring Mechanism (2) |
| Technical measures | • Testing Algorithms on Diverse Subsets (1)<br>• Using Analytics Systems to Judge Whether Decisions Are Equal/Fair (1)<br>• Generative Adversarial Networks and Other Techniques for Deriving Explanations from Outcomes (1)<br>• More Open Data (2) |
| Other measures | • Codes of Conduct (3)<br>• Education Campaigns (4)<br>• Employing 'Fairness' Officer or Ethics Board (3)<br>• Frameworks, Guidelines, and Toolkits (14)<br>• Grievance Mechanism (1)<br>• High-Level Expert Groups (6)<br>• Individual Action (2)<br>• International Framework (3)<br>• Investigative Journalism (3)<br>• NGO Coalitions (1)<br>• Open Letters (1)<br>• Public Policy Commitment (1)<br>• Self-Regulation (1)<br>• Stakeholder Dialogue and Scrutiny (3)<br>• Standardisation (3)<br>• Third-Party Testing and External Audits (2) |

**Table 3**
Overview of pros and cons of current mitigation measures.

| Pros |
|---|
| • Dialogue means we **learn from each other** |
| • Regulation has **power of enforcement** |
| • Transparency measures means **building ethics into the design** |
| • Education **enhances citizen/consumer power** |
| • Ethical Impact Assessments provide **clear methodology & tools** |
| • Standardisation has **objective set of criteria** |
| • Oversight **addresses human rights violations** |

| Cons |
|---|
| • **Lack of understanding** about roles & responsibilities |
| • **Risk of shifting burden of responsibility** to developers or consumers |
| • Measures are **too abstract** |
| • Creation & implementation is **resource intensive** |
| • Non-binding measures have **no enforcement** |
| • **No comprehensive approach** |
| • **Too complicated** to implement new ways of thinking |
| • Regulation has **limited application** |
| • Technology **development outpaces rule-making process** |
| • Measures **perceived as a hurdle** |
| • Measures are **public-sector focused** |
| • **Difficult to measure ethics objectively** |
| • Educational campaigns ineffective because **don't reach people who need it most** |

linked with concepts such as the singularity (the point where AI is expected to be able to improve itself, leading to an exponential growth of its capabilities (Kurzweil, 2006; Tipler, 2012)) and the future role of humans, as for example discussed by proponents of transhumanism (Livingstone, 2015). These are highly contested notions that attract attention and are frequently the subject of science fiction but whose current real-world relevance is subject to debate.

The identification of ethical issues and mitigation options thus did not provide many surprising insights. However, the way in which the respondents ranked them clearly defied our expectations. Given that the respondents were selected on the basis of their expertise, we expected them to prioritise specific and well-articulated mitigation options that would clearly address particular issues. Surprisingly, this did not turn out to be the case. The top options were predominantly broad and geared towards the creation of knowledge and awareness. Few were specific to AI. Most surprising in light of current discussions concerning the need for legislation to better govern AI in many national contexts and at the level of the European Union, not a single legislative option made it to the top 15. In fact, many of them ended up in the bottom 15 of the score (see Fig. 3).

Many of the more specific mitigation options that were identified are in the category of technical options. These scored at an intermediate level in round 2, when respondents were asked to rank all options. However, those that made it into the top 15 and therefore were included in round 3 of the study then went on to be selected most frequently, suggesting that such specific and targeted interventions are seen as useful when looked at in detail.

Further insights concerning respondents' reasoning can be gleaned from the way they differentiated between desirability, feasibility and probability. They mostly scored the desirability of options higher than their feasibility or their probability. This suggests that respondents believed that interventions are needed but their optimism about the possibility of achieving them and finding ways of implementing them was more limited.

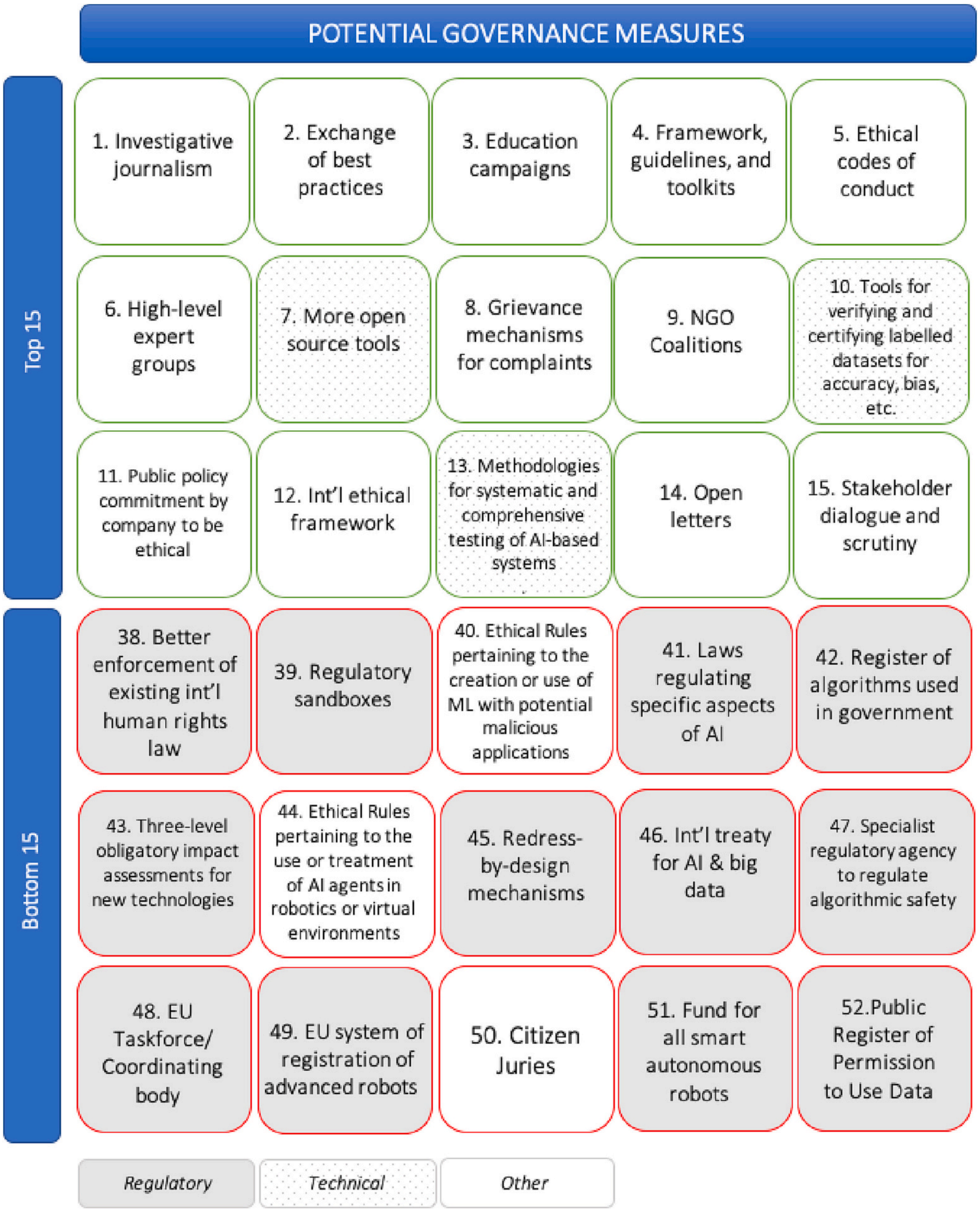
One reason for this evaluation may be that the exact pathways to the implementation of mitigation options are unclear. This may be caused or at least exacerbated by the complexity and multiplicity of stakeholders. The responses indicated that even in case of clearly defined options, our respondents did not agree on whose responsibility it was to further develop, promote and implement them. This is exacerbated by the lack of agreement on how successful implementation could be measured or ascertained. This lack of certainty about individual stakeholders and stakeholder groups may go some way to explaining why many of the top-ranked options are broad, such as education campaigns or exchanges of best practice.

One reasonable explanation for the unexpected outcomes of the study is that it is based on a set of assumptions that did not work out in practice. The logic of the Delphi study and of most of the ethics and AI debate was detailed earlier in the paper. It assumes that there is a recognisable set of technologies that falls under the heading of AI. These technologies have features that give rise to or exacerbate ethical and human rights concerns. These concerns can be clearly identified. Once identified, they are subject to resolution. Mitigation options addressing these issues can be developed, applied appropriately at the relevant stage of the technology lifecycle by clearly identifiable stakeholders. The result of this mitigation is that the issue is resolved or at least that responsibilities and liabilities are clearly assigned.

The findings of our Delphi study do not support this logic. While the respondents identified a plethora of ethical and human rights issues and mitigation options, they did not converge on well-defined ways of addressing ethical issues that can be implemented by specific stakeholders. There are several possible explanations for this.

The first explanation is that the concept of AI is too broad to provide a useful starting point for this discussion. In the literature review earlier in this paper, we highlighted the fundamentally different categories of technologies and socio-technical systems that fall under the heading of AI. Depending on the concept of AI that one uses, the resulting issues vary. Some of these issues may be subject to resolution, but many of them are either part of broader societal issues (e.g., inequality, justice, distribution) or touch on fundamental philosophical questions (nature of humanity, role of human beings in the world) that may not be subject to resolution at all.

The second explanation for the failure of the Delphi study to result in unambiguous outcomes is likely to be the systemic nature of AI and its ethical consequences. The issues that the respondents highlighted are not independent but have many mutual interdependencies. For example, privacy and data protection are not just a problem in their own right, but the protection of personal data in autonomous systems can have consequences for the discrimination of individuals, for the ability to profit



POTENTIAL GOVERNANCE MEASURES

Top 15

1. Investigative journalism

2. Exchange of best practices

3. Education campaigns

4. Framework, guidelines, and toolkits

5. Ethical codes of conduct

6. High-level expert groups

7. More open source tools

8. Grievance mechanisms for complaints

9. NGO Coalitions

10. Tools for verifying and certifying labelled datasets for accuracy, bias, etc.

11. Public policy commitment by company to be ethical

12. Int'l ethical framework

13. Methodologies for systematic and comprehensive testing of AI-based systems

14. Open letters

15. Stakeholder dialogue and scrutiny

Bottom 15

38. Better enforcement of existing int'l human rights law

39. Regulatory sandboxes

40. Ethical Rules pertaining to the creation or use of ML with potential malicious applications

41. Laws regulating specific aspects of AI

42. Register of algorithms used in government

43. Three-level obligatory impact assessments for new technologies

44. Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments

45. Redress-by-design mechanisms

46. Int'l treaty for AI & big data

47. Specialist regulatory agency to regulate algorithmic safety

48. EU Taskforce/ Coordinating body

49. EU system of registration of advanced robots

50. Citizen Juries

51. Fund for all smart autonomous robots

52.Public Register of Permission to Use Data

Regulatory

Technical

Other

**Fig. 3.** Top fifteen and bottom fifteen potential governance measures (from complete list of 52 potential measures) based on average overall scores (average of desirability, feasibility, and probability scores).

from such systems and thus for the distribution of risks and benefits. Similarly, the mitigation strategies are often closely interrelated and interdependent. A legislative approach may make use of standardisation for purposes of definition of liability, it can mandate good practice defined by professional bodies. Specific tools, e.g., for the assessment of bias in data, can form part of risk assessment. A further aspect of the systemic nature of the problem space can be found when looking at relevant stakeholders and stakeholder groups. An individual computer scientist may work as a developer of a system in a company. She may simultaneously contribute to a standardisation body and give expert feedback on legislative proposals. The company that employs her can be part of an industry association that develops good practice guidance and promotes it when lobbying policymakers. The individual may furthermore be part of a civil society organisation that argues for stronger privacy protection.

Adopting such a systems-oriented view would require different ways of thinking about AI and approaches for addressing relevant issues. Instead of asking which ethical or human rights issues can be addressed in which way, the question would need to be how to define and delineate a relevant AI system or sub-system and how such a system could be shaped to allow it to be sensitive to ethical issues and find ways of dealing with them. The idea of adopting a systems-oriented approach to AI has already gained significant traction, most notably through the application of the concept of an ecosystem to AI (Digital Catapult, 2020; European Commission, 2020a; OECD, 2019). While the basic idea of thinking about AI using a systems perspective is thus already established, it has yet to be spelled out in more detail how this can inform the way in which ethical and human rights issues are identified, interpreted and dealt with (Stahl, 2021).

## 6. Conclusion

In this paper, we presented a Delphi study that explored an expert view of ethical and human rights consequences of the development,

**Table 4**
Number of times each of the top 15 options were selected.

| Option | Times chosen |
|---|---|
| Methodologies for systematic and comprehensive testing of AI-based systems | 16 |
| Framework, guidelines, and toolkits for project management and development | 13 |
| Stakeholder dialogue and scrutiny | 10 |
| International ethical framework | 9 |
| Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties | 9 |
| Exchange of best practices | 8 |
| More open-source tools that allow for transparency, explainability, and bias mitigation | 8 |
| Public policy commitment by company to be ethical | 8 |
| Ethical codes of conduct | 7 |
| Education campaigns | 6 |
| Grievance mechanisms for complaints on SIS | 5 |
| High-level expert groups | 5 |
| Investigative journalism about issues concerning SIS | 5 |
| Open letters to governments and the public | 4 |
| NGO coalitions on particular issues | 4 |
| TOTAL | 117 |

deployment and use of AI. The three-round Delphi study asked respondents to identify ethical and human rights issues, name mitigation measures, rank these and prioritise the most important mitigation measures.

The study confirmed the issues and measures that can be found in the literature. While confirming some aspects of the current debate, it also offered some surprising and unexpected insights. The first aspect that we did not expect to find was a general lack of emphasis on legislation and formal regulation. This is surprising because at present there are numerous initiatives to develop AI-related legislation that are debated in various parliaments, but none of them was seen as a priority by our expert respondents. There was no strong favourite(s) among the mitigation measures when experts rated these, but the ones coming out on top were general measures that aim at raising awareness, promoting education and formulating general principles.

In light of the lack of agreement among experts, it is not surprising that there is no general agreement in the population about these questions. This raises the question of how policymakers, but also organisations and individuals, should engage with AI. One important insight from our study is that diverse routes of investigation of AI technologies are required to develop our understanding of AI. This growing understanding should be made available beyond expert audiences and inform formal and informal education and societal discourse.

Maybe even more important is that our Delphi study suggests that

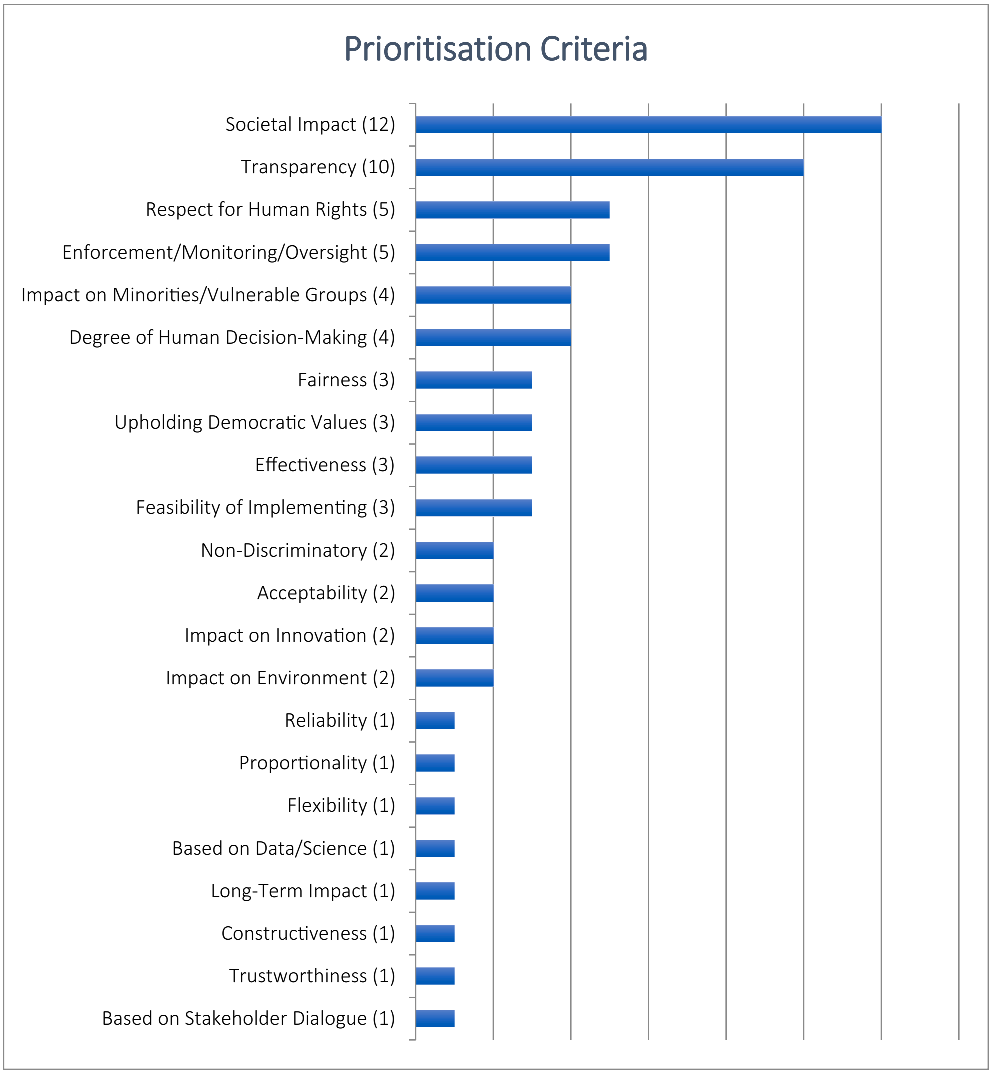


**Fig. 4.** Frequency analysis of response to R1-Q5.

the framing of the AI ethics discourse that informed the design of the study but also much of the policy and other activities may be inappropriate. In the discussion, we suggested that the concept of AI is too broad to be helpful and that this may explain some of the lack of convergence of the expert opinions. In addition, we suggested that the focus on individual technologies, mitigation options and stakeholder groups may need to be replaced by a broader and more encompassing view that takes a systems-level perspective. When considering such an alternative approach, the role of individual mitigation strategies would need to be reconsidered. Legislation, for example, may then have a crucial role not in addressing particular issues, but in shaping innovation ecosystems in ways that render them sensitive to appropriate interventions and capable of acting accordingly.

Our Delphi study, like any piece of research, has limitations. The ethics approval process required us to collect only anonymous data. It was therefore impossible for us to track the responses of individual respondents across the three rounds of the study, which could have provided more insights into possible links between perceptions of ethical issues and preferred mitigation strategies and their link to the demographics of the respondents. In addition, it became clear that the Likert-style questions that were asked in round 2 were difficult to answer and required respondents to spend a significant amount of time, thus limiting our response rate. Furthermore, there are fundamental limitations of Delphi studies, which are expert-oriented and do not allow drawing conclusions about wider populations.

Further research would be desirable in several directions. More detailed analyses of specific stakeholder groups or experts in specific areas would shed light on the question whether there are particular issues or mitigation measures of relevance to such groups. There are interesting questions around geographical spread of experts and concerns. The majority of the experts we invited to participate were of European origin or affiliation, which gives our findings a European flavour, which may not be replicated elsewhere. In light of our assumption that the concept of AI is a key problem, it would be important to follow up our study with similar studies using a more refined and specific concept, such as machine learning or artificial neural networks, or that would focus on particular application areas, such as transport, finance or medicine. Finally, if we are right in suggesting that a systems-oriented view might provide a better framing of the debate, it would be highly insightful to undertake a similar study based on a systems logic.

While there are many ways in which this study can be taken further, we believe that it makes important contributions in its own right. The confirmation that the ethical and human rights issues are comprehensively mapped is important from a theoretical as well as practical perspective. The failure of the Delphi study to clearly converge on a particular set of issues and mitigations provides important insights for scholarly debate, but it also indicates to policymakers and others who work on practical responses to these questions that a pluralistic and open approach is important that allows a dynamic development of the discourse (Kuhlmann et al., 2019). Maybe most important is our deduction of the limitation of the current framing of the debate. Taking a more holistic and systems-oriented perspective can stimulate scientific debate. It should encourage scholars working on AI ethics and human rights who come from different disciplinary backgrounds including research policy, technology ethics, science and technology studies, technology law and many others, to think beyond their disciplinary boundaries and familiar theories and incorporate thoughts from fields such as information systems, general systems theory or innovation studies. This will contribute to the scientific debate and the understanding of the relevant phenomena. Simultaneously, such a broader perspective may help practitioners and policymakers to shape AI systems in ways that are conducive to human flourishing.

## CRediT authorship contribution statement

Bernd Carsten Stahl: Funding acquisition, project administration, writing original draft, formal analysis, supervision.

Laurence Brooks: Writing, review and editing.

Tally Hatzakis: Writing, review and editing.

Nicole Santiago: data collection, data analysis, writing, review and editing.

David Wright: Writing, review and editing, supervision.

## Data availability

Data will be made available on request.

## Acknowledgements

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) and 945539 (Human Brain Project SGA3).This work was supported by the Engineering and Physical Sciences Research Council [Horizon Digital Economy Research 'Trusted Data Driven Products: EP/T022493/1]. This paper acknowledges the support that the authors received from collaborators, in particular the consortium members of the SHERPA project.

## Appendix A. Delphi questions

*Round 1*

1. What do you think are the three most important ethical or human rights issues raised by AI and/or big data?
2. Which current approaches methods, or tools for addressing these issues are you aware of? These may be organisational, regulatory, technical or other.
3. What do you think are the pros and cons of these current approaches, methods, or tools?
4. What would you propose to address such issues better?
5. Which should be the top 3 criteria for society to select and prioritise the most appropriate measures?

*Round 2*

Question 1:

The following ethical and human rights issues and possible mitigation measures (question 2,3,4) were taken from the DELPHI Round 1 responses. The issues were supplemented with issues identified in other activities of the SHERPA project, including analysis of case studies, stakeholder interviews, and an online survey.

Please rate the ethical and human rights issues in terms of:

- **Reach** (number of people affected)
- **Significance** (impact on individuals)
- **Attention** (likely to lead to public debate)

Issues should be rated on a score of 1 to 5. A low score (1) means the issue affects few (or no) individuals, is trivial, / or is not of serious concern. A high score (5) means the issue affects individuals worldwide, has vital consequences, / or is likely to generate robust public debate. In the last column, please provide a brief explanation of why you hold this opinion.

**Lack of Privacy**
Related to which type of data and how much data is collected, where from, and how it is used

**Misuse of Personal Data**
Related to concerns over how SIS might use personal data (e.g. commercialization, mass surveillance)

**Lack of Transparency**
Related to the public's need to know, understand, and inspect the mechanisms through which SIS make decisions and how those decisions affect individuals

**Bias and Discrimination**
Related primarily to how sample sets are collected/chosen/involved in generating data and how data features are produced for AI models; and how decisions are made (e.g. resource distribution) according to the guidance arising out of the data

**Unfairness**
Related to how data is collected and manipulated (i.e. how it is used), also who has access to the data and what they might do with it as well as how resources (e.g. Energy) might be distributed according to the guidance arising out of the data

**Impact on Justice Systems**
Related to use of SIS within judicial systems (e.g. AI used to ‘inform’ judicial reviews in areas such as probation)

**Impact on Democracy**
Related to the degree to which all involved feel they have an equal say in the outcomes, compared with the SIS

**Loss of Freedom and Individual Autonomy**
Related to how SIS affects how people perceive they are in control of decisions, how they analyse the world, how they make decisions (e.g. impact of manipulative power of algorithms to nudge towards preferred behaviours), how they interact with one another, and how they modify their perception of themselves and their social and political environment

**Human Contact**
Related to the potential for SIS to reduce the contact between people, as they take on more of the functions within a society

**Loss of Human Decision-Making**
Related to how SIS affects how people analyse the world, make decisions, interact with one another, and modify their perception of themselves and their social and political environment

**Control and Use of Data and Systems** Related to how data is used and commercialised, including malicious use (e.g. mass surveillance); how data is collected, owned, stored, and destroyed; and how consent is given

**Potential for Military Use**
Related to the use of SIS in future possible military scenarios (e.g. autonomous weapons), including the potential for dual-use applications (military and non-military)

**Potential for Criminal and Malicious Use**
Related to the use of SIS in criminal and malicious scenarios (e.g. cyber-attacks and cyber espionage)

**Ownership of Data**
Related to who owns data, and how transparent that is (e.g. when you give details to an organisation, who then ‘owns’ the data, you or that organisation?)

**Lack of Informed Consent**
Related to informed consent being difficult to uphold in SIS when the value and consequences of the information that is collected is not immediately known by users and other stakeholders, thus lowering the possibility of upfront notice

**Lack of Accountability and Liability**
Related to the rights and legal responsibilities (e.g. duty of care) for all actors (including SIS) from planning to implementation of SIS, including responsibility to identify errors or unexpected results

**Accuracy of Predictive Recommendations**
Related to the possibility of misinterpreting data, implementing biases, and diminishing the accuracy of SIS recommendations when SIS interprets an individual's personal data

**Accuracy of Non-Individualized Recommendations**
Related to the possibility of misinterpreting data, implementing biases, and diminishing the accuracy of SIS recommendations when SIS makes a decision based on data not specific to an individual

**Power Relations**
Related to the ability of individuals to frame and partake in dialogue about issues; and the fact that few powerful corporations develop technologies, influence political processes, and have know-how to ‘act above the law’

**Concentration of Economic Power**
Related to growing economic wealth of companies controlling SIS (e.g. big technology companies) and individuals, and unequal distribution of resources

**Power Asymmetries**
Related to the ability of individuals to frame and partake in dialogue about issues; and the fact that few powerful corporations develop technologies, influence political processes, and have know-how to ‘act above the law’

**Lack of Access to and Freedom of Information**
Related to quality and trustworthiness of information available to the public (e.g. fake news, deepfakes) and the way information is disseminated and accessed

**Accuracy of Data**
Related to using misrepresentative data or misrepresenting information (i.e. predictions are only as good as the

(*continued on next page*)

(*continued*)

underlying data) and how that affects end user views on what decisions are made (i.e. whether they trust the SIS and outcomes arising from it)

**Integrity**
Related to the internal integrity of the data used as well as the integrity of how the data is used by a SIS

**Impact on Health**
Related to the use of SIS to monitor an individual's health and how much control one can have over that

**Impact on Vulnerable Groups**
Related to how SIS creates or reinforces inequality and discrimination (e.g. impacting on the dignity and care for older people, for example how much a care robot might exert over an older person's life and 'tell them what to do')

**Violation of End-Users Fundamental Human Rights**
Related to how human rights are impacted for end-users (e.g. monitoring and control of health data impacting right to health; manipulative power of algorithms nudging towards some preferred behaviours, impacting rights to dignity and freedom)

**Violation of Fundamental Human Rights in Supply-Chain**
Related to how human rights are impacted for those further down the supply-chain extracting resources and manufacturing devices (e.g. impacts on health, labour violations, lack of free, prior and informed consent for extractives)

**Lack of Quality Data**
Related to using misrepresentative data or misrepresenting information in building AI models

**Disappearance of Jobs**
Related to concerns that use of SIS will lead to significant drop in the need to employ people

**Prioritisation of the "Wrong" Problems**
Related to the problems SIS is developed to 'solve' and who determines what the immediate problems are

**"Awakening" of AI**
Related to concerns about singularity, machine consciousness, super-intelligence etc. and the future relationship of humanity vis-a-vis technology

**Security**
Related to the vulnerabilities of SIS and their ability to function correctly under attacks or timely notify human operators about the need of response and recovery operations

**Lack of Trust**
Related to using misrepresentative data or misrepresenting information (i.e. predictions are only as good as the underlying data) and how that affects end user views on what decisions are made (i.e. whether they trust the SIS and outcomes arising from it); also related to informed consent and that helps with trust

**Access to Public Services**
Related to how SIS could change the delivery and accessibility of public services for all (e.g. through privatisation of services)

**Harm to Physical Integrity**
Related to the potential impacts on our physical bodies (e.g. from self-driving cars, autonomous weapons)

**Cost to Innovation**
Related to balancing the protection of rights and future technological innovation

**Unintended, Unforeseeable Adverse Impacts**
Related to future challenges and impacts that are yet known

**Impact on Environment**
Related to concern about the environmental consequences of infrastructures and devices needed to run SIS (e.g. demand for physical resources and energy)

**Question 2/3/4**: The following potential regulatory measures originated from the DELPHI Round 1 responses. The examples given were refined and supplemented by analysis conducted in other deliverables of the SHERPA project, including a report on regulatory options.

Please rate the following potential regulatory measures in terms of:

- **Desirability** (would you like to have this measure in place?)
- **Feasibility** (in theory, is it possible to have this measure in place?)
- **Probability** (in reality, is it likely that this measure would be put in place?)

Issues should be rated on a score of 1 to 5. A low score (1) means the measure will have a major negative effect, is very challenging to create, and/or is impossible to achieve. A high score (5) means the measure will have a very positive effect, is not difficult to create, and/or is very likely to happen.

| Potential regulatory measures (question 2) |
|---|
| **Creation of new international treaty for AI and Big Data**<br>(open for adoption by all countries) |
| **Better enforcement of existing international human rights law** |
| **Binding Framework Convention to ensure that AI is designed, developed and applied in line with European standards on human rights, democracy and the rule of law** (Council of Europe) including through a new ad hoc committee on AI (CAHAI) |
| **CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment** |
| **Legislative framework for independent and effective oversight over the human rights compliance of the development, deployment and use of AI systems by public authorities and private entities** (Council of Europe) |
| **General fund for all smart autonomous robots or individual fund for each and every robot category** (EU Parliament) |
| **Establishment of a comprehensive Union system of registration of advanced robots** within the Union's internal market where relevant and necessary for specific categories of robots and establishment of criteria for the classification of robots |

(*continued on next page*)

(*continued*)

| Potential regulatory measures (question 2) |
|---|
| **Algorithmic impact assessments** under the General Data Protection Regulation (GDPR) |
| Creation of new body: **EU Taskforce/Coordinating body of field-specific regulators for AI/big data** |
| **Redress-by-design mechanisms for AI** (High-Level Expert Group on Artificial Intelligence (AI HLEG)) |
| **New laws regulating specific aspects**, e.g., deepfakes, algorithmic accountability. |
| **Register of algorithms used in government** |
| **New national independent cross-sector advisory body** (e.g. UK Centre for Data Ethics and Innovation) |
| **New specialist regulatory agency to regulate algorithmic safety** |
| **Public Register of Permission to Use Data** (individuals provide affirmative permission in a public register for companies to use their data) |
| **Reporting Guidelines** (for publicly registered or traded companies based on corporate social responsibility reporting as described by GRI) |
| **Regulatory sandboxes for AI and big data** |
| **Three-level obligatory impact assessments for new technologies** |
| Potential technical measures (question 3) |
| **Methodologies for systematic and comprehensive testing of AI-based systems** (including fairness of decisions) |
| **Techniques for providing explanations for output of AI models** (e.g., Layerwise relevance propagation for neural networks) |
| **Easily understandable description of the model's inputs** (including input validity checks), training data, requirements, and potential limitations for services based on machine learning models |
| **AI-as-a-service, providing in-built mechanisms to mitigate against common adversarial attacks** (e.g. functionality to allow a model's owner to easily determine whether training data can be reverse-engineered from the model) |
| **Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties** |
| **Tools for verifying and certifying publicly available services based on machine learning models** |
| **Reputation information about publicly available services based on machine learning models** (e.g. including a black list of known faulty, vulnerable, inaccurate, etc. services and models) |
| **Tools capable of identifying synthetically created or manipulated content**, such as images, videos, speech, and written content (available and easy-to-use for the general public) |
| Other potential measures (question 4) |
| **Certification** (e.g. initiative for IEEE Ethics Certification Program for Autonomous and Intelligent Systems) |
| **Citizen Juries** to evaluate risk of various AI technologies and propose appropriate tools |
| **Education Campaigns** (e.g. Finnish Element of AI course; Dutch Nationale AI Cursus) |
| **Ethical Codes of Conduct** (e.g. EU High Level Expert Group Guidelines for Trustworthy AI, SHERPA guidelines) |
| **Ethical Mindset** adopted by companies |
| **Ethical Rules pertaining to the creation or use of machine learning models with potential malicious applications**, covering preventive and reactive cases (e.g. rules governing recommendation systems: how they should work, what they should not be used for, how they should be properly hardened against attacks, etc.) |
| **Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments** (e.g., AI robots resembling dogs, sex robots) |
| **Exchange of Best Practices** |
| **'Fairness' Officer or Ethics Board** employed within companies using/developing SIS |
| **Framework, Guidelines, and Toolkits** for project management and development (e.g. UK Data Ethics Framework; IBM AI Fairness 360 Open Source Toolkit; Dutch Data Ethics Decision Aid (DEDA) tool) |
| **Grievance Mechanisms** for complaints on SIS |
| **High-level Expert Groups** (e.g. UN AI for Good Global Summit) |
| **Individual Action** (e.g. participating in conferences to raise awareness; protecting oneself by refusing cookies online) |
| **International Ethical Framework** (e.g. OECD Principles on AI) |
| **Investigative Journalism** about issues concerning SIS |
| **More Open Source Tools** that allow for transparency, explainability, and bias mitigation |
| **NGO Coalitions** on particular issues (e.g. Campaign to Stop Killer Robots) |
| **Open Letters** to governments and the public (e.g. 2015 Open Letter on AI) |
| **Public Policy Commitment** by company to be ethical |
| **Public "Whistleblowing" Mechanisms** for the reporting of bias, inaccuracies, or ethical impacts of systems based on machine learning models |
| **Retaining 'Unsmart' Products and Services** by keeping them available to purchase and use |
| **Rules on how decisions in systems that have the capability to cause physical harm should be made in difficult situations** (e.g. self-driving vehicles and other systems) |
| **Self-Regulation by Company** (e.g. Twitter's self-imposed ban on political ads) |
| **Stakeholder Dialogue and Scrutiny** with scientists, programmers, developers, decision makers, politicians, and the public at large |
| **Standardisation** (e.g. IEEE P7000 series of standards for addressing ethical concerns during system design). |
| **Third-party Testing and External Audits** (e.g. of data used for training for quality, bias, and transparency) |

## *Top and bottom measures by category*

### *Regulatory measures*

In R1, regulation was the most frequently cited example of a possible 'approach, method, or tool' to address the ethical and human rights concerns associated with SIS. However, in R2, most potential regulatory measures scored low, both in absolute terms and relative to other types of potential measures. No regulatory measure was in the top fifteen measures, and twelve regulatory measures were in the bottom fifteen potential measures. This was because potential regulatory measures received the lowest average scores in all three criteria and overall. For the overall scores, no regulatory measures scored in the very high (4.5–5) or high (4–4.49) range. All of the top five regulatory measures (below) scored in the mid-high (3.5–3.99) range, which was lower than the top scoring technical and other measures. More significantly, potential regulatory measures had the highest percentage of measures scoring in the mid-low (3–3.49) to low (2–2.99) range for all three criteria. This was particularly true of probability, where 95 % of measures scored low. Within potential regulatory measures, the majority (16 of 18) were rated more desirable than feasible or probable.

**Table 3**
Top five and bottom five scoring potential regulatory measures based on overall scores.

| Top five regulatory measures | Bottom five regulatory measures |
|---|---|
| • Legislative framework for independent and effective oversight of human rights compliance (3.70)<br>• Algorithmic impact assessments (3.65)<br>• National independent cross-sector advisory body (3.59)<br>• Binding Framework Convention (3.51)<br>• Reporting Guidelines (3.50) | • Specialist regulatory agency to regulate algorithmic safety (3.07)<br>• EU Taskforce/Coordinating (3.06)<br>• EU system of registration of advanced robots (2.85)<br>• Funds for all smart autonomous robots (2.75)<br>• Public Register of Permission to Use Data (2.71) |

*Technical measures*

In R1, technical measures were rarely mentioned. However, in R2, technical measures scored relatively high, particularly in regard to desirability; all technical measures were very high (4.5–5) or high (4–4.49) for desirability. However, with lower average scores in feasibility and probability, only three technical measures were in the top fifteen measures. For the overall scores, all technical measures scored in the mid-high range (3.5–3.99). All potential technical measures were rated more desirable, then feasible, then probable.

**Table 4**
Top three and bottom three scoring potential technical measures based on overall scores.

| Top three technical measures | Bottom three technical measures |
|---|---|
| • Tools for verifying & certifying labelled datasets for accuracy, bias & other important properties (3.95)<br>• Methodologies for systematic & comprehensive testing of AI-based systems (3.90)<br>• Techniques for providing explanations for output of AI models (3.87) | • Reputation information about publicly available services based on machine learning models (3.63)<br>• Tools capable of identifying synthetically created or manipulated content (3.58)<br>• AI-as-a-service (3.52) |

As there were only eight potential technical measures, only the top and bottom three were highlighted.

*Other measures*

In R1, respondents cited a broad range of other measures. In R2, these other potential measures scored high, both in absolute terms and relative to the other two categories of measures. Twelve of the top fifteen measures were other measures. This was because other measures received the highest average scores in feasibility, probability, and overall. For the overall scores, seven measures were in the very high (4.5–5) range, sixteen in the high (4–4.49) range, and two in the mid-high (3.5–3.99) range. The only measure to score in the mid-low (3–3.49) range overall was citizen juries. The majority of measures (23 of 26) scored more desirable than feasible or probable.

**Table 5**
Top five and bottom five scoring other potential measures based on overall scores.

| Top five other measures | Bottom five other measures |
|---|---|
| • Investigative Journalism (4.48)<br>• Exchange of Best Practices (4.43)<br>• Education Campaigns (4.19)<br>• Framework, Guidelines, and Toolkits (4.14)<br>• Ethical Codes of Conduct (4.11) | • Self-Regulation by Company (3.58)<br>• Retaining “Unsmart” Products and Services (3.54)<br>• Ethical Rules pertaining to the creation or use of machine learning models with potential malicious applications (3.40)<br>• Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments (3.22)<br>• Citizen Juries (2.82) |

## References

Access Now, 2018. Human Rights in the Age of Artificial Intelligence. Access Now.

Access Now Policy Team, 2018. The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems. Access No, Toronto.

Adler, M., Ziglio, E. (Eds.), 1996. Gazing Into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health. Jessica Kingsley, London.

AI HLEG, 2019. Ethics Guidelines for Trustworthy AI. European Commission - Directorate-General for Communication, Brussels.

AIEI Group, 2020. From Principles to Practice - An Interdisciplinary Framework to Operationalise AI Ethics. VDE/Bertelsmann Stiftung.

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

Babuta, A., Oswald, M., Janjeva, A., 2020. Artificial Intelligence and UK National Security - Policy Considerations (Occasional Paper). Royal United Services Institute for Defence and Security Studies.

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

Bentham, J., 1789. An Introduction to the Principles of Morals and Legislation. Dover Publications Inc.

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

Boden, M.A., 2018. Artificial Intelligence: A Very Short Introduction, Reprint edition. OUP Oxford, Oxford, United Kingdom.

Borrás, S., Edler, J., 2020. The roles of the state in the governance of socio-technical systems'transformation. Res. Policy 49, 103971. https://doi.org/10.1016/j.respol.2020.103971.

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, 105–111. https://doi.org/10.1145/3072528.

British Academy, Royal Society, 2017. Data Management and Use: Governance in the 21st Century A Joint Report by the British Academy and the Royal Society. London.

Brooks, R.A., 2002. Flesh and Machines: How Robots Will Change Us. Pantheon Books, New York.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., hÉigeartaigh, S.Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., Amodei, D., 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation arXiv:1802.07228 [cs].

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

Buttarelli, G., 2018. Choose Humanity: Putting Dignity Back Into Digital.

Bynum, T., 2008. Computer and Information Ethics. Stanford Encyclopedia of Philosophy.

CDEI, 2019. Interim Report: Review Into Bias in Algorithmic Decision-making. Centre for Data Ethics and Innovation.

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

Coeckelbergh, M., 2019. Technology, narrative and performance in the social theatre. In: Kreps, D. (Ed.), Understanding Digital Events: Bergson, Whitehead, and the Experience of the Digital. Routledge, New York, pp. 13–27.

Coenen, C., Simakova, E., 2013. STS policy interactions, technology assessment and the governance of technovisionary sciences. Sci. Technol. Innov. Stud. 9, 3–20.

Council of Europe, 2019. Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights.

Dalkey, N.C., Brown, B.B., Cochran, S., 1969. The Delphi Method: An Experimental Study of Group Opinion. Rand Corporation, Santa Monica,CA.

Defense Innovation Board, 2019. AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. US Department of Defense.

Digital Catapult, 2020. Lessons in Practical AI Ethics: Taking the UK's AI Ecosystem From 'What' to 'How'. Digital Catapult, London.

Dignum, V., 2019. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. URL, 1st ed. Springer (accessed 2.10.18).

EDPS, 2020. A Preliminary Opinion on Data Protection and Scientific Research.

Eliot, L., 2022. Responsible AI Relishes Preeminent Boost Via AI Ethics Proclamation By Top Professional Society The ACM. Forbes.

Elsevier, 2018. Artificial Intelligence: How Knowledge Is Created, Transferred, and Used - Trends in China, Europe, and the United States. Elsevier, Amsterdam.

Ess, C., 2006. Ethical pluralism and global information ethics. Ethics Inf. Technol. 8, 215–226.

European Commission, 2021. Proposal for a Regulation on a European Approach for Artificial Intelligence (No. COM(2021) 206 final). European Commission, Brussels.

European Commission, 2020. White Paper on Artificial Intelligence: A European Approach to Excellence and Trust (White paper No. COM(2020) 65 final). Brussels.

European Commission, 2020. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS on A European Strategy for Data (No. COM/2020/66 final). European Commission, Brussels.

Executive Office of the President, 2016. Preparing for the Future of Artificial Intelligence. Executive Office of the President National Science and Technology Council Committee on Technology.

Executive Office of the President, 2016. Artificial Intelligence, Automation, and the Economy. Executive Office of the President National Science and Technology Council Committee on Technology.

Farah, M.J., 2005. Neuroethics: the practical and the philosophical. Trends Cogn. Sci. 9, 34–40.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M., 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI.

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

Floridi, L., 1999. Information ethics: on the philosophical foundation of computer ethics. Ethics Inf. Technol. 1, 33–52.

Foster-McGregor, N., Nomaler, Ö., Verspagen, B., 2021. Job automation risk, economic structure and trade: a European perspective. Res. Policy 104269. https://doi.org/10.1016/j.respol.2021.104269.

FRA, 2020. Getting the Future Right – Artificial Intelligence and Fundamental Rights. Eurepean Union Agency for Fundamental Rights, Luxembourg.

Friedman, B., Kahn, P., Borning, A., 2008. Value sensitive design and information systems. In: Himma, K., Tavani, H. (Eds.), The Handbook of Information and Computer Ethics. Wiley Blackwell, pp. 69–102.

Garriga, E., Melé, D., 2004. Corporate social responsibility theories: mapping the territory. J. Bus. Ethics 53, 51–71. https://doi.org/10.1023/B:BUSI.0000039399.90587.34.

Georghiou, L., Harper, J.C., Keenan, M., Miles, I., Popper, R., 2008. The Handbook of Technology Foresight: Concepts and Practice. Edward Elgar Publishing Ltd.

Hallamaa, J., Kalliokoski, T., 2022. AI ethics as applied ethics. Front. Comp. Sci. 4.

Haque, A., Milstein, A., Fei-Fei, L., 2020. Illuminating the dark spaces of healthcare with ambient intelligence. Nature 585, 193–202. https://doi.org/10.1038/s41586-020-2669-y.

IEEE, 2017. Ethically Aligned Design: A Vision for Prioritising Human Well-being With Autonomous and Intelligent Systems (Version 2 - For Public Discussion). IEEE.

IEEE, 2017. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [WWW Document]. URL. https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html (accessed 2.10.18).

Jelinek, T., Wallach, W., Kerimi, D., 2020. Policy brief: the creation of a G20 coordinating committee for the governance of artificial intelligence. AI Ethics. https://doi.org/10.1007/s43681-020-00019-y.

Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. Nat. Mach. Intell. 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Johnson, D.G., 2001. Computer Ethics, 3rd ed. Prentice Hall, Upper Saddle River, New Jersey.

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

Kant, I., 1788. Kritik der praktischen Vernunft. 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. Bus. Horiz. 62, 15–25.

Kuhlmann, S., Stegmaier, P., Konrad, K., 2019. The tentative governance of emerging science and technology—a conceptual introduction. Res. Policy 48, 1091–1097. https://doi.org/10.1016/j.respol.2019.01.006.

Kurzweil, R., 2006. The Singularity Is Near. Gerald Duckworth & Co Ltd, London.

Latonero, M., 2018. Governing Artificial Intelligence: Upholding Human Rights & Dignity. Data&Society.

Linstone, H.A., Turoff, M., 2011. Delphi: a brief look backward and forward. In: Technological Forecasting and Social Change, The Delphi Technique: Past, Present, and Future Prospects, 78, pp. 1712–1719. https://doi.org/10.1016/j.techfore.2010.09.011.

URL. In: Linstone, H.A., Turoff, M. (Eds.), 2002. The Delphi Method: Techniques and Applications. Addison-Wesley Publishing Company, Advanced Book Program (accessed 2.10.18).

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

Makridakis, S., 2017. The forthcomingArtificial Intelligence (AI) revolution: its impact on society and firms. Futures 90, 46–60. https://doi.org/10.1016/j.futures.2017.03.006.

Mantelero, A., Esposito, M.S., 2021. An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. Comput. Law Secur.Rev. 41, 105561 https://doi.org/10.1016/j.clsr.2021.105561.

Martin, B.R., 2010. The origins of the concept of "foresight" in science and technology: an insider's perspective. Technol. Forecast. Soc. Chang. 77, 1438–1447. https://doi.org/10.1016/j.techfore.2010.06.009.

McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., 2006. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Mag. 27, 12.

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

Miller, C., Ohrvik-Stott, J., 2018. Regulating for Responsible Technology - Capacity, Evidence and Redress: A New System for a Fairer Future. Doteveryone, London.

Morley, J., Floridi, L., Kinsey, L., Elhalal, A., 2019. From What to How- An Overview of AI Ethics Tools, Methods and Research to Translate Principles Into Practices arXive.

Muller, C., 2020. The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law (No. CAHAI(2020)06-fin). Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Strasbourg.

Müller, V.C., 2020. Ethics of artificial intelligence and robotics. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Murdick, D., Dunham, J., Melot, J., 2020. AI Definitions Affect Policymaking. Centre for Security and Emerging Technology, Georgetown University.

Murray, F., Stern, S., Campbell, G., MacCormack, A., 2012. Grand innovation prizes: a theoretical, normative, and empirical evaluation. Res. Policy 41, 1779–1792. https://doi.org/10.1016/j.respol.2012.06.013.

Nemitz, P., 2018. Constitutional democracy and technology in the age of artificial intelligence. Phil. Trans. R. Soc. A 376, 20180089. https://doi.org/10.1098/rsta.2018.0089.

OECD, 2019. Recommendation of the Council on Artificial Intelligence (OECD Legal Instruments). OECD.

Owen, R., Pansera, M., Macnaghten, P., Randles, S., 2021. Organisational institutionalisation of responsible innovation. Res. Policy 50, 104132. https://doi.org/10.1016/j.respol.2020.104132.

Pandza, K., Ellwood, P., 2013. Strategic and ethical foundations for responsible innovation. Res. Policy 42, 1112–1125. https://doi.org/10.1016/j.respol.2013.02.007.

Paré, G., Cameron, A.-F., Poba-Nzaou, P., Templier, M., 2013. A systematic assessment of rigor in information systems ranking-type Delphi studies. Inf. Manag. 50, 207–217. https://doi.org/10.1016/j.im.2013.03.003.

Peters, D., Vold, K., Robinson, D., Calvo, R.A., 2020. Responsible AI—two frameworks for ethical design practice. IEEE Trans.Technol.Soc. 1, 34–47. https://doi.org/10.1109/TTS.2020.2974991.

Powell, C., 2003. The Delphi technique: myths and realities. J. Adv. Nurs. 41, 376–382. https://doi.org/10.1046/j.1365-2648.2003.02537.x.

Rai, A., Constantinides, P., Sarker, S., 2019. Next-generation digital platforms: toward human–AI hybrids. MIS Q. 43, iii–x.

Rodrigues, R., Panagiotopoulos, A., Wright, D., Hatzakis, T., Laulhé Shaelou, S., Grant, A., Lundgren, B., Macnish, K., Ryan, M., Andreou, A., Zijlstra, T., Bijlsma, M., 2020. SHERPA Deliverable 3.3 Report on Regulatory Options (Online Resource No. Project Deliverable). URL. SHERPA Project. https://doi.org/10.21253/DMU.8181827.v2 (accessed 2.10.18).

Roessner, J.D., Frey, J., 1974. Methodology for technology assessment. Technol. Forecast. Soc. Chang. 6, 163–169. https://doi.org/10.1016/0040-1625(74)90015-8.

Rowe, E., Wright, G., Derbyshire, J., 2017. Enhancing horizon scanning by utilizing pre-developed scenarios: analysis of current practice and specification of a process improvement to aid the identification of important 'weak signals'. Technol. Forecast. Soc. Chang. 125, 224–235. https://doi.org/10.1016/j.techfore.2017.08.001.

Rowe, G., Wright, G., 2011. The Delphi technique: past, present, and future prospects — introduction to the special issue. In: Technological Forecasting and Social Change, The Delphi Technique: Past, Present, and Future Prospects, 78, pp. 1487–1490. https://doi.org/10.1016/j.techfore.2011.09.002.

Royakkers, L., Poel, I.van de, 2011. Ethics, Technology and Engineering: An Introduction. Wiley-Blackwell, Malden, Mass.

Russell, S.J., Norvig, P., 2016. Artificial Intelligence: A Modern Approach. Pearson Education Limited.

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

Sandrey, M.A., Bulger, S.M., 2008. The Delphi method: an approach for facilitating evidence based practice in athletic training. Athl. Train. Educ. J. 3, 135–142. https://doi.org/10.4085/1947-380X-3.4.135.

Sardar, Z., 2010. The namesake: futures; futures studies; futurology; futuristic; foresight—what's in a name? Futures 42, 177–184. https://doi.org/10.1016/j.futures.2009.11.001.

Schot, J., Rip, A., 1996. The past and future of constructive technology assessment. Technol. Forecast. Soc. Chang. 54, 251–268. https://doi.org/10.1016/S0040-1625(96)00180-1.

Stahl, B.C., 2021. Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies, SpringerBriefs in Research and Innovation Governance. Springer International Publishing. https://doi.org/10.1007/978-3-030-69978-9.

Stahl, B.C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., 2021. Artificial intelligence for human flourishing – Beyond principles for machine learning. J. Bus. Res. 124, 374–388. https://doi.org/10.1016/j.jbusres.2020.11.030.

Stahl, B.C., Rodrigues, R., Santiago, N., Macnish, K., 2022a. A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. Comput. Law Secur. Rev. 45, 105661 https://doi.org/10.1016/j.clsr.2022.105661.

Stahl, B.C., Antoniou, J., Ryan, M., Macnish, K., Jiya, T., 2022b. Organisational responses to the ethical issues of artificial intelligence. AI & Soc. 37, 23–37. https://doi.org/10.1007/s00146-021-01148-6.

Stahl, B.C., Markus, M.L., 2021. Let's claim the authority to speak out on the ethics of smart information systems. MIS Q. 45, 33–36. https://doi.org/10.25300/MISQ/2021/15434.1.6.

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

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., 2016. Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford University, Stanford, CA. Accessed: September 6, 2016. http://ai100.stanford.edu/2016-report.

Taddeo, M., Floridi, L., 2018. How AI can be a force for good. Science 361, 751–752. https://doi.org/10.1126/science.aat5991.

Tapio, P., Paloniemi, R., Varho, V., Vinnari, M., 2011. The unholy marriage? Integrating qualitative and quantitative information in Delphi processes. In: Technological Forecasting and Social Change, The Delphi Technique: Past, Present, and Future Prospects, 78, pp. 1616–1628. https://doi.org/10.1016/j.techfore.2011.03.016.

Tipler, F.J., 2012. Inevitable existence and inevitable goodness of the singularity. J. Conscious. Stud. 19, 183–193.

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

Vallor, S., 2016. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press.

Wiarda, M., van de Kaa, G., Yaghmaei, E., Doorn, N., 2021. A comprehensive appraisal of responsible research and innovation: from roots to leaves. Technol. Forecast. Soc. Chang. 172, 121053 https://doi.org/10.1016/j.techfore.2021.121053.

Willcocks, L., 2020. Robo-apocalypse cancelled? Reframing the automation and future of work debate. J. Inf. Technol. 35, 286–302. https://doi.org/10.1177/0268396220925830.

Winkler, T., Spiekermann, S., 2018. Twenty years of value sensitive design: a review of methodological practices in VSD projects. Ethics Inf. Technol. https://doi.org/10.1007/s10676-018-9476-2.

Wright, D., Stahl, B., Hatzakis, T., 2020. Policy scenarios as an instrument for policymakers. Technol. Forecast. Soc. Chang. 154, 119972 https://doi.org/10.1016/j.techfore.2020.119972.

Ziglio, E., 1996. The Delphi method and its contribution to decision making. In: Adler, M., Ziglio, E. (Eds.), Gazing Into the Oracle: The Delphi Method and its Application to Social Policy and Public Health. Jessica Kingsley, London, pp. 3–33.

Zuboff, P.S., 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, 1st edition. Profile Books.

**Bernd Carsten Stahl** is Professor of Critical Research in Technology at the School of Computer Science of the University of Nottingham. 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.

**Laurence Brooks** is Professor of Information Systems in the Information School (iSchool) in the University of Sheffield and Visiting Professor of Computing and Social Responsibility in the Centre for Computing and Social Responsibility in De Montfort University. His interests lie in the use of technology (both current and emerging) by and for society and organizations, from a socio-technical and ethical perspective.

**Tally Hatzakis** is a Senior Research Analyst at Trilateral Research. Tally led foresight research for AI futures and coordinated case study research for the use of AI in several domains and consulted AI-related fora. She has published more than 40 articles in peer-reviewed journals and worked on a wide range of technology-driven phenomena for projects funded by the EU, ESRC and EPSRC.

**Nicole Santiago** is a US-licensed attorney with a specialisation in international and human rights law. Her recent work has focused on the intersection of ethics, policy, and emerging technology, including contributions to EU-funded projects on AI governance. She consults for policy-makers and the private sector on technology governance and has taught at the University-level on issues related to law and technology.

**David Wright** is a Director and Chief Research Officer of Trilateral Research, a company he co-founded in 2004. He recently led a study on AI risk assessments for EY and a study for ENISA on a cybersecurity market analysis framework. He has co-edited and co-authored several books including Privacy Impact Assessment (Springer, 2012) and Surveillance in Europe (Routledge, 2015) and more than 75 articles in peer-reviewed journals.