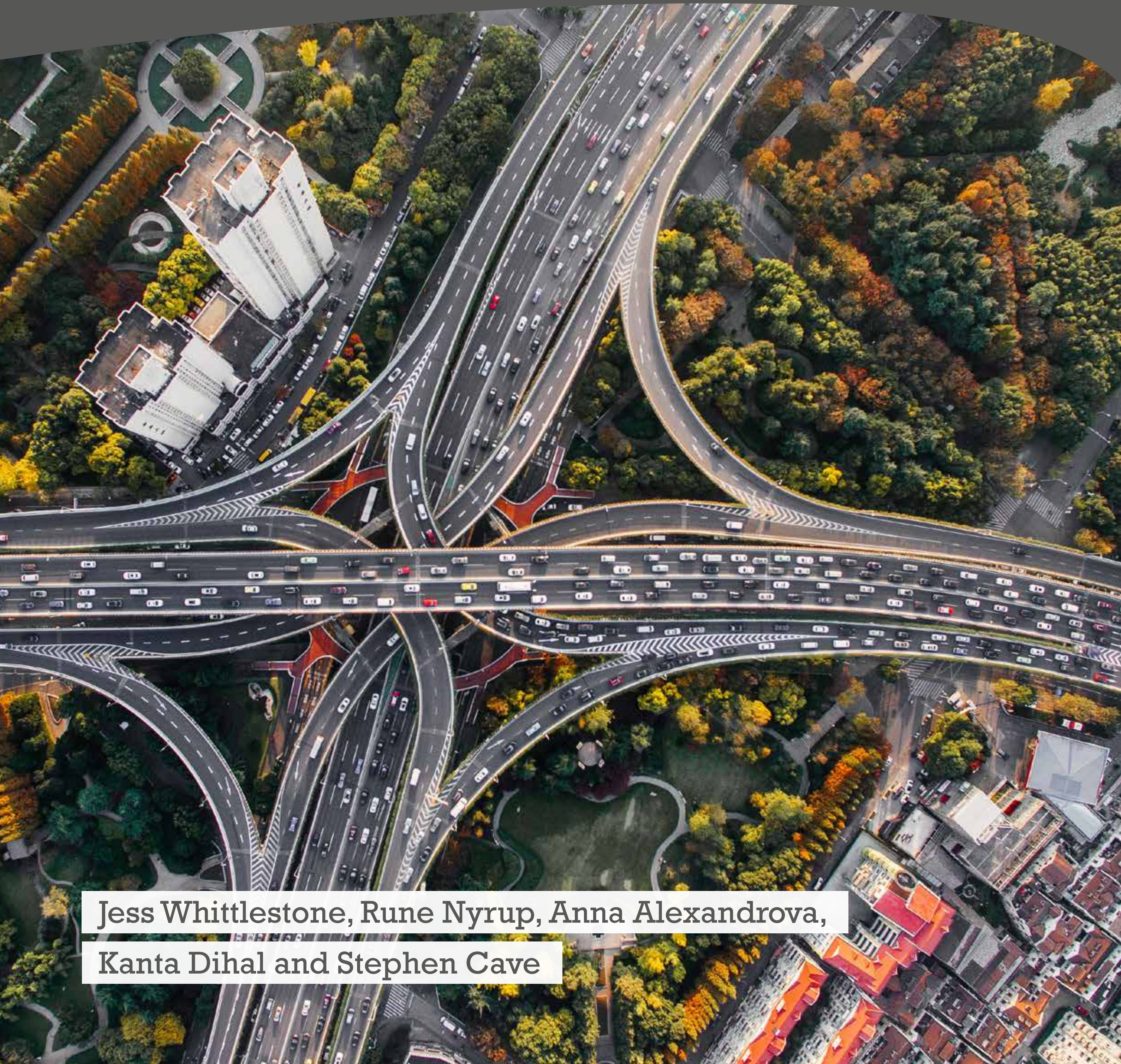
Nuffield Foundation

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research



Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal and Stephen Cave

Authors: Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, Stephen Cave, *Leverhulme Centre for the Future of Intelligence, University of Cambridge.*

With research assistance from: Ezinne Nwankwo, José Hernandez-Orallo, Karina Vold, Charlotte Stix.

Citation: Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019) *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* London: Nuffield Foundation.

ISBN: 978-1-9160211-0-5

The Nuffield Foundation has funded this project, but the views expressed are those of the authors and not necessarily those of the Foundation.

Acknowledgements

The authors are grateful to the following people for their input at workshops and valuable feedback on drafts of this report: Haydn Belfield, Jude Browne, Sarah Castell, Jennifer Cobbes, Damian Clifford, Matthew Crosby, Naomi Fern, Danit Gal, Julian Huppert, Stephen John, Martina Kunz, Nóra Ní Loideáin, Jonnie Penn, Huw Price, Diana Robinson, Henry Shevlin, Jeffrey Skopek, Adrian Weller, Alan Winfield.

Reviewers: Alan Wilson, Sophia Adams-Bhatti, Natalie Banner, Josh Cowls, Claire Craig, Tim Gardam, Helen Margetts, Natasha McCarthy, Imogen Parker, Reema Patel, Hetan Shah, Olivia Varley-Winter, Antony Walker.

About the Nuffield Foundation

The Nuffield Foundation funds research, analysis, and student programmes that advance educational opportunity and social well-being across the United Kingdom. The research we fund aims to improve the design and operation of social policy, particularly in Education, Welfare, and Justice. Our student programmes provide opportunities for young people, particularly those from disadvantaged backgrounds, to develop their skills and confidence in quantitative and scientific methods.

We have recently established the Ada Lovelace Institute, an independent research and deliberative body with a mission to ensure data and AI work for people and society. We are also the founder and co-funder of the Nuffield Council on Bioethics, which examines and reports on ethical issues in biology and medicine.

We are a financially and politically independent charitable trust established in 1943 by William Morris, Lord Nuffield, the founder of Morris Motors.

28 Bedford Square, London WC1B 3JS
T: 020 7631 0566

Registered charity 206601

www.nuffieldfoundation.org
www.adalovelaceInstitute.org
@NuffieldFound | @AdaLovelaceInst

Foreword

This report sets out a broad roadmap for work on the ethical and societal implications of algorithms, data and AI (ADA). Their impact on people and society shapes practically every question of public policy, but discussion is not necessarily based on a shared understanding of either the core ethical issues, or an agreed framework that might underpin an ethical approach to the development and deployment of ADA-based technologies. Even where there is a broad consensus on core issues, such as bias, transparency, ownership and consent, they can be subject to different meanings in different contexts – interpretation in technical applications differs to that in the judicial system, for example. Similarly, ethical values such as fairness can be subject to different definitions across different languages, cultures and political systems.

Clarifying these concepts, and resolving the tensions and trade-offs between the central principles and values in play, is crucial if we want ADA-based technologies to be developed and used for the benefit of society. The roadmap identifies, for the first time, the directions for research that need to be prioritised in order to build a knowledge base and shared discourse that can underpin an ethical approach. For each of the key tasks identified, the authors provide detailed questions that, if addressed, have the collective potential to inform and improve the standards, regulations and systems of oversight of ADA-based technologies.

The Nuffield Foundation has recently established – in partnership with others – the Ada Lovelace Institute (Ada), an independent research and deliberative body with a mission to ensure data and AI work for people and society. In commissioning this roadmap, our intention was to inform both Ada's work programme, and to help shape the research agenda on the increasingly important question of how society should equitably distribute the transformative power and benefits of data and AI while mitigating harm.

The message emerging from the roadmap is that the study of the questions it sets out must be plural, interdisciplinary, and connect different interests across academic research, public policy, the private sector and civil society. This is very much at the heart of the Ada Lovelace Institute's mission. One of Ada's core aims is to convene diverse voices to create a shared understanding of the ethical issues arising from data and AI, and an interdisciplinary and collaborative approach will be central to its operation.

As an independent funder with a mission to advance social well-being, the Nuffield Foundation is keen to fund more research in this area. The question of how digital technologies, and their distributional effects, can alleviate, exacerbate and shift vulnerability and affect concepts of trust, evidence, and authority is one of the themes prioritised in our strategy. We hope that this roadmap will help to generate relevant research proposals.

I thank the authors of this report for delivering an intellectually stimulating and, at the same time practical, contribution to this important field.

Tim Gardam.

Tim Gardam
Chief Executive

Executive Summary

The aim of this report is to offer a broad roadmap for work on the ethical and societal implications of algorithms, data, and AI (ADA) in the coming years. It is aimed at those involved in planning, funding, and pursuing research and policy work related to these technologies. We use the term 'ADA-based technologies' to capture a broad range of ethically and societally relevant technologies based on algorithms, data, and AI, recognising that these three concepts are not totally separable from one another and will often overlap.

A shared set of key concepts and concerns is emerging, with widespread agreement on some of the core issues (such as bias) and values (such as fairness) that an ethics of algorithms, data, and AI should focus on. Over the last two years, these have begun to be codified in various codes and sets of 'principles'. Agreeing on these issues, values and high-level principles is an important step for ensuring that ADA-based technologies are developed and used for the benefit of society.

However, we see three main gaps in this existing work: (i) a lack of clarity or consensus around the meaning of central ethical concepts and how they apply in specific situations; (ii) insufficient attention given to tensions between ideals and values; (iii) insufficient evidence on both (a) key technological capabilities and impacts, and (b) the perspectives of different publics.

In order to address these problems, we recommend that future research should prioritise the following broad directions (more detailed recommendations can be found in section 6 of the report):

1. **Uncovering and resolving the ambiguity inherent in commonly used terms** (such as privacy, bias, and explainability), by:

   a. Analysing their different interpretations.

   b. Identifying how they are used in practice in different disciplines, sectors, publics, and cultures.

   c. Building consensus around their use, in ways that are culturally and ethically sensitive.

   d. Explicitly recognising key differences where consensus cannot easily be reached, and developing terminology to prevent people from different disciplines, sectors, publics, and cultures talking past one another.

2. **Identifying and resolving tensions between the ways technology may both threaten and support different values,** by:

   a. Exploring concrete instances of the following tensions central to current applications of ADA:

      i. Using algorithms to make decisions and predictions more accurate **versus** ensuring fair and equal treatment.

      ii. Reaping the benefits of increased personalisation in the digital sphere **versus** enhancing solidarity and citizenship.

      iii. Using data to improve the quality and efficiency of services **versus** respecting the privacy and informational autonomy of individuals.

      iv. Using automation to make people's lives more convenient **versus** promoting self-actualisation and dignity.

   b. Identifying further tensions by considering where:

      i. The costs and benefits of ADA-based technologies may be unequally distributed across groups, demarcated by gender, class, (dis)ability, or ethnicity.

      ii. Short-term benefits of technology may come at the cost of longer-term values.

      iii. ADA-based technologies may benefit individuals or groups but create problems at a collective level.

   c. Investigating different ways to resolve different kinds of tensions, distinguishing in particular between those tensions that reflect a fundamental conflict between values and those that are either illusory or permit practical solutions.

3. **Building a more rigorous evidence base for discussion of ethical and societal issues,** by:

   a. Drawing on a deeper understanding of what is technologically possible, in order to assess the risks and opportunities of ADA for society, and to think more clearly about trade-offs between values.

b. Establishing a stronger evidence base on the current use and impacts of ADA-based technologies in different sectors and on different groups – particularly those that might be disadvantaged, or underrepresented in relevant sectors (such as women and people of colour) or vulnerable (such as children or older people) – and to think more concretely about where and how tensions between values are most likely to arise and how they can be resolved.

c. Building on existing public engagement work to understand the perspectives of different publics, especially those of marginalised groups, on important issues, in order to build consensus where possible.

Contents

1. Introduction

1.1 Aims, approach and outline

The aim of this report is to offer a roadmap for work on the ethical and societal implications of algorithms, data, and AI (ADA) in the coming years. We review what progress has been made in understanding these issues across academia, policy, and industry, identify gaps in the current research landscape, and assess the strengths and limitations of existing work. On this basis, we recommend three broad areas of research, and highlight specific priority questions within each of the three areas. These recommendations, and the report in general, are aimed at individuals and organisations involved in planning, funding, and pursuing research and policy work related to the emerging ethical and societal challenges raised by algorithms, data and AI. Our focus is on the short- to medium-term issues that have already emerged or are in the process of emerging at this time; we do not focus on solutions that involve radical political or technological transformation.[1] We also focus primarily on priorities for research rather than for the role of policy or regulation. However, we urge that these options are also researched and point out in this report how this might be done.

To arrive at these recommendations, we began by conducting a wide-ranging literature review of relevant work in the English language: covering over 100 academic papers, both theoretical and empirical, from disciplines including (but not limited to) computer science, ethics, human-computer interaction, law, and philosophy. We also reviewed key policy documents from across several continents, and some of the most commonly cited popular news and media articles from the last few years.[2] We held three workshops (each bringing together at least twenty different experts from a range of relevant fields), and a series of smaller discussion and brainstorming sessions in groups of between five and 10 people.

This report is organised as follows:

- Section 2 provides a high-level summary of the current landscape, based on a more detailed literature review, which can be found in appendix 1. We highlight some of the successes of research so far, and some of the gaps that still exist. We conclude that the road forward is in particular need of three broad types of work: concept building, identifying and resolving tensions and trade-offs, and building a stronger evidence base around these tensions.

- Sections 3–5 focus on each of these recommended areas of work in turn: explaining in more detail why this is a priority, what research in this area should consider in general, and what specific questions or areas seem particularly important. Section 6 draws the conclusions of the preceding sections together to present a 'roadmap': a set of high-level recommended research directions.

1.2 Definitions and key terms

The scope of this investigation was broad: to consider both the ethical and societal implications of algorithms, data, and artificial intelligence.

Ethical and societal implications

We adopt a broad definition of 'ethical and societal implications', to consider the ways that algorithms, data and AI may impact various parts of society, and how these impacts may either enhance or threaten widely held values. We deliberately use the term 'implications' to capture that we are not just interested in the *negative* impacts of these technologies (as alternative terms like 'issues', 'risks', or 'challenges' might suggest), but also the

[1] Other groups have focused on prioritising research on related, longer-term challenges of advanced AI systems, most notably the Future of Humanity Institute at Oxford University, which recently published a research agenda for long-term AI governance. It would be valuable for future work to look at the interrelations between short- and longer-term research priorities, and how the two might better learn from one another.

[2] We chose media articles that we either found to be cited repeatedly in the academic literature we reviewed, or those that were written in the last year in high-profile outlets such as the New York Times, the Guardian, or TechCrunch. A more systematic and comprehensive review of media outlets was beyond the scope of this initial review, but a broader analysis in follow-up work could strengthen our assessment of the space.

positive impacts they might produce. This is crucial, as we will later emphasise the importance of considering the *tensions* that arise between the opportunities and risks of technologies based on algorithms, data, and AI.

Values

When we speak about recognising conflicts between different values enhanced or endangered by new technologies, we use the term 'values' to pick out commitments that are deeply held and reasonably widely shared. Values are not mere desires, revealed preferences[3] or pleasures. They are goals and ideals that people endorse thoughtfully and defend to each other as appropriate, and which motivate ways in which they organise communal life.[4] Here we concentrate on those values that have been invoked especially frequently in the recent anglophone debates about emerging AI and data-based technologies, but also try to identify those that resonate more widely across cultures.

Algorithms

In mathematics and computer science, the term 'algorithm' means an unambiguous procedure for solving a given class of problems. In this report, we primarily use 'algorithm' to mean something closer to 'automated algorithm': a procedure used to automate reasoning or decision-making processes, typically carried out by a digital computer. Often we will simply use 'algorithm' as a shorthand to refer to the software that implements this procedure, and terms like 'algorithmic decision-making' more or less as a synonym for computerised decision-making. For our purposes, the key aspect of algorithms is that they can be automated, and can be executed systematically at much higher speeds than humans, also automatically triggering many other procedures as a result.

Data

We define data as 'encoded information about one or more target phenomena' (such as objects, events, processes, or persons, to name a few possibilities). Today, data is usually encoded digitally rather than analogically. Data is ethically and societally relevant for three reasons. First, the process of collecting and organising data itself requires making assumptions about what is significant, worthy of attention, or useful. Since these assumptions are unlikely to hold in all contexts, no dataset is fully complete, accurate, or neutral. Second, digitally encoded data allows information to be duplicated, transferred, and transformed much more efficiently than ever before. Third, new forms of analysis allows those possessing large amounts of data to acquire novel insights.

Artificial Intelligence

Of the three key terms explored in this report, 'artificial intelligence' (AI) is probably the hardest and most controversial to define. The word 'intelligence' is used in many different ways both in ordinary discourse and across a number of different academic fields, often with politically loaded connotations.[5] For the purpose of this report, we take 'artificial intelligence' to refer to any technology that performs tasks that might be considered intelligent – while recognising that our beliefs about what counts as 'intelligent' may change over time. For example, we don't intuitively think of visual perception or walking as particularly 'intelligent' tasks, because they are things we do with little conscious effort: but attempting to replicate these abilities in machines has shown they actually rely on incredibly complex processes. We also consider the key feature of AI most relevant to ethics and society: the fact that AI can often be used to optimise processes and may be developed to operate autonomously, creating complex behaviours that go beyond what is explicitly programmed.

Publics

The term 'public' is often taken for granted as a catch-all term for 'every person in society'. We, on the other hand, use the term 'publics' in plural to emphasise that different interest groups (scientists, mediators, decision-makers, activists, etc.) bring their own distinct perspectives.[6]

[3] That is, we do not think it is possible to infer values simply by observing how they behave in a market place.

[4] Tiberius (2018) provides a fuller definition.

[5] See Cave (2017).

[6] Here we follow Burns et al. (2003), who distinguish different publics as being relevant for different contexts.

This allows us to avoid focusing on the dominant views and attitudes at the expense of those coming from the margins. We do not use the term 'lay public' in opposition to 'experts' in recognition of the fact that many different groups have relevant expertise.

With these definitions in hand, we can clarify why the ethical and societal implications of ADA-based technologies motivate concern. ADA-based technologies are dual-use in nature: the purpose to which they are initially developed can easily be changed and transferred, often radically altering their moral valence. For example, image recognition techniques have clearly positive applications such as in the identification of malignant tumours, but can also be repurposed in ways that could be harmful, such as for mass surveillance (Bloomberg News, 2018). Relatedly, ADA-based technologies involve inherently domain-neutral capacities, such as information processing, knowledge acquisition and decision-making. Thus, the same techniques can be applied to almost any task, making them increasingly pervasive and permeable across different parts of society. The same technology could carry very different risks and benefits in different application areas, for different publics, touching on different values.

These features, together with the remarkable speed with which powerful private companies have pioneered new applications of ADA-based technologies in the recent decade, explain the increase in focus on the need to regulate and guide the ethics of ADA in the right direction.

2. The current landscape

Discussion of the ethical and societal implications of algorithms, data, and AI does not fit neatly into a single academic discipline, or a single sector. To understand the full range of recent coverage of these issues, we therefore need to look very broadly: at academic publications from philosophy and political science to machine learning, and beyond academia to policy, industry, and media reports. Our review focused on understanding two main things: (1) what specific issues and concerns are given attention across different types of publication, and (2) what attempts have been made to synthesise issues across disciplines and sectors.

We drew two main conclusions from this review. First, a shared set of key concepts and concerns *is* emerging, but the terms used are often ambiguous or used unreflectively. Second, several different attempts to synthesise issues into frameworks and sets of principles exist,[7] but are often unsystematic or too high-level to guide practical action.[8]

2.1 Identifying key concepts and concerns

While the existing literature covers a wide range of issues, a shared set of key concepts and concerns is nonetheless emerging. Concerns about algorithmic bias and ensuring that machine learning that supports decisions about individuals is used fairly, for example, have become a centrepiece of these discussions, as has an emphasis on the importance of making 'black box' systems transparent and explainable. Issues of personal data privacy also arise repeatedly, as do questions of how we maintain accountability and responsibility as more and more decisions become automated. The impact of ADA-based technologies on the economy and implications for the future of work are further themes that arise frequently. See figure 1 for an illustration of the most common terms used in recent attempts to list the key issues arising from ADA. This word cloud is based on the frequency of terms as they arise in the various frameworks and categories we reviewed, with larger words occurring more frequently.

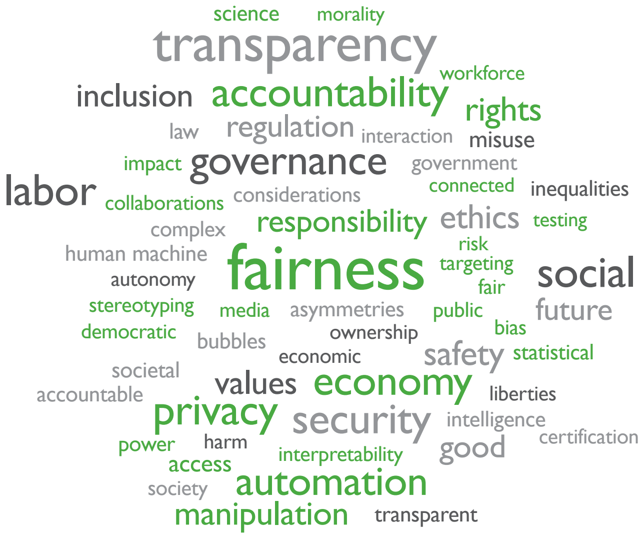


Figure 1. Word cloud of emerging shared concepts, based on the frequency of words used in the groupings and frameworks of several key reports and organisations. See appendix 2 for details.

However, as terms rise in popularity, they may be used unreflectively or ambiguously. For example, commentators frequently champion the importance of 'transparency' without clarifying exactly what they mean by it or why it is important. There is also inconsistency in the meanings attached to these terms in different contexts: for example, 'bias' might mean something quite precise in a technical paper, but something more vague in a policy report. We discuss how different uses and interpretations of the same terms may cause problems in section 3.

Although consensus on key issues is emerging, disciplines of course differ in their areas of emphasis. Unsurprisingly, computer science and machine learning research focuses mostly on those ethical issues that can most easily be framed in technical terms: including how to make machine learning systems more interpretable and reliable, and issues of privacy and data protection. Philosophy and ethics papers often focus on questions about the moral significance of more advanced AI systems that could exist in the future, with less attention paid to the ethical challenges of current

[7] See for example Cowls and Floridi (2018) for a framework-focused approach, and the House of Lords Select Committee on AI's (2018) report for a principles-focused approach.

[8] This section is restricted to providing a high-level assessment of the current landscape of ADA ethics and societal impacts. For more detailed descriptions and assessments, see appendices 1–4.

technologies – though a body of literature on these more near-term issues is emerging in fields such as information and technology ethics. Academic law literature does much more than other areas we reviewed to try to pull apart different interpretations of different terms such as 'privacy' and 'fairness', and to discuss the implications of these different meanings.

When we look beyond research papers tackling specific issues, towards high-level attempts to synthesise a range of issues, we find many take similar approaches to grouping or categorising these issues.[9] For example, many similarities can be seen between the categories that DeepMind Ethics and Society (DMES) and the Partnership on AI (PAI) use to define their research areas:

DMES Research Themes[10]	**PAI Thematic Pillars**[11]
Privacy, transparency and fairness	Fair, transparent and accountable AI
Economic impact, inclusion, and equality	AI, labor, and the economy
Governance and accountability	Social and societal influences of AI
AI morality and values	AI and social good
Managing AI risk, misuse, and unintended consequences	Safety-critical AI
AI and the world's complex challenges	Collaborations between people and AI systems

Though these groupings hint at some underlying structure, as they stand they are relatively unsystematic. This is illustrated by the subtle differences in how different groups place the boundaries of their categories. Does 'accountability' belong in the same category as 'fairness and transparency' (as the Partnership on AI have it), or should it fall into a separate category with issues of governance and regulation (as DeepMind have it)? Should 'trust' be categorised with either 'transparency', or 'fairness', or 'privacy' – or should all these issues be lumped together? Should 'AI for social good' be a category of its own or does it cut across all the other categories? What issues might not fit neatly into any of these categories at all (such as AI Now's notion of 'rights and liberties'[12])?

Without an understanding of why these issues and categories have been chosen and not others, it is difficult to be confident that all the relevant issues have been captured. It is not clear whose values and priorities are being promoted, and whether the concerns of all members of society – including minority groups – are being represented. Some groups and papers are beginning to take an approach that starts with a more fundamental map of the ethical landscape: for example, a 2018 report from the EDPS Ethics Advisory Group, 'Towards a Digital Ethics', systematically considers each of the 'European' values, and how they might be threatened by the features of an increasingly digital world. This highlights some questions that have not been given so much attention, such as how individualised profiling might threaten solidarity in society, or how the availability of data might worsen power imbalances between governments and companies on the one hand, and individuals on the other.

Efforts like these aim to produce a single theoretical framework that can be presented as a single list of principles and values. While unity is valuable for some endeavours (e.g. for coordination and public accountability), it can also restrict attention: highlighting some issues at the expense of masking others. For now, it is clear that there remain many possible ways to carve up this space, each of which will have different advantages and disadvantages, and prioritise some values above others.[13]

[9] See appendix 2 for recently proposed lists of key issues.

[10] https://deepmind.com/applied/deepmind-ethics-society/research/

[11] www.partnershiponai.org/about/#our-work

[12] https://ainowinstitute.org/research.html

[13] For a more detailed assessment of the strengths and weaknesses of different approaches to organising the issues, see appendix 2. Appendix 3 contains several of perspectives which can be used to restrict discussions to a more limited set of issues.

2.2. Formulating ethical principles

In addition to these explorations of key concepts, various groups have also begun to publish sets of prescriptive principles or codes to guide the development and use of ADA-based technologies. These principles often overlap with and include concepts mentioned in the previous section, but focus less on articulating what the 'issues' are, and instead on articulating some goals for the use and development of technology.

For example, the Asilomar AI principles developed in 2017 in conjunction with the Asilomar conference for Beneficial AI,[14] outline guidelines on how research should be conducted, ethics and values that AI must respect, and important considerations for thinking about longer-term issues. The principles were signed by several thousand AI researchers and others, including many academic ethicists and social scientists. The Partnership on AI has also established a set of 'tenets' to guide the development and use of AI technologies, which all members – including many of the most prominent technology companies – endeavour to uphold.[15]

In addition, governments and international bodies are developing their own principles: a recent report from the Lords Select Committee on Artificial Intelligence, 'AI in the UK: ready, willing, and able?'[16] suggests five principles for a cross-sector AI code which could be adopted internationally. The IEEE Standards Association has also launched a 'Global Initiative on Ethics of Autonomous and Intelligent Systems'[17], and has developed a set of general principles to guide ethical governance of these technologies. Industry is also getting involved: most prominently with Google publishing its 'AI ethics principles' in June this year.[18] Figure 2 illustrates the key terms that arise across all these sets of principles we reviewed, again where word size corresponds to frequency.

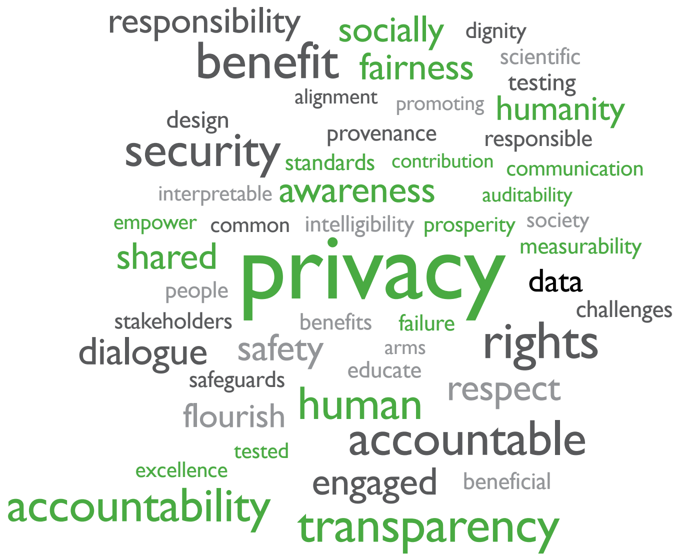


Figure 2. Word cloud of concepts frequently occurring in principles and codes, based on the frequency of words used in the principles outlined in appendix 2.

There is substantial overlap between these different sets of principles. For example, there is widespread agreement that ADA-based technologies should be used for the common good, should not be used to harm people or undermine their rights, and should respect some of the widely-held values mentioned above such as fairness, privacy, and autonomy. There have also been attempts to synthesise them into a short list of key principles (e.g. beneficence, non-maleficence, autonomy, justice, and explicability)[19] modelled on a prominent tradition within biomedical ethics.

Principles are a valuable part of any applied ethics: they help to condense complex ethical issues into a few central elements which can allow widespread commitment to a shared set of values. They can also provide an informal means of holding people and organisations accountable, to reassure public concerns. For example, the machine learning community has mobilised over the issue of autonomous weapons in the past year, with many groups and individual researchers making public commitments not to be involved in their development. This is the case where joint commitment to a specific and action-guiding principle can have a real impact on the ethical implications of technology.

However, most of the principles proposed for AI ethics are not specific enough to be action-guiding. While these

[14] https://futureoflife.org/ai-principles/. Some of the authors of this report were present at that conference and involved in the development of the principles.

[15] www.partnershiponai.org/tenets/. The Leverhulme Centre for the Future of Intelligence, at which the authors of this report are based, is a member of the Partnership on AI.

[16] Some of the authors of this report gave evidence to the Committee.

[17] https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html. Some of the authors of this report have been involved with this initiative.

[18] https://ai.google/principles/

[19] Cowls and Floridi (2018).

principles do reflect agreement about which aims are important and desirable as the development and use of ADA-based technologies advances, they do not provide practical guidance to think through new and challenging situations. The real challenge is recognising and navigating the tensions *between* principles that will arise in practice. For example, a truly beneficial application of AI that could save lives might involve using personal data in a way that threatens commonly held notions of privacy, or might require us to use algorithms that we cannot entirely explain. Discussion of principles must begin to acknowledge these tensions, and provide guidelines for how to navigate the trade-offs they introduce.[20]

2.3 Underlying assumptions and knowledge gaps

It is also worth noting several assumptions implicit in current discussions, as they reveal gaps in the existing knowledge about what is, and what will be, technically possible, and about the values of different groups in society.

For example, concerns about algorithmic bias often presuppose that the human decision-makers being replaced by algorithms are not equally or even more biased. While this comparison is sometimes raised, it is rarely investigated systematically when we should expect algorithmic systems to do better than humans, and when they merely perpetuate or reinforce human biases. Emphasis on algorithmic transparency assumes that some kind of 'explainability' is important to all kinds of people, but there has been very little attempt to build up evidence on which *kinds* of explanations are desirable to which people in which contexts. Discussions of the future of work are often underpinned by assumptions about the benefits and harms of different forms of automation, but lack substantive evidence on either the objective benefits and harms of automation so far, or public opinion on these topics.

Putting principles into practice and resolving tensions will require us to identify these kinds of assumptions and fill knowledge gaps around technological capabilities, the impacts of technology on society, and public opinion. Without understanding current applications of ADA-based technologies and their impacts on society, we cannot clearly identify the issues and tensions which are most pressing. Without understanding what is technologically feasible, it is difficult to have a meaningful discussion about what trade-offs exist and how they might be navigated. And without understanding the perspectives of various different groups in society, we risk making trade-offs that favour the values and needs of the majority at the expense of minorities. It is not enough to agree that we must preserve human autonomy, for example: we need to develop a rigorous understanding of the specific ways that technology might undermine autonomy now and in future, and in what contexts different people might be willing to sacrifice some amount of autonomy for other goods.

2.4 Summary and recommendations

To summarise:

- A useful set of shared concepts is emerging, but is currently based on ambiguous terms often used unreflectively. There are important ambiguities in many of the terms often used, which may mask significant differences in how concepts are understood by different disciplines, sectors, publics and cultures.

- Important codes and principles are being established, but there is little recognition of the tensions that will inevitably be encountered in putting these principles into practice: when values come into conflict with one another, when there are conflicts between the needs of different groups, or when there are resource limitations.

- Current discussion of issues and principles often rely on implicit assumptions about what is technically possible, how technology is impacting society, and what values society should prioritise. To put principles into practice and resolve these tensions, it is crucial to identify and challenge these assumptions, building a stronger and more objective evidence base for understanding underlying technological capabilities, societal impacts and societal needs.

Substantial progress has been made over the last few years on understanding the ethical and societal implications of ADA, the challenges and questions these raise, and how we might address them. The road ahead needs to focus on:

- **Building a shared understanding of key concepts** that acknowledges and resolves ambiguities, and bridges disciplines, sectors, publics and cultures. In section 3, we begin to unpack some of the terminological overlaps, different uses and

[20] We make a full case for this in Whittlestone et al (2019).

interpretations, and conceptual complexities which contribute to confusion and disagreement.

- **Identifying and exploring the tensions** that arise when we try to put agreed-upon principles into practice. In section 4 of this report, we begin to do exactly this: identifying and unpacking in detail several tensions that are illustrative of the conflicts emerging in this space more broadly, and outlining some guidelines for resolving these tensions.

- **Deepening understanding of technological capabilities, societal impacts, and the perspectives of different groups**, in order to better understand the issues that arise and how to resolve them. In section 5, we explain why understanding and challenging assumptions about technology and society is crucial for resolving tensions, and highlight some priority areas for research.

In each of the proceeding sections, we also highlight research priorities and recommendations for future work.

3. Concept building

An important obstacle to progress on the ethical and societal issues raised by ADA is the ambiguity of many central concepts currently used to identify salient issues. As reviewed in section 2, concepts like 'fairness', 'transparency' and 'privacy' figure prominently in the existing literature. While they have served to highlight common themes emerging from case studies, many of these terms are overlapping and ambiguous. This stems partly from the fact that different fields, disciplines, sectors, and cultures can use these concepts in substantially different ways, and partly from inherent complexities in the concepts themselves. As a result, discussions of the ethical and societal impacts of ADA risk being hampered by different people talking past each other.

Making constructive progress in this space requires conceptual clarity, to bring into sharper focus the values and interests at stake. In this section we outline in detail the different challenges that need to be overcome in order to achieve this conceptual clarity.

3.1 Terminological overlaps

One challenge is that different terms are often used to express overlapping (though not necessarily identical) phenomena.

For example, the terms 'transparency', 'explainability', 'interpretability', and 'intelligibility' are often used interchangeably to refer to what 'black-box' algorithms are thought to be missing. Commentators have pointed out that these terms can refer to a number of distinct problems.[21] Is the problem that companies or state agencies refuse to share their algorithms? Or that the models themselves are too complex for humans to parse? And are we talking about any human or merely people with the relevant scientific knowledge or expertise? While all of these questions may in a loose sense be said to involve problems of transparency, they raise different kinds of challenges and call for different kinds of remedies.

Similarly, the terms 'bias', 'fairness', and 'discrimination' are often used to refer to problems involving datasets or algorithms which (in some sense) disadvantage certain individuals or groups. Again, it is not clear that all cases referred to by these terms involve the same type of problem.[22]

Some research has begun to untangle these overlaps.[23] For example, Barocas (2014) distinguishes three kinds of concerns for algorithms based on data-mining, which have been raised under the heading 'discrimination':

1. Cases where deployers of an algorithm deliberately attempt to disadvantage certain users and make this difficult to detect (e.g. by hiding the critical bit of code within a complicated algorithm).

2. Cases where data-mining techniques produce errors which disadvantage certain users (e.g. due to unreliable input data or users drawing faulty inferences from the algorithms' output).

3. Cases where an algorithm enhances decision-makers' ability to distinguish and make differential decisions between people (e.g. allowing them to more accurately identify and target financially vulnerable individuals for further exploitation).

Disentangling different issues lumped together under a single term in this way is an important first step towards conceptual clarification, as different types of issues arguably require different types of remedy.

3.2 Differences between disciplines

A further challenge stems from the fact that some of the most widely used terms have different connotations and meanings in different contexts.

For example, in statistics a 'biased sample' means a sample that does not adequately represent the distribution of features in the reference population (e.g. it contains a higher proportion of young men than in the overall population). In law and social psychology,

[21] Burrell 2016; Lipton (2016); Weller (2017); Selbst & Barocas (2018).

[22] Barocas (2014); Binns (2017).

[23] E.g. Barocas (2014); Burrell (2016); Weller (2017); Zarsky (2016); Mittelstadt et al (2016).

by contrast, the term 'bias' often carries the connotation of negative attitudes or prejudices towards a particular group. In this sense, a dataset which is 'unbiased' (in the statistical sense) may nonetheless encode common biases (in the social sense) towards certain individuals or social groups. Distinguishing these different uses of the same term is important to avoid cross-talk.[24]

Apart from these terminological issues, different disciplines also embody different research cultures that can affect the clarification and refining of ambiguous concepts. For instance, many machine learning researchers would naturally seek to construct a mathematically precise definition of, say, 'fairness',[25] whereas qualitative social scientists would often seek to highlight the rich differences in how different stakeholders understand the concept. Similarly, philosophical ethicists often seek to highlight inherent dilemmas and in-principle problems for different definitions of a concept, whereas many lawyers and researchers from other policy-oriented disciplines would look for operational definitions that are good enough to resolve in-practice problems.

These differences in approach are in part motivated by what problems the methodologies available to different disciplines are best suited to solve, and the kinds of research that are valued within different fields. Furthermore, different strategies for concept building tend to align with different strategies for resolving ethical and societal problems. For example, conceiving such problems as purely technical in nature, where value judgements are used only in the specification of the problem, as opposed to conceiving them as political problems, which require stakeholders to negotiate and compromise.

Attempts to clarify key concepts relating to the ethical and societal challenges of ADA should take heed of these disciplinary differences and not inadvertently prioritise specific research or policy approaches by default.

3.3 Differences across cultures and publics

In addition to disciplinary differences, key concepts may be understood differently or carry different connotations across different cultures and publics.

One example that has been highlighted is the concept of privacy. Whereas modern Western ethical traditions (e.g. Kantianism) tend to conceive of individual privacy as an intrinsic good, this is often not the case in Eastern traditions. In Confucianism, which tends to emphasise the collective good over the individual, the notion of individual privacy (as opposed to the collective privacy e.g. of a family) has traditionally been given less attention (and may even carry negative connotations e.g. of shameful secrets).
In a different vein, as traditional Buddhism regards the belief in an autonomous self as a pernicious illusion, some Buddhist traditions have argued one should actively share one's secrets as a means to achieving a lack of self.[26]

Of course, it is important to avoid the assumption that everyone within a cultural tradition shares the same concept, and that the attitudes of an entire culture can be reduced to whatever is expressed in its dominant philosophical or religious traditions. As long as they are recognised as *tendencies*, exploring these differences will be important for understanding the varied connotations for different groups of the concepts used to discuss ADA-based technologies. However, there is also a need for more empirical work (e.g. surveys, interviews, anthropological studies) on conceptual variations between and within different countries and cultures.

These points are not only applicable to different cultures as defined for example by a specific national, linguistic or religious community. Key ethical and political concepts may also be more or less visible, and may have different connotations in, different intersecting groups or publics within and across cultures, such as gender, sexuality, class, ethnicity, and so on. A useful illustration is the argument by second-wave feminists encapsulated in the slogan 'the

[24] Barocas & Selbst (2016); London and Danks (2017).

[25] For example, Zafar et al. (2017); Kusner et al. (2017); Kearns et al. (2017).

[26] For more on differences between Eastern and Western conceptions of privacy, see Ess (2006). See also the IEEE's *Ethically Aligned Design*, v.2, pp. 193–216, which discusses the implications for ADA of several ethical traditions, including both secular traditions (e.g. utilitarianism, virtue ethics, deontology) and religious/cultural traditions such as Buddhism, Confucianism, African Ubuntu and Japanese Shinto.

personal is political'.[27] Second-wave feminists criticised the traditional conception of the private sphere as personal and apolitical in contrast to the political public sphere, a distinction which in Western thought traces back to ancient Greece (Burch 2012, ch. 8). Among other things, this traditional conception has often led nonmarket housework and childcare to be considered irrelevant (or simply ignored) in discussions of labour and economics (for instance, these are not measured in GDP).[28] It also resulted in failure to name certain phenomena only visible to the marginalised (such as sexual harassment).

This example illustrates how debates about the future of work in particular, and technology in general, should take into account a broad range of perspectives on what is involved and valuable in concepts such as 'labour', 'leisure' or 'spare time'. More generally, different publics within society will differ in their understanding and valuation of key concepts involved in debates about ADA. Understanding these differences, and ensuring that the values of all members of society are represented, will be key to navigating these debates.

3.4 Conceptual complexity

However, merely distinguishing different uses and interpretations is in itself unlikely to resolve these conceptual tangles. While many morally significant terms can seem intuitively clear and unproblematic, philosophical analyses often reveal deeper conceptual complexities.

Take the concept of fairness again. This is often highlighted as being the key value at stake in cases of algorithmic bias. The terms 'bias' and 'fairness' are often conflated, with some discussions of such cases simply defining bias as unfair discrimination.[29] Yet there is no uniform consensus within philosophy on an exact definition of fairness. Political philosophers have defended several different definitions, each drawing on different intuitions associated with the concept.

Some theories focus on achieving a fair distribution of outcomes between groups. Of course, we still need to say what it is that makes a distribution of outcomes fair: different subtheories argue that the fairest distributions are ones that maximise overall benefit (utilitarianism), ones that are as equal as possible (egalitarianism), or ones that benefit the worst-off the most (minimax). Other theories of fairness focus less on any particular distribution of outcomes and instead emphasise *how* those outcomes are determined: whether the benefits or disadvantages an individual receives are the result of their own free choices, or result from unlucky circumstances beyond their control such as historical injustices towards specific groups or individuals.[30]

These differences are relevant to how we think about the impact of ADA on fairness. For instance, suppose we are concerned with whether an algorithm used to make healthcare decisions is fair to all patients. On a purely egalitarian conception of fairness, we ought then to assess whether the algorithm produces equal outcomes for all users (or all relevant subgroups – at which point we have to ask which are the relevant subgroups). On a minimax conception (i.e. maximising benefits for the worst off), by contrast, we should instead ensure the algorithm results in the best outcomes for the worst off user group, even if this leads to a greater disparity between the outcomes for different groups, or produces worse results on average. Adopting a conception of fairness based on free choice would instead require us to decide which conditions are truly free choices and which are merely lucky circumstance. For example, is smoking, or obesity, a free choice? Simply stating that the algorithm should be 'fair' would fail to distinguish between these different potential meanings of the concept.[31]

Similar conceptual complexities can be found in most of the key terms framing debates around the impacts of ADA. What does it mean for an algorithmic decision-making system to have 'intelligibility' and why is this an important feature for such systems to possess? What counts as 'personal' data and why is it important to

[27] See, for example, Hanisch (2006, 1969). Similar arguments and slogans were used across a number of political movements in the 1960s and 1970s, Crenshaw (1995).

[28] GPI Atlantic (1999).

[29] Friedman and Nissenbaum (1996).

[30] Some legal systems grant special protections against discrimination or unequal treatment to groups defined by certain 'protected characteristics', such gender, ethnicity or religion. This is sometimes justified on the grounds that these groups have historically been subjected to unfair discrimination. However, what makes discrimination against such groups especially wrong is disputed. See, for example, Altman (2015).

[31] See Binns (2017) for a fuller survey of different theories of fairness and their relation to machine learning.

protect the privacy of these (as opposed to 'non-personal' data)? What counts as 'meaningful' consent?

While rich philosophical literatures exist on most of these concepts, there is relatively little work spelling out their application to how we talk and think about the ethical implications of ADA.

Although clarification is an important step to making constructive progress, doing this will not always be straightforward. Differences in the understanding of key concepts sometimes reflect deeper, substantial disagreements between groups who endorse fundamentally different values or have conflicting interests. For example, when libertarians prefer a choice-based conception of fairness while social democrats prefer a distribution-based conception, this is not merely a terminological dispute. Rather, they fundamentally disagree about what justice requires us to prioritise.

Merely analysing and highlighting these differences is unlikely to yield uncontroversial solutions to these disagreements. Navigating such disagreements will often require political solutions, rather than mere conceptual analysis. For example, by designing political processes or institutions which can be recognised as legitimate by different publics or interest groups even when they disagree with individual decisions. Clarifying and analysing the key concepts can however help distinguish cases where disputes are merely terminological, and identify where further work is needed to resolve or navigate substantial disagreements.

3.5 Summary and recommendations

Making progress in debates on ADA ethics and societal impacts requires disentangling the different meanings of key terms used to frame these debates. Three kinds of work are necessary to make progress on this task:

1. Mapping and clarifying ambiguities

The first kind of work needed is to understand the differences and ambiguities in the use of key concepts surrounding debates of ADA.

An important step towards this will be **mapping exercises** of the kind mentioned in 3.1, which seek to disentangle and classify different types of problems or cases that are currently lumped together under the same terminology. These mapping exercises will need to clarify both (a) different *possible* interpretations and uses of a given concept, such as 'transparency', and also (b) how important concepts are, in practice, used by different groups and communities. To achieve (a), in-depth **philosophical analyses** will sometimes be needed to uncover the conceptual complexities hiding under commonly used concepts. To achieve (b), this work will need to intersect with relevant **technical research**, for example work on different possible mathematical definitions of fairness, and **empirical social sciences research** to elucidate different understandings of similar concepts within different disciplines and between different cultures. We discuss such empirical work further in section 5.

Most work of this kind has centred on the conceptual clusters of bias/fairness/discrimination and transparency/explainability, intelligibility/interpretability, and to some extent privacy and responsibility/accountability. More studies of this kind would be welcome and should be extended more systematically to other important concepts in this space (including those that we discuss in section 4, namely, dignity, solidarity, citizenship, convenience and self-actualisation).

2. Bridging disciplines, sectors, publics and cultures

Analysing and bringing these complexities and divergences into focus will help to mitigate the risk of cross-talking. However, we also need constructive work aiming to *bridge* these differences, i.e. engaging relevant practitioners and stakeholders to make them aware of relevant differences and actively *enabling* communication across these divides.

In terms of crossing disciplinary divides, mapping and communicating differences in use will help identify situations where researchers or practitioners are misunderstanding each other. Actively enabling communication will furthermore require **interdisciplinary collaborations** where researchers can help each other translate their findings to different target audiences. Examples of such collaborations already taking place include papers co-authored between lawyers and technical researchers. These can be taken as a template for further collaborations. Workshops that bring together different disciplines to discuss key concepts could be another model for bridging these terminological and language differences.

Much of the current international debate around ADA ethics emanates from Western countries and is framed in terms of Western intellectual traditions.[32] Work is however ongoing in other cultural spheres, in particular East Asia, at the forefront of ADA research. An important step to integrate a fuller range of perspectives into international debates will be to translate important policy documents and research literature – both from other languages into English, and the reverse. Ensuring that major conferences and meetings have delegates from a wide range of countries and other backgrounds will also be important. Furthermore, work should be done to identify research from other countries, in particular from developing countries, whose perspectives are currently not strongly represented. This should include building collaborations with researchers and policy makers in those countries.

3. Building consensus and managing disagreements

Finally, work should be done to build consensus around the best ways to conceptualise the ethical and societal challenges raised by ADA. We should seek to find common understandings and pieces of shared conceptual machinery. These need not replace the existing frameworks within disciplines or cultures, but should be ones that different stakeholders can agree are good enough for joint constructive action. Though we may not always be able to agree on a single precise definition for every term related to the ethics of AI, we can clarify meaningful disagreements and prevent people from talking past one another.

Many of the recommendations discussed in relation to **mapping and clarifying ambiguities** and **bridging disciplines and cultures** will contribute to this. However, while traditional conceptual analysis provides an important starting point for resolving ambiguities, settling on definitions will require engaging with all stakeholders influenced by technologies, including the public. We discuss ways of involving the public in section 4.4.2.

It should be stressed that not all ethically relevant disagreements can be resolved by purely conceptual means. Some conceptual differences reflect deeper disciplinary, cultural or political disagreements. To address these, we will need to consider how to manage the tensions, trade-offs and dilemmas to which these disagreements give rise. We explore these in the next section.

[32] See appendix 1 for comments on how discussion of these issues differs in developed versus developing countries.

4. Exploring and addressing tensions

The conceptual work described in section 3 aims to build clarity and consensus around the key concepts and principles of ADA ethics. This is an important starting point, but is not enough if these principles cannot be put into practice, and it is not yet clear that the very high-level principles proposed for ADA ethics can guide action in concrete cases. In addition, applying principles to concrete cases often reveals obstacles to their implementation: they may be technically unrealisable, overly demanding, or implementing them might endanger other things we value. For instance, recent attempts to construct definitions of fairness that are sufficiently mathematically precise to be implemented in machine learning systems have highlighted that it is often mathematically impossible to optimise for different, intuitively plausible dimensions of fairness.[33] It is therefore far from clear what it would mean to ensure that AI, data, and algorithms 'operate on principles of fairness'[34] in practice.

To think clearly about ADA-based technologies and their impacts, we need to shift our focus to exploring and addressing the tensions that arise between different principles and values when trying to implement them in practice. While several of the existing discussions recognise the importance of facing these conflicts, none do so systematically. For example, the Montréal Declaration on Responsible AI (2018) states that its principles 'must be interpreted consistently to prevent any conflict that could prevent them from being applied', but it is not clear *how* one is supposed to prevent such a conflict in practice. Similarly, Cowls and Floridi (2018) recognise that using AI for social good requires 'resolving the tension between incorporating the benefits and mitigating the potential harms of AI', but do not talk about specific tensions in detail or how to resolve them.

4.1. Values and tensions

As highlighted in section 2, existing collections of principles invoke a number of values that can be at stake in applications of ADA. These express different kind of aims which either motivate the use of ADA-based technologies for various purposes, or which such technologies ought to preserve. Importantly, these aims are multiple rather than one overall goal such as utility, goodness or human flourishing.

These values are attractive ideals, but in practice they can come into conflict, meaning that prioritising one value can require sacrificing another. Developing more complex algorithms that improve our ability to make accurate predictions about important questions may reduce our ability to understand how they work, for instance. The use of data-driven technologies might also make it impossible for us to fully guarantee desirable levels of data privacy. But if the potential gains of these technologies are significant enough – new and highly effective cancer treatments, say – communities might decide a somewhat higher risk of privacy breaches is a price worth paying.

We use the umbrella term 'tension' to refer to different ways in which values can be in conflict, some more fundamentally than others (as elaborated in section 4.4.1). Note that when we talk about tensions between values, we mean tensions between the pursuit of different values in technological applications rather than an abstract tension between the values themselves. The goals of efficiency and privacy are not fundamentally in conflict across all scenarios, for example, but do come into conflict in the context of certain data-driven technologies. Given the right contextual factors, ADA-based technologies might create tensions between any two (or more) of these values – or even simultaneously threaten and enhance the same value in different ways.

Some of these tensions are more visible and of higher priority than others. The table below highlights some key tensions between values that arise from current applications of ADA-based technologies:

[33] See Friedler et al. (2016); Kleinberg et al. (2016); Chouldechova (2018); Binns (2017).

[34] As per principle 2 of the Lords' Select Committee on AI's proposed 'cross-sector AI code'.

EXAMPLES OF TENSIONS BETWEEN VALUES

Quality of services *versus* privacy: using personal data may improve public services by tailoring them based on personal characteristics or demographics, but compromise personal privacy because of high data demands.

Personalisation *versus* solidarity: increasing personalisation of services and information may bring economic and individual benefits, but risks creating or furthering divisions and undermining community solidarity.

Convenience *versus* dignity: increasing automation and quantification could make lives more convenient, but risks undermining those unquantifiable values and skills that constitute human dignity and individuality.

Privacy *versus* transparency: the need to respect privacy or intellectual property may make it difficult to provide fully satisfying information about an algorithm or the data on which it was trained.

Accuracy *versus* explainability: the most accurate algorithms may be based on complex methods (such as deep learning), the internal logic of which its developers or users do not fully understand.

Accuracy *versus* fairness: an algorithm which is most accurate on average may systematically discriminate against a specific minority.

Satisfaction of preferences *versus* equality: automation and AI could invigorate industries and spearhead new technologies, but also exacerbate exclusion and poverty.

Efficiency *versus* safety and sustainability: pursuing technological progress as quickly as possible may not leave enough time to ensure that developments are safe, robust and reliable.

Given the wide scope of possible applications of ADA-based technologies, and the variety of values that may be impacted (positively or negatively) by these applications, there is unlikely to be any simple, exhaustive list of all possible tensions arising from ADA in all contexts. It would go beyond the scope of this report to attempt to systematically map all of these. We have therefore limited our discussion to four tensions that are central to current debates, as summarised in table 1.

The tensions in the first two rows reflect how goods offered by ADA technologies may come into conflict with the societal ideals of fairness and solidarity – we therefore refer to these tensions as societal:

1. Using algorithms to make decisions and predictions more accurate **versus** ensuring fair and equal treatment.

2. Reaping the benefits of increased personalisation in the digital sphere **versus** enhancing solidarity and citizenship.

This first societal tension, between accuracy and fairness, has been widely discussed in controversies and case studies involving ADA-based technologies in recent years. The second tension, between personalisation and solidarity, has received less explicit attention – but we believe it is also fundamental to ethical concerns surrounding the application of ADA-based technologies in society.

The next two rows concern ideals of individual life, so we refer to them as individual tensions:

3. Using data to improve the quality and efficiency of services **versus** respecting the privacy and informational autonomy of individuals.

4. Using automation to make people's lives more convenient **versus** promoting self-actualisation and dignity.

Again, we highlight one tension that has already been widely recognised, between the quality and efficiency of services and the informational autonomy of individuals, and one that has been discussed less, between the convenience offered by automation on the one hand and the threat to self-actualisation on the other.

All four tensions, however, share the following crucial similarities: they arise across a wide range of sectors and they touch upon the deepest ethical and political ideals of modernity. Between them, they cover a broad spectrum of issues where further research is likely to be valuable for managing the impacts of current and foreseeable applications of ADA.

4.2 Unpacking four central tensions

Tension 1: Using algorithms to make decisions and predictions more accurate **versus** ensuring fair and equal treatment.

TABLE 1. KEY TENSIONS ARISING BETWEEN THE GOODS OFFERED BY ADA TECHNOLOGIES AND IMPORTANT SOCIETAL AND INDIVIDUAL VALUES

Goods offered by ADA technologies	**Core values in tension with those goods**	
Accuracy	Fairness	Societal values
Personalisation	Solidarity	
Quality & efficiency	Informational autonomy	Individual values
Convenience	Self-actualisation	

This tension arises when various public or private bodies base decisions on predictions about future behaviour of individuals (e.g. when probation officers estimate risk of reoffending, or school boards evaluate teachers),[35] and when they employ ADA-based technologies to improve their predictions. The use of blunt quantitative tools for evaluating something as complex as human behaviour or quality of teaching can be misguided, as these algorithms can only pick out easily measurable proxies.[36] Nonetheless, these algorithms can sometimes be more accurate on some measures than alternatives, especially as systematic bias afflicts judgments made by humans too. This raises questions of whether and when it is fair to make decisions affecting an individual's life based on an algorithm that inevitably makes generalisations, which may be missing important information and which, in addition to this, can systematically disadvantage some groups over others. An additional way in which algorithms can undermine fairness and equality is that it is often difficult to explain why they work – either because they are based on 'black box' methods, or because they use proprietary software – thus taking away individuals' ability to challenge these life-altering decisions.

Hypothetical illustration: to assist in decisions about whether to release defendants on bail or to grant parole, a jurisdiction adopts an algorithm that estimates the 'recidivism risk' of criminal defendants, i.e. their likelihood of re-offending. Although it is highly accurate on average, it systematically discriminates against black defendants, because the 'false positives' – the rate of individuals classed as high risk who did not go on to reoffend – is almost twice as high for black as for white defendants.[37] Since the inner workings of the algorithm is a trade secret of the company that produced it (and in any case is too complex for any individual to understand), the defendants have little to no recourse to challenging the verdict that have huge consequences on their lives.

Tension 2: Reaping the benefits of increased personalisation in the digital sphere **versus** enhancing solidarity and citizenship.

Companies and governments can now use people's personal data to draw inferences about their characteristics or preferences, which can then be used to tailor the messages, options and services they see. This personalisation is the end of crude 'one size fits all'

[35] Angwin, J., et al. (2016).

[36] For more on this topic see the work of, for example, Cathy O'Neil: www.bloomberg.com/view/articles/2018–06–27/here-s-how-not-to-improve-public-schools

[37] Angwin, J., et al. (2016).

solutions and enables individuals to find the right products and services for them, with large potential gains for health and well-being. However, this risks threatening the guiding ideals of democracy and the welfare state, namely citizenship and solidarity.[38] These ideals invite us to think of ourselves as citizens and not just individual consumers, and to provide for each other in the face of unexpected blows of fate beyond individual control. Public commitments that certain goods should be ensured for citizens irrespective of their ability to pay (education, healthcare, security, housing, basic sustenance, public information) depend on there being a genuine uncertainty about which ones of us will fall ill, lose employment, or suffer in other ways. This uncertainty underpins commitments to risk-pooling and without it there is an increased tension between promoting individual benefit and collective goods.

Hypothetical illustration: a company markets a new personalised insurance scheme, using an algorithm trained on rich datasets that can differentiate between people in ways that are so fine-grained as to forecast effectively their future medical, educational, and care needs. The company is thus able to offer fully individualised treatment, better suited to personal needs and preferences. The success of this scheme leads to the weakening of publicly funded services because the advantaged individuals no longer see reasons to support the ones with greater needs.

Tension 3: Using data to improve the quality and efficiency of services **versus** respecting the privacy and informational autonomy of individuals.

This tension arises when machine learning and big data are used to improve a range of different services: public ones such as healthcare, education, social care, and policing, or any service offered privately. These technologies could enable service providers to tailor services exactly to customers' needs, improving both quality of services as well as efficient use of taxpayers' money. However, the heavy demand on individuals' personal data raises concerns about loss of privacy and autonomy of individuals over their information (we shall use the term 'informational autonomy' to denote this value).[39]

Hypothetical illustration: a cash-strapped public hospital gives a private company access to patient data (scans, behaviours, and medical history) in exchange for implementing a machine learning algorithm that vastly improves doctors' ability to diagnose dangerous conditions quickly and safely. The algorithm will only be successful if the data is plentiful and transferable, which makes it hard to predict how the data will be used in advance, and hard to guarantee privacy and to ensure meaningful consent for patients.

Tension 4: Using automation to make people's lives more convenient **versus** promoting self-actualisation and dignity.

Many ADA-based technologies are currently developed by private commercial entities working to disrupt existing practices and replace them with more efficient solutions convenient to as many customers as possible. These solutions may genuinely improve people's lives by saving them time on mundane tasks that could be better spent on more rewarding activities, and by empowering those previously excluded from many activities. But automated solutions also risk disrupting an important part of what makes us human.[40] Literature and arts have long explored anxieties about humans relying on technology so much that they lose their creative, intellectual, and emotional capacities.[41] These capacities are essential to individuals' ability to realise their life plans autonomously and thoughtfully – an ideal that is often referred to as self-actualisation and dignity. The fast rise of ever more effective and comprehensive AI systems makes the possibility of human decline and obsolescence – and associated fears of deskilling, atrophy, homogenisation, and loss of cultural diversity – more vivid and realistic. These fears also arise in displacement of human labour and employment by AI and robots because, in addition to livelihood, work is a source of meaning and identity.

[38] See Prainsack and Buyx (2017) on personalisation and solidarity in healthcare and biomedicine.

[39] A recent example of this tension is the case of DeepMind and the Royal Free hospital: www.theguardian.com/technology/2017/jul/03/google-deepmind-16m-patient-royal-free-deal-data-protection-act

[40] Turkle (2016 and 2017) explores these trends in depth.

[41] E.M. Forster's dystopian short story 'The Machine Stops' (1909, see Forster 1947) and the 2008 animated film Wall-E illustrate this concern.

Hypothetical illustration: AI makes possible an all-purpose automated personal assistant that can translate between languages, find the answer to any scientific question in moments, and produce artwork or literature for the users' pleasure, among other things. Its users gain unprecedented access to the fruits of human civilization but they no longer need to acquire and refine these skills through regular practice and experimentation. These practices progressively become homogenised and ossified and their past diversity is now represented by a set menu of options ranked by convenience and popularity.

4.3 Identifying further tensions

The four tensions outlined above are central to thinking about the ethical and societal implications of ADA broadly and as they stand today. However, other tensions can and should be identified, particularly when focusing more narrowly on specific aspects of ADA ethics, and as the impacts of technology on society change over time.

Our approach to identifying tensions begins with a list of important values and principles we want our use of ADA-based technologies to respect. We then consider what obstacles might arise to realising these values in practice, and ways that using technology to enhance or promote one value might undermine another.

This approach could usefully be extended or repeated by others to identify additional tensions to those outlined above, as different perspectives will inevitably unearth slightly different tensions. Appendix 3 presents some different ways of carving up issues, publics, and sectors, which could be used to help identify a variety of tensions. Repeating this process over time will also be important, as the ways that technology may be used to enhance or threaten key values changes, and even the very values we prioritise as a society may change.

Thinking about tensions could also be enhanced by systematically considering different *ways* that tensions are likely to arise. We outline some conceptual lenses that serve this purpose:

- **Winners versus losers.** Tensions sometimes arise because the costs and benefits of ADA-based technologies are unequally distributed across different groups and communities.
  - A technology which benefits the majority may systematically discriminate against a minority: predictive algorithms in a healthcare setting may improve outcomes overall, but worsen outcomes for a minority group for whom representative data is not easily accessible, for example.
  - Automation may enrich the lives of the most privileged, liberating them to do more worthwhile pursuits, while wiping out the livelihood of those whose labour is replaced and who do not have other options. In addition to shifts in distribution of material resources, prestige, power, and political influence are also affected.
- **Short term versus long term.** Tensions can arise because values or opportunities that can be enhanced by ADA-based technologies in the short term may compromise other values in the long term. For example:
  - Technology which makes our lives better and more convenient in the short term could have hard-to-predict impacts on societal values in the long term: as outlined above, for example, increasing personalisation could make our lives easier and more convenient, but might undermine autonomy, equality and solidarity in the long term.
  - Speeding up innovation could create greater benefits for those alive today, while introducing greater risks in the long term: there is a trade-off between getting the benefits of AI as quickly as possible, and taking extreme caution with the safety and robustness of advanced systems.
- **Local versus global.** Tensions may arise when applications that are defensible from a narrow or individualistic view produce negative externalities, exacerbating existing collective action problems or creating new ones. For example:
  - Technology that is optimised to meet individual needs might create unforeseen risks on a collective level: a healthcare algorithm might recommend against vaccination for individuals, which could have huge negative impacts on global health.

4.4 Resolving the tensions

So far we have been using the single term 'tension' to denote what is in fact several different kinds of conflicts between values: some fundamental, others merely practical. Because these differences matter to how these tensions should be resolved, we spell them out here before discussing the solutions.

4.4.1 Kinds of tensions

The quintessential ethical conflict is a **true dilemma**. A true dilemma is a conflict between two or more duties, obligations, or values, both of which an agent would ordinarily have reason to pursue but cannot. These are instances when no genuine resolution is possible because the very acts that further one value (say, Antigone's duty to bury her dead brother) take away from the other value (her duty to obey the king). We call these true dilemmas because the conflict is inherent in the very nature of the values in question and hence cannot be avoided by clever practical solutions. Sometimes the tensions we discussed above will take the form of such a dilemma in which it is genuinely impossible to, say, implement a new automating technology without devaluing and undermining certain human skills and capacities. In true dilemmas, a choice has to be made to prioritise one set of values, say speed, efficiency and convenience, over another, say achievement or privacy.

However, sometimes what appears like a tough choice necessitating the sacrifice of important values is not one in reality. Claims of dilemmas can be exaggerated or go unexamined, such as when a company claims that privacy needs to be sacrificed without properly studying how their goals might be achieved without this sacrifice. In many cases the tension we face is a **dilemma in practice**, where the tension exists not inherently, but due to our current technological capabilities and constraints, including the time and resources we have available for finding a solution. The tension between transparency and accuracy is a useful illustration. These two ideals are not fundamentally in conflict with one another (in the same way that some of the conflicting definitions of fairness are, for example.) The conflict here is a more practical one: generally, producing the most accurate algorithm possible will tend to result in models that are more complex and therefore more difficult to make fully intelligible to humans. However, it is an open empirical question to what extent we must be forced to make a trade-off between these two ideals, and methods are beginning to be developed which increase transparency without compromising accuracy.[42]

This in turn highlights that some apparent tensions may in fact be **false dilemmas**. These are situations where there exists a third set of options beyond having to choose between two important values. We can commit more time and resources to developing a solution which avoids having to sacrifice either value, or to delay implementing a new technology until further research makes available better technologies. False dilemmas can arise when we fail to recognise either the extent to which our current technological capabilities are in fact able to resolve a tension, or there are no overriding constraints that force us to implement a given technology immediately.

The best approach to resolving a tension will depend on the nature of the tension in question.

4.4.2 Trade-offs and true dilemmas

To the extent that we face a true dilemma between two values, any solution will require making **trade-offs** between those values: choosing to prioritise one value at the expense of another. For example, if we determined that each of the tensions presented above could not be dissolved by practical means, we would need to consider trade-offs such as the following:

- Trade-off 1: Judging when it is acceptable to use an algorithm that performs worse for a specific subgroup, if that algorithm is more accurate on average across a population.
- Trade-off 2: Judging how much we should restrict personalisation of advertising and public services for the sake of preserving ideals of citizenship and solidarity.
- Trade-off 3: Judging what risks to privacy it is acceptable to incur for the sake of better disease screening or greater public health.
- Trade-off 4: Judging what kinds of skills should always remain in human hands, and therefore where to reject innovative automation technologies.

The difficult question is how such trade-off judgments should be made. In business and economics, solutions to trade-offs are traditionally derived using cost-benefit analysis (CBA): where all the costs and benefits of a

[42] See, for example, Adel et al. (2018).

given policy are converted to units on the same scale (be it monetary or some other utility scale such as well-being) and a recommendation is made on the basis of whether the benefits outweigh the costs. These methods are used almost universally in governance, industry, and commerce because they provide clear procedures and appear objective. It will be tempting to transfer these same methods to the dilemmas above, demanding data on how much value all stakeholders put on each of the ideals involved in any given dilemma and crunching the numbers thereafter.

We caution against this. Cost-benefit analysis can be *part* of the process of exploring trade-offs. The process is transparent and mechanical and can generate useful data to input into decision-making. But CBA alone should not be seen as the answer: it is technocratic, it does not recognise the fact that values are vague and unquantifiable and that numbers themselves can hide controversial value judgments, and finally, the very act of economic valuation of a good can change people's attitude to it (this explains why applying CBA to environmental or other complex and public goods attracts so much controversy).[43]

Resolution of these dilemmas can take a variety of forms depending on precise political arrangements. But one approach we wish to highlight (and one that is also relevant to the cases discussed in section 3), is that legitimacy of any emerging solution can be achieved through consultation and inclusive public deliberation. Methods for implementing such deliberations, where small groups of citizens are guided through controversies by experts and moderators, are emerging in political science and in environmental and medical research where public participation matters.[44] In case of ADA-based technologies, such consultations are not yet well established but they are much needed.[45] Their goals should be as follows:

1. To give voice to all stakeholders and to articulate their interests with rigour and respect (data about potential costs and benefits of technologies can be useful for this).
2. To identify acceptable and legitimate trade-offs that are compatible with rights and entitlements of those affected by these technologies.
3. To arrive at resolutions that, even when imperfect, are at least publicly defensible.

Faced with tragic choices between different ideals of virtue and good life, such an approach accepts that human judgment, protest, contestation, and consensus-building are all unavoidable and no technocratic process can replace it.[46] We talk more about the process of public deliberation in section 5.2.

4.4.3 Dilemmas in practice

On the other hand, to the extent that we face a dilemma in practice, we lack the knowledge or tools to advance the conflicting values without sacrificing one or the other. In this case, trade-offs may or may not be inevitable, depending on how quickly and with what resources we need to implement a policy or a technology. Data-driven methods for improving the efficiency of public services *and* securing high levels of informational privacy may be possible in principle, for example, but not available at the moment. For each of the four tensions highlighted, it is possible that with more knowledge or better methods the tension would dissolve or at least be alleviated.

In these situations we face a choice:

- To put the technology to use in its current state. In this case, we will need to determine and implement some legitimate trade-off that sacrifices one value for another. This will involve the same kind of work as described for true dilemmas in section 4.4.2.
- To hold off implementing this technology and instead invest in research on how to make it serve all the values we endorse equally and maximally.

This choice can be thought of as involving its own tension, of the short-term versus long-term kind discussed in

[43] For controversies surrounding cost benefit analysis see Frank (2000), Alexandrova and Haybron (2011), Haybron and Alexandrova (2013). For complexities of identifying and measuring what counts as benefit and well-being see Alexandrova (2017).

[44] Stanford's Centre for Deliberative Democracy is a pioneer http://cdd.stanford.edu/what-is-deliberative-polling/

[45] Though some work on consultation relating to ADA has begun, led particularly by the RSA and Royal Society, as we will discuss in section 5.2.

[46] See Moore (2017) and Alexandrova (2018) among others on the crucial role of consultation and contestation of expertise in democracies.

section 4.3: to what extent should we postpone the benefits of new technologies and instead invest the time and resources necessary to better resolve these tensions? This is not a binary choice, of course. Rather, we might choose to strike a balance: try to navigate the trade-offs required to make decisions about how to use technology today, while investing in research to explore whether the tension might be fully resolvable in future.

4.4.4 Better understanding tensions

As this discussion highlights, in order to make progress on the tensions arising in relation to ADA-based technologies, it is crucial to be clear about the nature of the tensions – do they involve true dilemmas, practical dilemma or even false dilemmas?[47]

In addition to developing methods for balancing trade-offs and investing in better technology, we should therefore also invest in research to better understand the nature of important tensions. We can explore this by asking the following questions:

- **Can the most accurate predictive algorithms be used in a way that respects fairness and equality?** Where specific predictive algorithms are currently used (e.g. in healthcare, crime, employment), to what extent do they discriminate against or disadvantage specific minorities?

- **Can the benefits of personalisation be reaped without undermining citizenship and solidarity?** In what specific ways might different forms of personalisation undermine these important ideals in future? How can this be addressed or prevented?

- **Can personal data be used to improve the quality and efficiency of public services without compromising informational autonomy?** To what extent do current methods allow the use of personal data in aggregate for overall social benefits, while protecting the privacy of individuals' data?

- **Can automation make lives more convenient without threatening self-actualisation?** Can we draw a clear line between contexts where automation will be beneficial or minimally harmful and tasks or abilities should not be automated?

Answering these questions will in part involve conceptual research of the kind discussed in section 3. For instance, clarifying what kind of algorithmic 'fairness' is most important is an important first step towards deciding whether this is achievable by technical means. In addition, since these are largely empirical questions about what is in fact possible, answering them will often require drawing on evidence about what is technically feasible, as described in detail in the next section. In some cases, current or potential technology may be able to resolve or lessen some of the tensions.

Finally, these tensions will only be resolved in practice if there are sound and appropriate institutions, laws, and governance structures to undergird and implement these efforts. Standards, regulations, and systems of oversight concerning the ADA technologies are currently in flux and much uncertainty surrounds their future.[48] We urge that new approaches to governance and regulation be duly sensitive to the tensions described above and devise legitimate institutions that will help communities to navigate whatever tensions arise and at whatever levels.

4.5 Summary and recommendations

This section introduces the idea and the importance of thinking about **tensions** between values that different principles of ADA ethics embody, in order to ensure that these principles can be action-guiding in concrete cases. The following high-level recommendation follows immediately:

- Move the focus of ADA ethics towards identifying the tensions arising from implementation of ethical practice involving ADA.

The four tensions we propose as priorities in section 4.2 encompass controversies and case studies that commentators across different fields and sectors are beginning to explore. Hence, our next set of recommendations are to:

- Investigate instances of the **four tensions** highlighted in this report across different sectors of society, exploring specific cases where these tensions arise:

[47] Concrete cases may well involve a combination of all three kinds of dilemmas, once we distinguish at a more fine-grained level between the different values held by different stakeholders in a given case.

[48] See Wachter and Mittelstadt (2018) for the uncertainty surrounding the implications and implementation of the GDPR, for example.

- Using algorithms to make decisions and predictions more accurate **versus** ensuring fair and equal treatment.

- Reaping the benefits of increased personalisation in the digital sphere **versus** enhancing solidarity and citizenship.

- Using data to improve the quality and efficiency of services **versus** respecting the privacy and informational autonomy of individuals.

- Using automation to make people's lives more convenient **versus** promoting self-actualisation and dignity.

- Identify **further tensions** based on other value conflicts and their underlying causes using the following questions:

  - Where might the costs and benefits of ADA-based technologies be distributed unequally across groups?

  - Where might short-term benefits come at the cost of longer-term values?

  - Where might ADA-based technologies benefit the individual or groups but raise problems at a collective level?

Articulating the tensions that apply in a given case is the first step in implementing ethical technologies, but the next step should be towards resolving these conflicts. How we do so depends on the nature of any given tension. We therefore recommend that further research should aim to:

- Identify the extent to which key tensions involve **true dilemmas**, **dilemmas in practice** or **false dilemmas**. Often this will involve investigating specific instances of the tension, and considering ways to resolve it without sacrificing either of the key values.

- Where we face dilemmas in practice, conduct research into **how these dilemmas might be dissolved**, for example by advancing the frontiers of what is technically possible such that we can get more of both the values we care about.

- Where we face true dilemmas between values, or practical dilemmas that we are forced to act on now, conduct research into **dilemma resolution through legitimation of trade-offs** in public deliberations and regulatory institutions adapted specially to ADA technologies.

5. Developing an evidence base

Current discussions of the ethical and societal implications of ADA suffer from gaps in our understanding: of what is technologically possible, of how different technologies will impact society, and what different parts of society want and need. To make progress in using ADA-based technologies for the good of society, we need to build a stronger evidence base in all of these areas. Building this stronger evidence base will be particularly important for those developing practical frameworks and guidelines for AI ethics, including government bodies, legislators, and standards-setting bodies.

For example, the tension between using data to improve public services and the need to protect personal privacy is difficult in part because discussions of this topic are lacking good evidence on the following:

- How much machine learning and 'big data' could improve public services – and to what extent and in what ways personal privacy might be compromised by doing so.

- To what extent different publics value better healthcare relative to data privacy, and in what contexts they are happy for their data to be used.

- What the longer-run consequences of increased use of personal data by authorities might be.

Ensuring that algorithms, data and AI are used to benefit society is not a one-off task but an ongoing process. This means that, as well as understanding technological capabilities and societal needs as they stand *today*, we also need to think about how these things might evolve in the future so that we can develop adaptive strategies that take future uncertainty into account.

This section outlines some of the general areas of research that will be needed to develop a stronger evidence base, and highlights some priority questions based on the tensions discussed in section 4. Our focus is on what kinds of questions need answering and the general directions of research. While we highlight some promising methods, these are not meant to be exhaustive. We have not attempted to survey all existing or possible methods for studying these questions and for some questions, new and innovative research strategies may be needed. In general, a plurality of disciplinary perspectives and innovative methodological thinking is likely to provide the best possible evidence base.

5.1 Understanding technological capabilities and impacts

5.1.1 Technological capabilities – what is possible?

Understanding technological capabilities is a vital foundation for understanding what the real risks and opportunities of different technologies are. For example, in order to assess where data-based targeting may pose the greatest opportunities, and what risks it might introduce, we need to understand what technical steps are involved in collecting and using personalised data to target an intervention, and the limitations of existing approaches.[49] In order to assess the threat of technological unemployment and design effective policies to tackle it, we need to understand on what kinds of tasks machines are currently able to outperform humans, and the ways we might expect this to change over coming years.

Understanding technological capabilities helps us to think more clearly about the ethical tensions described in section 4 in several ways: by showing whether these tensions are true dilemmas or dilemmas in practice, by helping us to estimate the specific costs and benefits of a technology in a given context, and by giving grounds for plausible trade-offs between values that a technology promotes or threatens. This kind of evidence will be crucial for policymakers and regulators working on the governance of AI-based technologies, as well as helping researchers to identify gaps and priorities for future research.

We also need research focused on forecasting future capabilities, not just measuring existing ones, so we can anticipate and adapt to new challenges.

For the four central tensions, some key questions that will need to be answered include:

- Accuracy **versus** fair and equal treatment
  - To what degree does accuracy trade off against different definitions of fairness?

[49] It is not clear that many of the claims about Cambridge Analytica's use of 'psychographic microtargeting' stand up to rigorous technical scrutiny, for example: see Resnick (2018).

- What forms of interpretability are desirable and can be ensured in state-of-the-art models?
- To what extent is it possible to ensure adequate interpretability without sacrificing accuracy (or other values, such as privacy)?

- Personalisation **versus** solidarity and citizenship
  - Are there any in-principle or in-practice limits to how fine-grained personalisation can become (using current or foreseeable technology)?
  - To what extent is personalisation able to affect relevant outcomes in a meaningful way (e.g. user satisfaction, consumer behaviour, voting patterns)?
- Quality and efficiency of services **versus** privacy and informational autonomy
  - By how much could machine learning and 'big data' improve different public services? Can potential gains be quantified?
  - What are the best current methods for securing data privacy, and what are the technical constraints?
- Convenience **versus** self-actualisation and dignity
  - What types of tasks can feasibly be automated using current or foreseeable technologies?
  - What would the costs (e.g. energy and infrastructure requirements) be for widespread automation of a given task?
- In addition, there are **overarching questions** to be investigated, which touch upon all four of these tensions and could be applied to others:
  - What do we need to understand about technological capabilities and limitations in order to assess the risks and opportunities they pose in different ethical and societal contexts?
  - How might advances in technological capabilities help resolve tensions between values in applications of ADA, and what are the limitations of technology to do so?

The questions are phrased at a generic level. To help resolve tensions in practice, such questions will need to be tailored to the specific problem domain, as illustrated in the following scenario:

Hypothetical scenario: Imagine the Department for Health and Social Care (DHSC) is developing guidelines on the level of interpretability that should be required for algorithms to be used in different healthcare applications, and how to balance this against potential costs to accuracy. To do this well they need to understand both:

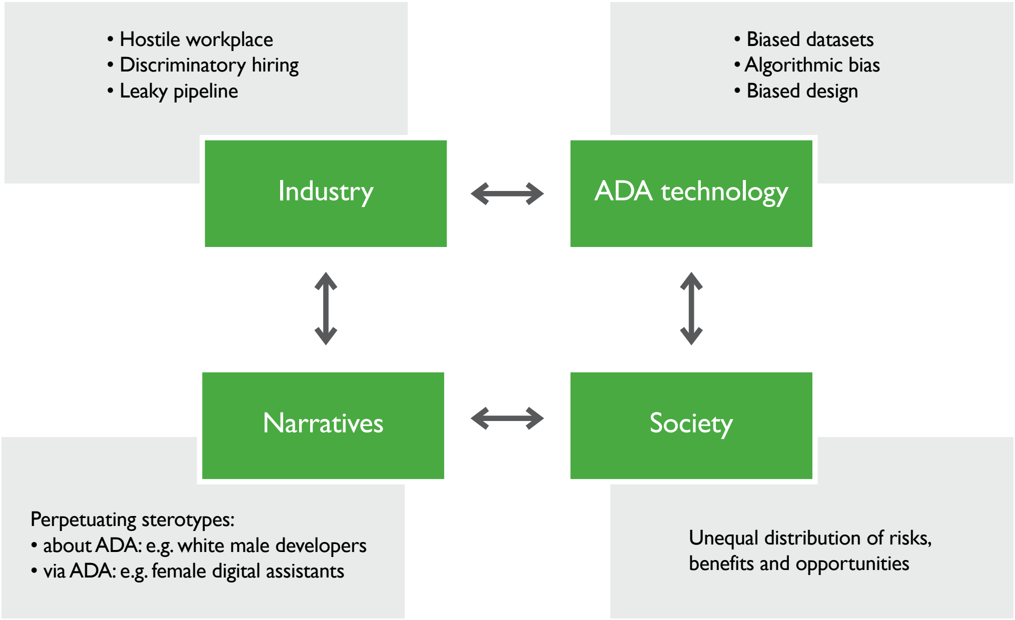
- What the options are for a given application. What different models could be used to analyse radiological imaging, for example, and to what extent and in what ways is each interpretable, and at what cost to accuracy?
- The various costs and benefits in a given context. In some cases, a drop in accuracy might be much more costly than in others, for example if incorrect diagnoses could threaten lives. And the importance of different forms of interpretability will also vary by situation (depending on whether there are other ways to test the reliability of an algorithm, or whether decisions frequently need to be explained to patients, for example).

Without understanding these technical details, the DHSC risks producing highly general guidelines that are at best difficult or impossible to implement in practice, and at worst harmful (advising never using a model that cannot be fully explained might prevent some clearly beneficial applications, for example).

These questions in part concern the technology itself, and so involve what is possible from the perspective of computer science, machine learning and other technical fields of research. Many of these are rapidly advancing fields, making it critical to stay aware of current technical constraints and developments. One way to collect evidence on technological capabilities would be to talk to or survey experts in relevant domains.[50] As a single researcher's opinions on what is 'state-of-the-art' might not be representative, surveying a wide range of technical experts is preferable to just asking one or two for their opinion. A key challenge here will be to describe the technical state of research with sufficient accuracy and detail in ways that are useful and comprehensible to non-technical people working in ethics and policy.

[50] See, for example, Grace et al. (2018).

Figure 3. An illustration of how cycles and injustice can be reinforced in how technology is developed, applied, and understood by members of society.



© AI Narratives and Justice Research Programme, Leverhulme Centre for the Future of Intelligence, 2018

However, these questions go beyond the technology itself; they also involve the effects and impacts these technologies have on humans. To answer them fully will also require research of a more psychological or sociological nature. The field of human-computer interaction studies many questions regarding the impacts of ADA-based technology on humans, often using psychology and social science methodologies.

Finally, some of these questions ask not just about current capabilities of technology, but also how these could evolve in future. Excellent work on measuring and forecasting technological capabilities already exists.[51] However, research on the ethical and societal challenges of those technologies could do much more to draw on and build on this work, to help ensure that our understanding of these broader challenges starts from rigorous thinking about what is – and what could be – technically possible.

Research into technological capabilities and impacts will therefore likely require collaborations between experts from technical ADA research, psychology/social science, forecasting, policy and ethics, as well as people able to translate between these different fields.

5.1.2 Current uses and impacts – what is happening?

In addition to understanding what is technologically possible, there is also a need to better understand: (1) how different technologies are being used, and what kinds of impacts these are in fact having, and (2) what kinds of causes, mechanisms or other influences underlie these impacts.

Regarding (1), at the moment, many debates about the ethics of ADA are motivated either by case studies, such as those uncovered by investigative journalists and social commentators, or hypothetical scenarios about how technologies might be used. While these are crucial to highlighting the potential ethical and societal impacts of ADA technologies, it is unclear to what extent these are representative of current or future developments. There is a risk of over-estimating the frequency of some applications and impacts, while missing others.

One important type of research would be to map and quantify how different ADA technologies are used on a sector-by-sector basis, looking at how they are used in finance, energy, health care, etc.[52] Another would be to identify the extent to which the kinds of positive or negative impacts often discussed actually occur in practice in different sectors. A potential challenge that will need

[51] See, for example, the AI Index, and the Electronic Frontier Foundation's work on AI Progress Measurement.

[52] See appendix 3 for further ways to break down the space of ethical and societal impacts of ADA.

to be addressed is the extent to which private or public entities are willing to disclose this information.

Regarding (2), understanding how potential impacts come about is crucial to determining the kinds of interventions that can best mitigate them, as explained in the case study below.

CASE STUDY

Cycles of injustice: Race and gender

Tackling algorithmic bias and discrimination requires better understanding of how they fit into a broader cycle of injustice, in which different problems reinforce each other, as illustrated in figure 3. For instance, discriminatory or biased outputs from the AI industry are caused both by a lack of diversity among researchers and developers, and by pre-existing social biases that are reflected in many data-sets (e.g. gender-stereotypical correlations between words in linguistic corpora). The deployment of these biased systems leads to the exacerbation of existing social injustices (e.g. systems advising on which prisoners get parole that use racially biased historical data and result in people of colour staying in prison longer).

These injustices affect who is able to shape the narratives surrounding the technology, which in turn impacts on both who is able to enter the industry and the original social injustices. For example, creators of AI are invariably portrayed as men, potentially affecting both whether women are motivated to apply and whether they are hired, and digital assistants are invariably framed as female, perpetuating the view that women are subservient. Understanding these interrelations is key to determining how best to address the resulting problems.

To better understand the ways in which different technologies are being used, their impacts on society, and the mechanisms underlying these impacts, we can ask the following questions in relation to our four central tensions:

Accuracy **versus** fair and equal treatment

- In what sectors and applications are ADA being used to inform decisions with implications for people's lives?
- Is it possible to determine how often these result in differential treatment of different socially salient groups?
- How easy are these algorithms to interpret, and what recourse do individuals have for challenging decisions?

Personalisation **versus** solidarity and citizenship

- What kinds of messages, interventions and services are already being personalised using machine learning, and in what sectors?
- How 'fine-grained' is this personalisation, and on what kinds of categories is it based?
- What evidence is there that this personalisation can substantially affect attitudes or behaviour?

Quality and efficiency of services **versus** privacy and informational autonomy

- In what sectors and applications are ADA being used to improve the efficiency of public services?
- What impacts are these specific applications having on autonomy and privacy?

Convenience **versus** self-actualisation and dignity

- What tasks and jobs have been automated in recent years, and what might we expect to be automated in the near future?
- What effects is automation already having on people's daily lives?

In addition, there are overarching questions to be investigated, which touch upon all four of these tensions and could be applied to others:

- Across different sectors (energy, health, law, etc.), what kinds of ADA-based technologies are already being used, and to what extent?
- What are the societal impacts of these specific applications, in particular on those that might be disadvantaged, or underrepresented in relevant sectors (such as women and people of colour) or vulnerable (such as children or older people)?

5.2 Understanding the needs and values of affected communities

In order to make progress on the ethical and societal implications of ADA technologies, it is necessary to understand the perspectives of those who are or will be affected by those technologies. In particular, negotiating trade-offs between values can only happen when these values, and the related hopes and concerns, of everyone who is going to be impacted by these technologies are identified and considered. Identifying these perspectives requires consultation with these end users, or at least demographically representative groups of members of different publics.[53]

It must be noted that fostering public understanding of technology alone is far from sufficient. Indeed, some science communication experts argue that it often does not matter whether non-scientists know very little about science:[54] A full understanding of how the technology works is not necessary for end users to understand its impact on their lives. Public engagement, which includes public deliberation, polling, and dialogues, is much more important: that is, fostering mutual understanding between researchers, developers, policymakers, and end users. It involves mutual interaction between these groups aimed at understanding not only the science and technology, but also its societal impacts, limits, trade-offs, and pitfalls.

For present purposes, public engagement is crucial for resolving trade-offs and dilemmas in a way that is defensible to all members of society, *especially* for trade-offs that arise because there are conflicts between the interests of different groups. On any given issue, citizens will rarely all share the same values and perspectives. However, there is evidence that when groups are able to reflect on and articulate what they care about, it is possible to reduce conflict and reach compromise.[55] It is important to note, however, that while understanding relevant public values is important to resolving trade-offs, it is not in itself the solution, but only one part of a more complex political process.[56]

There is a wide range of methods available to foster such engagement.[57] These methods can be deployed to elicit a range of views, from uninformed to informed. While uninformed polling aims to gather opinions that the surveyed groups currently hold, informed views can be elicited through engagement strategies that aim first to increase the knowledge base of the surveyed groups before investigating their informed opinions.

Public engagement that aims to resolve trade-offs can take the following forms:

- **Quantitative surveys**. Such surveys are frequently employed for 'understanding public understanding', i.e. to understand how much the surveyed groups already know about a topic, and how this informs their opinions and attitudes towards a technology.

- **Collaborative online consultation**. One example is the recent consultation put out by the UK's Centre for Data Ethics and Innovation.[58] Using trade-offs and conjoint analysis, possibly in gamified form, this could capture the views of many thousands of citizens and obtain wide-ranging top-of-mind views on how AI might play out in society and how people respond to different ethical decisions.

- **Qualitative surveys and interviews**. When used to complement quantitative work, this application of qualitative methods is particularly useful for exploring people's existing motivations and the meanings they attach to their interactions with a technology. These methods can also be deployed in combination with an educational element, in order to gather informed views.

- **Public dialogue with scenario planning**. This typically involves the input of a group of experts, including forecasters and technical analysts, who systematically map out the key uncertainties within a specified time frame. The task for the public then becomes easier – rather than having to engage in the abstract with the risks and benefits of different aspects of complex technologies, they simply have to react to different extrapolated outcomes and talk about how individuals and society would fare in different possible scenarios.

- **Citizen fora**. The RSA emphasises citizen fora as a particularly important form of public dialogue. These are not just a one-way process of gaining information from the public, but focus on an iterative dialogue where expert stakeholders and citizens work together to produce recommendations for policymakers.

[53] We are grateful to Sarah Castell (Ipsos Mori) for valuable input on this section.

[54] Hallman in Jamieson et al (2017).

[55] Royal Society for the encouragement of Arts, Manufactures and Commerce (RSA, 2018).

[56] Which might also involve cost-benefit analysis, drawing on expert perspectives, and evidence on the concrete impacts of technology on society.

[57] International Association for Public Participation. www.dvrpc.org/GetInvolved/PublicParticipation/pdf/IAP2_public_participationToolbox.pdf

[58] www.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation

This is often used where a problem requires navigating trade-offs and considering multiple different plausible solutions. This form of public dialogue is particularly well-suited to exploring and resolving some of the trade-offs we discussed in section 4.

In all these forms of public engagement, the resulting views represent only a snapshot taken at a single moment in time: it will be important also to keep track of how values and perspectives change over time. Over the next few years, for example, concerns around data privacy might grow stronger – or they might dissipate entirely.

We outline existing public engagement work in more detail as part of the literature review in appendix 1, section 4. Based on the work that has already been done, we can identify examples of specific questions for public engagement around the four central tensions (although in order to explore in-depth attitudes to any given technology, many more questions will be relevant).

Accuracy **versus** fair and equal treatment

- How do individuals experience situations when major decisions about them are being taken with the aid of ADA technology?
- Under what circumstances are people willing to accept differential effectiveness of a technology for different groups?
- What do people consider to be 'fair and equal treatment' in different contexts?

Personalisation **versus** solidarity and citizenship

- In what contexts do people seek out or endorse individualised information or options specifically tailored to a certain 'profile' they fit?
- How does this change depending on the level of personal benefit?
- How does it change depending on the field (e.g. health, entertainment, political advertising)?
- How do people experience changes in the public sphere due to automation?

Quality and efficiency of services **versus** privacy and informational autonomy

- When do people endorse the use of their personal data to make services more efficient?
- How do these attitudes differ depending on exactly what data is being used, who is making use of it, and for what purpose?
- How do these attitudes differ between groups?

Convenience **versus** self-actualisation and dignity

- How do people experience loss of different jobs or tasks to automation?
- How do answers to this question differ by demographic factors?
- In the light of increasing automation, what would people's ideal working patterns be?
- How would people like to interact with ADA technologies in the workplace? Which tasks would they prefer to be taken over by these technologies?

In addition, there are several overarching questions to be investigated, which touch upon all four of these tensions and could be applied to others:

- Why, and to what extent, is it important for publics to understand a given technology (including its mechanisms, purposes, owners and creators, etc.)?
- If algorithms are being used as part of making decisions that significantly impact people's lives, what kinds of explanations of its decisions are needed and appropriate? Does this differ depending on the type of decision, or who is ultimately in charge of it?
- What do the public see as the biggest opportunities and risks of different technologies, and how do they think about trade-offs between the two? How does this differ based on demographic factors? How does this differ based on people's personal experience with different technologies?

5.3 Applying evidence to resolve tensions

Having highlighted some specific examples of questions in each of the previous subsections, we now pull all of this together to highlight how a stronger evidence base can help unpack and resolve our four central tensions.

Accuracy **versus** fair and equal treatment

This tension arises when users embed an algorithm as part of a decision-making process, but there are trade-offs between the benefits an algorithm brings (increased accuracy, for example), and its potential costs (potential biases which may lead to unfair outcomes). To make progress here, we need not only to understand the strengths and limitations of a given algorithm in a specific context, but also to *compare* them to the relative strengths and limitations of human decision-makers. More research comparing the predictive accuracy and biases of algorithms compared to those of humans in different contexts would make it clearer when accuracy and fair treatment are really in conflict, and make it easier to decide when using an algorithm is appropriate.

Understanding different societal perspectives will also be a crucial part of navigating the trade-offs that arise when we use algorithms in decision processes. Does automation of life-changing decisions reduce or strengthen trust in public institutions? What level and types of explainability do different groups need to trust algorithms that impact their lives? What kinds of information and characteristics is it acceptable for an algorithm to use in making different types of decisions, and what kinds might be considered unfair?[59]

Personalisation **versus** citizenship and solidarity

Here a tension arises because data and machine learning can be used to personalise services and information, with both positive and negative implications for the common good of democracies. In order to understand the trade-offs here, however, we need better evidence than the current, often sensationalist, headlines on what is currently technically possible: what kinds of inferences about groups and individuals can be drawn from publicly or privately available data? What evidence is there that using these conclusions to target information and services is more effective and with respect to what purposes? What kinds of influence on attitudes might this make possible?

We also need to collect better evidence on the attitudes towards increasing personalisation: where do people see this as benefiting their lives, where is it harmful, and is it possible to draw a clear line between the two? Personalisation is sometimes welcome and sometimes 'creepy' and we need to know when and why. To the extent that people reject personalisation in a given domain, which concerns underlie this attitude – for example are they concerned about no longer sharing the same informational sphere as other members of society, or are they more worried about whether the ability to tailor information gives organisations too much power to manipulate individuals? As with privacy, we might expect attitudes around personalisation and solidarity to change over time: we need to consider what these changes might be, and how they might change the tensions that arise. Scholarship on the wider social and political implications of personalisation for democracy, the welfare state, and political engagement are also essential.

Quality and efficiency of services **versus** privacy and informational autonomy

As mentioned, this tension arises because personal data may be used to improve public services, but doing so raises challenges for privacy and autonomy of individuals over their information. However, technical methods exist for drawing inferences from aggregate data while protecting the privacy of individual subjects, such as differential privacy.[60] The more successful these methods are, and the more we are able to implement new models of consent, the less there is a tension between innovative uses of data and privacy. Understanding the current status of technical research in this area, and what it can and cannot do, will therefore be important for understanding this tension.

Where a trade-off between quality of service and privacy remains, understanding public opinion, both uninformed and informed, will be essential for resolving it. It might be that publics endorse the use of their personal data in some cases – lifesaving medical applications, say – but not in others. Notions of privacy and its importance may also evolve over time, changing what aspects of it become more or less central. Expert judgment about the broader social and legal implications of privacy violations and enhancement should supplement these studies.

Convenience **versus** self-actualisation and dignity

At the heart of this tension lies the fact that automation has clear benefits: saving people time and effort spent on mundane tasks, increased convenience and access, but *too much* automation could threaten our sense of achievement, self-actualisation, and dignity as humans. To explore this tension we therefore need to start with clearer thinking about where automation is seen to be largely beneficial (perhaps because the tasks in question

[59] Building on work such as Grgić-Hlača et al. (2018), who study human perceptions of fairness in algorithmic decision-making in the context of criminal risk prediction, proposing a framework to understand why people perceive certain features as fair or unfair in algorithmic decisions.

[60] Differential privacy methods aim to maximise the accuracy of inferences drawn from a database while minimising the chance of identifying individual records, by ensuring that the addition or removal of a single datapoint does not substantially change the outcome. Though differential privacy is not an absolute guarantee of privacy, it ensures that the risk to an individual of having their data part of a database is limited. For reviews, see for example, Hilton and Dwork (2008).

are mindless and alienating), and where it is threatening and inappropriate on moral or prudential grounds (e.g. automating complex tasks involved in education, warfare, immigration, justice, and relationships may be offensive even if narrowly effective). Understanding the perspectives of a wide range of different groups on this question will be especially important, because the activities that are highly valued in one age group or culture may be very different from another.

If we can begin to agree on some tasks that it would be beneficial to automate, then we can begin to collect evidence on current technological capabilities in these areas, and to assess what is needed to make progress. By contrast, if we can more clearly identify the abilities that are central to human flourishing or we otherwise do *not* want to automate, then measuring current capabilities in these areas can help us better assess any threats, and think about potential responses.

In addition to research on the tensions we identify here, we also welcome bold multidisciplinary studies of tensions that explore more radical political and technological changes to come: for example, how ADA-based technologies could look if they were not pursued primarily for profit or for geopolitical advantage, and what socio-economic arrangements alternative to capitalism these technologies could make possible.

5.4 Summary and recommendations

We recommend that research and policy work on the ethics of ADA-based technologies should invest in developing a stronger evidence-base around (a) current and potential technological capabilities, and (b) societal attitudes and needs, identifying and challenging the many assumptions. In particular:

- **Deepening understanding of technological capabilities and limitations** in areas particularly relevant to key ethical and societal issues.

Often discussion of ethical and societal issues is founded on unexamined assumptions about what is currently technologically possible. To assess confidently the risks and opportunities of ADA for society, and to think more clearly about trade-offs between values, we need more critical examination of these assumptions.

- **Building a stronger evidence base on the current uses and impacts** of ADA-based technologies, especially around key tensions and as they affect marginalised or underrepresented groups.

Understanding specific applications of ADA-based technologies will help us to think more concretely about where and how tensions between values are most likely to arise, and how they might be resolved. Evidence on current societal impacts of technology will provide a stronger basis on which to assess the risks, and to predict possible future impacts.

- Building on existing public engagement work to **better understand the perspectives of different members of society** on important issues and trade-offs.

As we have emphasised, navigating the ethical and societal implications of ADA requires us to acknowledge tensions between values that use of these technologies promote, and values they might threaten. Since different publics will be affected differently by technology, and may hold different values, resolving these tensions requires us to understand varied public opinion on questions related to tensions and trade-offs.

6. Conclusion: A roadmap for research

In this report, we have explored the state of current research and debates on ethical and societal impacts of algorithms, data, and AI, to identify what has been achieved so far and what needs to be done next.

In section 2, we identified a number of key concepts used to categorise the issues raised by ADA-based technologies and a number of ethical principles and values that most actors agree are important. We also identified three key tasks that we believe need to be prioritised in order to move these discussions forward, namely:

- **Task 1 – Concept building**: Addressing the vagueness and ambiguities in the central concepts used in discussions of ADA, identifying important differences in how terms are used and understood across disciplines, sectors, publics and cultures, and working to build bridges and consensus around these where possible.

- **Task 2 – Resolving tensions and trade-offs:** Recognising and articulating tensions between the different principles and values at stake in debates about ADA, determining which of these tensions can be overcome through better technologies or other practical solutions, and developing legitimate methods for the resolution of any trade-offs that have to be made.

- **Task 3 – Developing an evidence base**: Building a stronger evidence base on technological capabilities, applications, and societal needs relevant to ADA, and using these to resolve tensions and trade-offs.

Throughout the report, we have made a number of recommendations and suggested questions for research relevant to achieving each of these tasks. We summarise these below. These are by no means meant to be exhaustive of the questions that could be fruitfully pursued in relation to the ethical and societal impacts of ADA. However, they highlight areas where we believe there is a strong potential for future research to provide high-value contributions to this field.

We envisage the study of the ethical and societal impacts of ADA as a pluralistic interdisciplinary and intersectoral enterprise, drawing on the best of the available methods of the humanities, social sciences and technical disciplines, as well as the expertise of practitioners. Together, the recommendations yield a roadmap for research that strikes a balance between respecting and learning from differences between stakeholders and disciplines, and encouraging consistent and productive criticism that provides relevant and practical knowledge. The point of this knowledge base is to improve the standards, regulations, and systems of oversight of the ADA technologies, which are currently uncertain and in flux. We urge that new approaches to governance and regulation be duly sensitive to the tensions described above and devise legitimate and inclusive institutions that will help communities to identify, articulate, and navigate these tensions, and others as they arise, in the context of greater and more pervasive automation of their lives.

Questions for research

Task 1: Concept Building

To clarify and resolve ambiguities and disagreements in the use of key terms:

- What are the different meanings of key terms in debates about ADA? Such terms include, but are not limited to: fairness, bias, discrimination, transparency, explainability, interpretability, privacy, accountability, dignity, solidarity, convenience, empowerment, and self-actualisation.

- How are these terms used interchangeably, or with overlapping meaning?

- Where are different types of issues being conflated under similar terminology?

- How are key terms used divergently across disciplines, sectors, cultures and publics?

To build conceptual bridges between disciplines and cultures:

- What other cultural perspectives, particularly those from the developing world and marginalised groups, are not currently strongly represented in research and policy work around ADA ethics? How can these perspectives be included, for example by translating relevant policy and research literature, or by building collaborations on specific issues?

- What relevant academic disciplines are currently underrepresented in research on ADA ethics, and what kinds of interdisciplinary research collaborations could help include these disciplines?

To build consensus and manage disagreements:

- Where ambiguities and differences in use of key terms exist, how can consensus and areas of common understanding be reached?

- Where consensus cannot easily be reached, how can we acknowledge, and work productively with, important dimensions of disagreement?

Task 2: Tensions and Trade-offs

To better understand the four central tensions:

- To what extent are we facing true dilemmas, dilemmas in practice, or false dilemmas?

- For the four central tensions, this includes asking:

  - How can the most accurate predictive algorithms be used in a way that does not violate fairness and equality?

  - How can we get the benefits of personalisation and respect the ideals of solidarity and citizenship?

  - How can we use personal data to improve public services and preserve or enhance privacy and informational autonomy?

  - How can we use automation to make our lives more convenient and at the same time promote self-actualisation and dignity?

To legitimate trade-offs:

- How do we best give voice to all stakeholders affected by ADA and articulate their interests with rigour and respect?

- What are acceptable and legitimate trade-offs that are compatible with rights and entitlements of those affected by these technologies?

- Which mechanisms of resolution are most likely to receive broad acceptance?

- For the four central tensions, this includes asking:

  - When, if ever, is it acceptable to use an algorithm that performs worse for a specific subgroup, if that algorithm is more accurate on average across a population?

  - How much should we restrict personalisation of advertising and public services for the sake of preserving ideals of citizenship and solidarity?

  - What risks to privacy and informational autonomy is it acceptable to incur for the sake of better disease screening or greater public health?

  - What kinds of skills should always remain in human hands, and therefore where should we reject innovative automation technologies?

To identify new tensions beyond those highlighted in this report:

- Where might the harms and benefits of ADA-based technologies be unequally distributed across different groups?

- Where might uses of ADA-based technologies present opportunities in the near term but risk compromising important values in the long term?

- Where might we be thinking too narrowly about the impacts of technology? Where might applications that are beneficial from a narrow or individualistic view produce negative externalities?

Task 3: Developing an evidence base

To deepen our understanding of technological capabilities and limitations:

Overarching questions

- What do we need to understand about technological capabilities and limitations in order to assess meaningfully the risks and opportunities they pose in different ethical and societal contexts?

- How might advances in technological capabilities help resolve tensions between values in applications of ADA, and what are the limitations of technology for this purpose?

Applying these overarching questions to our four specific tensions:

- Accuracy **versu**s fair and equal treatment
  - To what extent does accuracy trade off against different definitions of fairness?
  - What forms of interpretability are desirable from the perspective of different stakeholders?

  - What forms of interpretability can be ensured in state-of-the-art models?
  - To what extent is it possible to ensure adequate interpretability without sacrificing accuracy (or other properties, e.g. privacy)?

- Personalisation **versus** solidarity and citizenship
  - Are there any in-principle or in-practice limits to how fine-grained personalisation can become (using current or foreseeable technology)?
  - To what extent does personalisation meaningfully affect relevant outcomes (e.g. user satisfaction, consumer behaviour, voting patterns)?

- Quality and efficiency of services **versus** privacy and informational autonomy
  - How much could machine learning and 'big data' improve different public services? Can potential gains be quantified?
  - To what extent do current methods allow the use of personal data in aggregate, while protecting the privacy of individuals' data?
  - What are the best methods for ensuring meaningful consent?

- Convenience **versus** self-actualisation and dignity
  - What types of tasks can feasibly be automated using current or foreseeable technologies?
  - What would the costs (e.g. energy and infrastructure requirements) be for widespread automation of a given task?

To build a stronger evidence base on the current uses and impacts of technology:

Overarching questions

- Across different sectors (energy, health, law, etc.), what kinds of ADA-based technologies are already being used, and to what extent?

- What are the societal impacts of these specific applications, in particular on groups that might be disadvantaged (such as people of colour), underrepresented (such as women) or vulnerable (such as children or older people)?

Applying these overarching questions to our four specific tensions:

- Accuracy **versus** fair and equal treatment
  - In what sectors and applications are ADA being used to inform decisions/predictions with implications for people's lives?
  - Is it possible to determine how often these result in differential treatment of different socially salient groups?
  - How easy to interpret are the algorithms being used to inform decisions that have implications for people's lives? And what recourse do individuals have for challenging decisions?

- Personalisation **versus** solidarity and citizenship
  - What kinds of messages, interventions and services are already being personalised using machine learning, and in what sectors?
  - How 'fine-grained' is this personalisation, and on what kinds of categories is it based?
  - What evidence is there that this personalisation can substantially affect attitudes or behaviour?

- Quality and efficiency of services **versus** privacy and informational autonomy
  - In what specific sectors and applications are ADA being used to improve the efficiency of public services?
  - What impacts are these specific applications having on autonomy and privacy?

- Convenience **versus** self-actualisation and dignity
  - What effects is automation already having on daily living activities of different publics?

To better understand the perspectives of different interest groups:

Overarching questions:

- What are the publics' preferences about understanding a given technology (including its mechanisms, purposes, owners and creators, etc.)?

- If algorithms are being used as part of making decisions that significantly impact people's lives, what kinds of explanations of these decisions would people like to be able to access? Does this differ depending on the type of decision, or who is ultimately in charge of it?

- What do different publics see as the biggest opportunities and risks of different technologies, and how do they think about trade-offs between the two? How does this differ based on demographic factors? How does this differ based on people's personal experience with different technologies?

Applying these overarching questions to our four specific tensions:

- Accuracy **versus** fair and equal treatment
  - How do different publics experience differential effectiveness of a technology?
  - What do people consider to be 'fair and equal treatment' in different contexts?

- Personalisation **versus** solidarity and citizenship
  - In what contexts do people seek out or endorse individualised information or options specifically tailored to a certain 'profile' they fit?
  - How do people experience changes in the public sphere due to automation?

- Quality and efficiency of services **versus** privacy and informational autonomy
  - When do publics endorse the use of their personal data to make public services more efficient?
  - How are these attitudes different depending on exactly what data is being used, who is making use of it, and for what purpose?
  - How do these attitudes differ across groups?

- Convenience **versus** self-actualisation and dignity
  - What tasks and jobs are people most concerned about losing to automation? How do answers to this question differ by demographic factors?
  - In the light of increasing automation, what would ideal working patterns be?
    - How would people like to interact with ADA technologies in the workplace?
    - Which tasks is it ethically and prudentially appropriate for technologies to take over?

Bibliography

Acs, G., Melis, L., Castelluccia, C., & De Cristofaro, E. (2018). Differentially private mixture of generative neural networks. *IEEE Transactions on Knowledge and Data Engineering*.

Adams, F. and Aizawa, K. (2001). The Bounds of Cognition. *Philosophical Psychology*, 14(1): 43–64.

Adel, T., Ghahramani, Z., & Weller, A. (2018). Discovering Interpretable Representations for Both Deep Generative and Discriminative Models. In *International Conference on Machine Learning*: 50–59.

Aditya, S. (2017). Explainable Image Understanding Using Vision and Reasoning. Paper presented at the AAAI.

Aha, D.W., & Coman, A. (2017). The AI Rebellion: Changing the Narrative. Paper presented at the AAAI.

AI Now Institute. (2017). AI Now Symposium 2017 Report.

Alekseev, A. (2017). Artificial intelligence and ethics: Russian theory and communication practices. *Russian Journal of Communication*, 9(3): 294–296.

Alexandrova A. and D. Haybron (2011) High fidelity economics, in *Elgar Companion to Recent Economic Methodology* (edited by John Davis and Wade Hands), Edward Elgar, 94–117.

Alexandrova A. (2017) *A Philosophy for the Science of Well-being*, New York: Oxford University Press.

Alexandrova, A. (2018) Can the Science of Well-Being Be Objective?, *The British Journal for the Philosophy of Science*, 69(2): 421–445. https://doi.org/10.1093/bjps/axw027

Altman, A. 2015. Discrimination, *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/entries/discrimination/

Alkoby, S., & Sarne, D. (2017). The Benefit in Free Information Disclosure When Selling Information to People. Paper presented at the AAAI.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3): 251–261.

American Honda Motor Co. (2017). ASIMO: The World's Most Advanced Humanoid Robot. Available online: http://asimo.honda.com

Anderson, B., & Horvath, B. (2017). The rise of the weaponized ai propaganda machine. *Scout*, February, 12.

Anderson, M., & Anderson, S. L. (2007). The status of machine ethics: a report from the AAAI Symposium. *Minds and Machines*, 17(1): 1–10.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, May, 23.

Ovanessoff, A. and Plastino, E. (2017). How Can AI Drive South America's Growth? Accenture Research Report.

Arney, C. (2016). Our Final Invention: Artificial Intelligence and the End of the Human Era. *Mathematics and Computer Education*, 50(3): 227.

ASI Data Science and Slaughter & May. (2017). *Superhuman Resources: Responsible Deployment of AI in Business*.

Australian Computing Society. (2017). Australia's Digital Pulse in 2017.

Barocas, S. (2014). Data mining and the discourse on discrimination. Proceedings of the Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining (KDD). https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review,* 104: 671.

Becker, B. (2006). Social robots-emotional agents: Some remarks on naturalizing man-machine interaction. *International Review of Information Ethics* 6: 37–45.

Bei, X., Chen, N., Huzhang, G., Tao, B., & Wu, J. (2017). Cake cutting: envy and truth. Paper presented at the Proceedings of the 26th International Joint Conference on Artificial Intelligence.

Bei, X., Qiao, Y., & Zhang, S. (2017). *Networked fairness in cake cutting*. arXiv preprint arXiv:1707.02033.

Belle, V. (2017). Logic meets probability: towards explainable AI systems for uncertain worlds. Paper presented at the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI.

Bess, M. (2010). Enhanced Humans versus "Normal People": Elusive Definitions, *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 35(6): 641–655. https://doi.org/10.1093/jmp/jhq053

Binns, R. (2017). *Fairness in Machine Learning: Lessons from Political Philosophy*. arXiv preprint arXiv:1712.03586.

Biran, O., & McKeown, K. R. (2017). Human-Centric Justification of Machine Learning Predictions. Paper presented at the IJCAI.

Bloomberg News. (2018). *China Now Has the Most Valuable AI Startup in the World*.

Boddington, P., Millican, P., & Wooldridge, M. (2017). Minds and Machines Special Issue: Ethics and Artificial Intelligence. *Minds and Machines*, 27(4): 569–574.

Bogaerts, B., Vennekens, J., & Denecker, M. (2017). Safe inductions: An algebraic study. Paper presented at the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI).

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277–284.

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Filar, B. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv preprint arXiv:1802.07228.

Burch, K.T. (2012). *Democratic transformations: Eight conflicts in the negotiation of American identity*. A&C Black.

Burns, T.W., O'Connor, D.J., & Stocklmayer, S.M. (2003) Science communication: a contemporary definition. *Public Understanding of Science*, 12: 183–202.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12.

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). *Ethical considerations in artificial intelligence courses*. arXiv preprint arXiv:1701.07769.

Bygrave, L. A. (2001). Automated profiling: minding the machine: article 15 of the ec data protection directive and automated profiling. *Computer Law & Security Review*, 17(1): 17–24.

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2): 277–292.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Cave, S. (2017) *Intelligence: A History*. Aeon.

Cave, S., & Dihal, K. (2018) Ancient dreams of intelligent machines: 3,000 years of robots. *Nature*, 559: 473–475.

Cech, E. A. (2014). Culture of disengagement in engineering education? *Science, Technology, & Human Values*, 39(1): 42–72.

Chace, C. (2015). *Surviving AI: The promise and peril of artificial intelligence*. Bradford: Three Cs Publishing.

Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S. (2017). *Plan explanations as model reconciliation: Moving beyond explanation as soliloquy*. arXiv preprint arXiv:1701.08317.

Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). *Fair clustering through fairlets*. Paper presented at the Advances in Neural Information Processing Systems.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Clark, A. (1996). *Being There: Putting Brain, Body, and World Together Again*. Cambridge: MIT Press.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.

Clark, A. and D. Chalmers. (1998). The Extended Mind. *Analysis*, 58: 7–19.

Clifford, D., Graef, I., & Valcke, P. (2018). *Pre-Formulated Declarations of Data Subject Consent–Citizen-Consumer Empowerment and the Alignment of Data, Consumer and Competition Law Protections*.

Coeckelbergh, M., Pop, C., Simut, R., Peca, A., Pintea, S., David, D. & Vanderborght, B. (2016). A Survey of Expectations About the Role of Robots in Robot-Assisted Therapy for Children with ASD: Ethical Acceptability, Trust, Sociability, Appearance, and Attachment. *Science and Engineering Ethics* 22 (1): 47–65.

Coggon, J., and J. Miola. (2011). Autonomy, Liberty, and Medical Decision-Making. *The Cambridge Law Journal*, 70(3): 523–547.

Collins, S. and A. Ruina. (2005). A bipedal walking robot with efficient and human-like gait. Proceedings IEEE International Conference on Robotics and Automation, Barcelona, Spain.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral Decision Making Frameworks for Artificial Intelligence. Paper presented at the AAAI.

Cowls, J., & Floridi, L. (2018). Prolegomena to a White Paper on an Ethical Framework for a Good AI Society. SSRN Electronic Journal.

Crawford, K. (2016). Artificial intelligence's white guy problem. *The New York Times*, 25.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, 538(7625): 311.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Stanford Law Review*, 43(6): 1241–1299.

Dafoe, A. (2018). *AI Governance: A Research Agenda*. University of Oxford.

Daly, A. (2016). *Private power, online information flows and EU law: Mind the gap*. Bloomsbury Publishing.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. Paper presented at the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence.

Davies, J. (2016). Program good ethics into artificial intelligence. *Nature News*.

Dawkins, R. (1982). *The Extended Phenotype*. New York: Oxford Press.

Devlin, H. (2017). AI programs exhibit racial and gender biases, research reveals. *The Guardian*, 13.

Dietterich, T. G. (2017). Steps toward robust artificial intelligence. *AI Magazine*, 38(3): 3–24.

Dignum, V. (2018). *Ethics in artificial intelligence: introduction to the special issue*.

Ding, J. (2018). *Deciphering China's AI Dream*. University of Oxford.

Dunbar, M. (2017). To Be a Machine: Adventures Among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death. *The Humanist*, 77(3): 42.

Dwork, C. (2008). Differential privacy: A survey of results. Paper presented at the International Conference on Theory and Applications of Models of Computation.

Dwork, C. (2017). What's Fair? Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Paper presented at the Proceedings of the 3rd innovations in theoretical computer science conference.

Edwards, L., & Veale, M. (2017). Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking for. *Duke Law and Technology Review* 16(1): 18.

Ess, C. (2006). Ethical pluralism and global information ethics. *Ethics and Information Technology,* 8(4): 215–226.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* 542(7639): 115.

EU EDPS Ethics Advisory Group. (2018). Towards a digital ethics.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Group on Ethics in Science and New Technologies. (2018). Statement on Artificial Intelligence, Robotics and "Autonomous" Systems.

Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. Paper presented at the AAAI.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Fish, B., Kun, J., & Lelkes, Á. D. (2016). A confidence-based approach for balancing fairness and accuracy. Paper presented at the Proceedings of the 2016 SIAM International Conference on Data Mining.

Fisher, D. H. (2017). A Selected Summary of AI for Computational Sustainability. Paper presented at the AAAI.

Forster, E. M. (1947). *Collected short stories of EM Forster.* Sidgwick and Jackson.

Frank, R. H. (2000). Why is cost-benefit analysis so controversial? *The Journal of Legal Studies*, 29(S2): 913–930.

Freuder, E. C. (2017). Explaining Ourselves: Human-Aware Constraint Reasoning. Paper presented at the AAAI.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation? *Technological forecasting and social change*, 114: 254–280.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). *On the (im) possibility of fairness*. arXiv preprint arXiv:1609.07236.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems* (TOIS), 14(3): 330–347.

Future Advocacy and The Wellcome Trust. (2018). *Ethical, social and political challenges of artificial intelligence in health*.

Garrett, R. K., E. C. Nisbet, and E. K. Lynch. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication* 63(4): 617–637.

Garrett, R. K., E. C. Weeks, and R. L. Neo. (2016). Driving a wedge between evidence and beliefs: How online ideological news exposure promotes political misperceptions. Journal of Computer-Mediated *Communication* 21(5): 331–348.

Gellert, R. (2015). Data protection: a risk regulation? Between the risk management of everything and the precautionary alternative. *International Data Privacy* Law, 5(1): 3.

Gellert, R. (2018). Understanding the notion of risk in the General Data Protection Regulation. *Computer Law & Security Review*, 34(2): 279–288.

Goh, G., Cotter, A., Gupta, M., & Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. Paper presented at the Advances in Neural Information Processing Systems.

Goldsmith, J., & Burton, E. (2017). Why Teaching Ethics to AI Practitioners Is Important. Paper presented at the AAAI.

Goodman, B., & Flaxman, S. (2016). *European Union regulations on algorithmic decision-making and a "right to explanation"*. arXiv preprint arXiv:1606.08813.

Government Office for Science. (2016). *Artificial intelligence: opportunities and implications for the future of decision making.*

Government Office for Science. (2017). *The Futures Toolkit: Tools for Futures Thinking and Foresight Across UK Government.*

GPI Atlantic. (1999). Gender Equality in the Genuine Progress Index. Made to Measure Symposium Synthesis Paper, Halifax, October 3–6.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research*, 62: 729–754.

Graef, I. (2016). *EU Competition Law, Data Protection and Online Platforms: Data as Essential Facility*: Kluwer Law International.

Grafman, J. and I. Litvan. (1999). Evidence for Four Forms of Neuroplasticity. In *Neuronal Plasticity: Building a Bridge from the Laboratory to the Clinic*. J. Grafman and Y. Christen (eds.). Springer-Verlag Publishers.

Grbovic, M., Radosavljevic, V., Djuric, N., Bhamidipati, N., & Nagarajan, A. (2015). Gender and interest targeting for sponsored post advertising at tumblr. Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). *Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction*. arXiv preprint arXiv:1802.09548.

Greenwald, A. G. (2017). An AI stereotype catcher. *Science*, 356(6334): 133–134.

Gribbin, J. (2013). *Computing with quantum cats: From Colossus to Qubits*. Random House.

Gunkel, D. J., & Bryson, J. (2014). Introduction to the special issue on machine morality: The machine as moral agent and patient. *Philosophy & Technology*, 27(1): 5–8.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). *The off-switch game.* arXiv preprint arXiv:1611.08219.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7): 1445–1459.

Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., & Giannotti, F. (2015). Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery*, 29(6): 1733–1782.

Hanisch, C. (1969). *The personal is political.* Available at www.carolhanisch.org/CHwritings/PIP.html

Hanisch, C. (2006). *The personal is political: The women's liberation movement classic with a new explanatory introduction.* Women of the World, Unite.

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow.* Random House.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Paper presented at the Advances in neural information processing systems.

Harel, Y., Gal, I. B., & Elovici, Y. (2017). Cyber Security and the Role of Intelligent Systems in Addressing its Challenges. *ACM Transactions on Intelligent Systems and Technology* (TIST), 8(4): 49.

Haybron, D. M., & Alexandrova, A. (2013). Paternalism in economics. *Paternalism: Theory and practice*, (eds Christian Coons and Michael Weber), Cambridge University Press, 157–177.

Helberger, N., Zuiderveen Borgesius, F. J., & Reyna, A. (2017). *The perfect match? A closer look at the relationship between EU consumer law and data protection law.*

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Zwitter, A. (2017). Will democracy survive big data and artificial intelligence? *Scientific American*, 25.

Hilton, M. *Differential privacy: a historical survey.* Cal Poly State University.

House of Commons Science and Technology Committee, The Big Data Dilemma. 12 February 2016, HC 468 2015–16.

Hurley, S. L. (1998). Vehicles, Contents, Conceptual Structure and Externalism. *Analysis* 58: 1–6.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2018) Ethically aligned design: a vision for prioritizing human wellbeing with artificial intelligence and autonomous systems.

Imberman, S. P., McManus, J., & Otts, G. (2017). Creating Serious Robots That Improve Society. Paper presented at the AAAI.

Institute of Technology and Society in Rio. (2017). *Big Data in the Global South: Report on the Brazilian Case Studies.*

Ipsos MORI and the Royal Society. (2017). *Public views of Machine Learning.*

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., & Roth, A. (2016). *Fairness in reinforcement learning.* arXiv preprint arXiv:1611.03071.

Johndrow, J. E., & Lum, K. (2017). *An algorithm for removing sensitive information: application to race-independent recidivism prediction*. arXiv preprint arXiv:1703.04957.

Jotterand, F. and V. Dubljevic (Eds). (2016). *Cognitive Enhancement: Ethical and Policy Implications in International Perspectives*. Oxford University Press.

Kamarinou, D., Millard, C., & Singh, J. (2016). Machine Learning with Personal Data. *Queen Mary School of Law Legal Studies Research Paper*, 247.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1–33.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. Paper presented at the Data Mining (ICDM), 2010 IEEE 10th International Conference on Computer and Information Technology.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. Paper presented at the Data Mining (ICDM), 2012 IEEE 12th International Conference on Dating Mining.

Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information systems*, 35(3): 613–644.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Kaplan, J. (2015). *Humans need not apply: A guide to wealth and work in the age of artificial intelligence*. Yale University Press.

Kaplan, J. (2016). *Artificial Intelligence: What everyone needs to know*. Oxford University Press.

Kearns, M., Roth, A., & Wu, Z. S. (2017). Meritocratic fairness for cross-population selection. Paper presented at the International Conference on Machine Learning.

Kleinberg, J., Ludwig, J., Mullainathan, S. (2016). *A Guide to Solving Social Problems with Machine Learning*. Harvard Business Review.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. arXiv preprint arXiv:1609.05807.

Koops, B. (2013). *On decision transparency, or how to enhance data protection after the computational turn. Privacy, due process and the computational turn: the philosophy of law meets the philosophy of technology*, 189–213.

Kraemer, F., Van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, 13(3): 251–260.

Kristoffersson, A., Coradeschi, S., Loutfi, A., & Severinson-Eklundh, K. (2014). Assessment of interaction quality in mobile robotic telepresence: An elderly perspective. *Interaction Studies* 15(2): 343–357.

Kuner, C., Svantesson, D. J. B., Cate, F. H., Lynskey, O., & Millard, C. (2017). Machine learning with personal data: is data protection law smart enough to meet the challenge? *International Data Privacy Law*, 7(1), 1–2.

Kurzweil, R. (2013). *How to create a mind: The secret of human thought revealed*. Penguin.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Paper presented at the Advances in Neural Information Processing Systems.

Kuzelka, O., Davis, J., & Schockaert, S. (2017). *Induction of interpretable possibilistic logic theories from relational data*. arXiv preprint arXiv:1705.07095.

Kökciyan, N., & Yolum, P. (2017). Context-Based Reasoning on Privacy in Internet of Things. Paper presented at the IJCAI.

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable Agency for Intelligent Autonomous Systems. Paper presented at the AAAI.

Lehmann, H., Iacono, I., Dautenhahn, K., Marti, P. and Robins, B. (2014). Robot companions for children with down syndrome: A case study. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 15(1), pp. 99–112.

Levy, N. Rethinking Neuroethics in the Light of the Extended Mind Thesis, *The American Journal of Bioethics*, 7(9): 3–11.

Lewis-Kraus, G. (2016). The great AI awakening. *The New York Times Magazine*, 14.

Li, Fei-Fei. (2018). How to Make A.I. That's Good for People. *The New York Times*.

Lipton, Z. C. (2016). The Mythos of Model Interpretability. ICML 2016 Workshop on Human Interpretability in Machine Learning.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. Paper presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.

Lyons, J. B., Clark, M. A., Wagner, A. R., & Schuelke, M. J. (2017). Certifiable Trust in Autonomous Systems: Making the Intractable Tangible. *AI Magazine*, 38(3).

Marcus, G. (2012). Will a Robot Take Your Job? *The New Yorker*.

Marcus, G. (2013). Why we should think about the threat of artificial intelligence. *The New Yorker*.

Marien, M. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. *Cadmus*, 2(2): 174.

Mattu, S. and Hill, K. (2018) *The House That Spied on Me*. Gizmodo.

McAllister, R., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R., & Weller, A. V. (2017). *Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning*.

McFarland, D. (2009). *Guilty robots, happy dogs: the question of alien minds*. Oxford University Press.

Mei, J.-P., Yu, H., Shen, Z., & Miao, C. (2017). A social influence based trust model for recommender systems. *Intelligent Data Analysis*, 21(2): 263–277.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbounded*. Palgrave Macmillan.

Mendoza, I., & Bygrave, L. A. (2017). The Right not to be Subject to Automated Decisions based on Profiling. In *EU Internet Law*: 77–98.

Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). *Should robots be obedient?* arXiv preprint arXiv:1705.09990.

Mindell, D. (2002). *Between human and machine: feedback, control, and computing before cybernetics*. Baltimore: Johns Hopkins University Press.

Minsky, M. (1982). *Semantic information processing*: MIT Press.

Minton, S. N. (2017). The Value of AI Tools: Some Lessons Learned. *AI Magazine*, 38(3).

Monbiot, G. (2017). Big data's power is terrifying. That could be good news for democracy. *The Guardian*.

Montréal Declaration on Responsible AI. (2018). Montréal Declaration for a Responsible Development of Artificial Intelligence. Available at www.montrealdeclaration-responsibleai.com/the-declaration

Moore, A. (2017). *Critical elitism: Deliberation, democracy, and the problem of expertise*. Cambridge University Press.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Mukherjee, S. (2017). A.I. Versus M.D.: What happens when a diagnosis is automated? *The New Yorker*.

Müller, V. C. (2014). Risks of artificial general intelligence. *Journal of Experimental and Theoretical Artificial Intelligence*, 26(3): 297-301.

Noble, S. U. (2018). *Algorithms of Oppression: How search engines reinforce racism*. NYU Press.

Noë, A. (2009). *Out of our heads*. Hill and Wang.

Novitske, L (2018). The AI Invasion is Coming to Africa and It's a Good Thing. *Stanford Social Innovation Review*.

Nushi, B., Kamar, E., Horvitz, E., & Kossmann, D. (2017). On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems. Paper presented at the AAAI.

Omidyar Network and Upturn. (2018). Public scrutiny of automated decisions: early lessons and emerging methods. Available online at www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Open Data Institute. (2017). *Helping organisations navigate concerns in their data practices*. Available online at https://theodi.org/article/data-ethics-canvas/

Pagallo, U. (2017). From automation to autonomous systems: a legal phenomenology with problems of accountability. Paper presented at the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

Parens, E. (1998). *Enhancing human traits: Ethical and social implications* (Hastings Center studies in ethics). Washington, D.C.: Georgetown University Press.

Parens, E. (2015). *Shaping ourselves: On technology, flourishing, and a habit of thinking.* Oxford University Press.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press.

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. Paper presented at the Proceedings of the 2009 SIAM International Conference on Data Mining.

Phan, N., Wu, X., Hu, H., & Dou, D. (2017). Adaptive laplace mechanism: differential privacy preservation in deep learning. Paper presented at the 2017 IEEE International Conference on Data Mining (ICDM).

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. Paper presented at the Advances in Neural Information Processing Systems.

Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. *Health and technology,* 7(4): 351–367.

Prainsack, B., & Buyx, A. (2017). *Solidarity in biomedicine and beyond* (Vol. 33). Cambridge University Press.

National Science and Technology Council, Obama White House. (2016). *Preparing for the Future of Artificial Intelligence.*

Purtova, N. (2018). The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology,* 10(1): 40–81.

Quadrianto, N., & Sharmanska, V. (2017). Recycling privileged learning and distribution matching for fairness. Paper presented at the Advances in Neural Information Processing Systems.

Reed, C., Kennedy, E., & Silva, S. (2016). Responsibility, Autonomy and Accountability: legal liability for machine learning. *Queen Mary School of Law Legal Studies Research Paper* No. 243/2016. Available at SSRN: https://ssrn.com/abstract=2853462

Resnick, B. (2018). Cambridge Analytica's "psychographic microtargeting": what's bullshit and what's legit. *Vox.*

Richert, A., Müller, S., Schröder, S., and Jeschke, S. (2018). Anthropomorphism in social robotics: empirical results on human–robot interaction in hybrid production workplaces. *AI and Society* 33(3): 413–424.

Robins, B., Dautenhahn, K., and Dubowski, J., (2006). Does appearance matter in the interaction of children with autism with a humanoid robot? Interaction Studies. *Social Behaviour and Communication in Biological and Artificial Systems* 7(3): 509–542.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review,* 29(5): 582–638.

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). *Right for the right reasons: Training differentiable models by constraining their explanations.* arXiv preprint arXiv:1703.03717.

Royal Society. (2017). *Machine learning: the power and promise of computers that learn by example.*

Royal Society and The British Academy. (2017). *Data management and use: Governance in the 21st century.*

Royal Society for the encouragement of Arts, Manufactures and Commerce (RSA). (2018). *Artificial Intelligence: Real Public Engagement.*

Russell, C., Kusner, M. J., Loftus, J., & Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. Paper presented at the Advances in Neural Information Processing Systems.

Schermer, M. (2013). Health, Happiness and Human Enhancement – Dealing with Unexpected Effects of Deep Brain Stimulation. *Neuroethics,* 6(3): 435–445.

Scheutz, M. (2017). The case for explicit ethical agents. *AI Magazine,* 38(4): 57–64.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences,* 3: 417–457.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review.*

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law,* 7(4): 233–242.

Select Committee on Artificial Intelligence. (2018). *AI in the UK: Ready, Willing, and Able?* HL 100 2017–19. London: House of Lords.

Shapiro, L. A. (2004). *The Mind Incarnate.* MIT Press.

Sharkey, A. (2014). Robots and human dignity: a consideration of the effects of robot care on the dignity of older people. *Ethics and Information Technology* 16(1): 63–75.

Sharkey, A., and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology* 14(1): 27–40.

Shirk, J. L., H. L. Ballard, C. C. Wilderman, T. Phillips, A. Wiggins, R. Jordan, E. McCallie, M. Minarchek, B. V. Lewenstein, M. E. Krasny, and R. Bonney. (2012). Public participation in scientific research: a framework for deliberate design. *Ecology and Society* 17(2): 29. http://dx.doi.org/10.5751/ES-04705–170229

Shariff, A., Rahwan, I., and Bonnefon, J. (2016). Whose Life Should Your Car Save? *New York Times.*

Shell International BV. (2008). *Scenarios: An Explorer's Guide.*

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., Krasny, M. E. (2012). Public participation in scientific research: a framework for deliberate design. *Ecology and Society,* 17(2).

Simon, H. (1969). *The Sciences of the Artificial.* MIT Press.

Sintov, N., Kar, D., Nguyen, T., Fang, F., Hoffman, K., Lyet, A., & Tambe, M. (2017). Keeping It Real: Using Real-World Problems to Teach AI to Diverse Audiences. *AI Magazine,* 38(2).

Such, J. M. (2017). Privacy and autonomous systems. Paper presented at the Proceedings of the 26th International Joint Conference on Artificial Intelligence.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue,* 11(3): 10.

Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., Rusu, A. S. and David, D. O. (2012). Children with autism social engagement in interaction with Nao, an imitative robot: A series of single case experiments. *Interaction Studies* 13(3): 315–347.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* Knopf.

Tene, O., & Polonetsky, J. (2017). Taming the Golem: Challenges of Ethical Algorithmic Decision-Making. *NC Journal of Law and Technology,* 19(1): 125.

Thelisson, E. (2017). Towards trust, transparency, and liability in AI/AS systems. Paper presented at the Proceedings of the 26th International Joint Conference on Artificial Intelligence.

Tiberius, V. (2018). *Well-Being As Value Fulfillment: How We Can Help Each Other to Live Well.* Oxford University Press.

Tolomei, G., Silvestri, F., Haines, A., & Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking. Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Turkle, S. (2016). *Reclaiming conversation: The power of talk in a digital age.* Penguin.

Turkle, S. (2017). *Alone together: Why we expect more from technology and less from each other.* Hachette UK.

Vanderborght, B., Simut, R, Saldien, J., Pop, C., Rusu, A. S., Pintea, S., Lefeber, D. and David, D.O. (2012). Using the social robot probo as a social story telling agent for children with ASD. *Interaction Studies* 13(3): 348–372.

Varakin, D.A., Levin, D.T. and Fidler, R. (2004). Unseen and unaware: Implications of recent research on failures of visual awareness for human-computer interface design. *Human-Computer Interaction* 19(4): 389–422.

Vempati, S. S. (2016). *India and the Artificial Intelligence Revolution*. Carnegie Endowment for International Peace.

Vold, K. (2015). The Parity Argument for Extended Consciousness. *Journal of Consciousness Studies,* 22(3–4): 16–33.

Wachter, S. and Mittelstadt, B.D. (2018) A right to reasonable inferences: re-thinking data protection in the age of big data and AI, *Columbia Business Law Review.*

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2): 76–99.

Wachter-Boettcher, S. (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*: WW Norton & Company.

Walden, J., Jung, E., Sundar, S., and Johnson, A. (2015). Mental models of robots among senior citizens: An interview study of interaction expectations and design implications. *Interaction Studies* 16(1): 68–88.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong.* Oxford University Press.

Walsh, T. (2016). *The singularity may never be near*. arXiv preprint arXiv:1602.06462.

Walsh, T. (2017). *Android Dreams: The Past, Present and Future of Artificial Intelligence*. Oxford University Press.

Weller, A. (2017). *Challenges for transparency.* arXiv preprint arXiv:1708.01870.

Weiskopf, D. (2008). Patrolling the mind's boundaries. *Erkenntnis*, 68(2): 265–76.

Whitfield, C. (2018). *The Ethics of Artificial Intelligence.* PwC Australia.

Whittlestone, J., Nyrup, R., Alexandrova, A., and Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. Forthcoming in Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society.

Wheeler, M. (2010). Minds, Things, and Materiality. In L. Malafouris and C. Renfrew. (eds.), *The Cognitive Life of Things: Recasting the Boundaries of the Mind.* Cambridge: McDonald Institute Monographs. (Reprinted in J. Schulkin (ed.), *Action, Perception and the Brain: Adaptation and Cephalic Expression*. Basingstoke: Palgrave Macmillan.)

Wilson, R. A. (1994). Wide Computationalism. *Mind*, 103(411): 351–72.

Wilson, R. A. and A. Clark. (2009). How to situate cognition: Letting nature take its course. In Murat Aydede and P. Robbins (eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press, 55–77.

The Wilson Centre. (2017). *Artificial Intelligence: A Policy-Oriented Introduction.*

Wood, L., Lehmann, H., Dautenhahn, K., Robins, B., Rainer, A., and Syrdal, D. (2016). Robot-mediated interviews with children. *Interaction Studies* 17(3): 438–460.

World Wide Web Foundation. (2017). *Artificial Intelligence: Starting the Policy Dialogue in Africa.*

Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. Paper presented at the Advances in Neural Information Processing Systems.

Yuste, R. et al. (2017). Four Ethical Priorities for Neurotechnologies and AI. Nature News, *Nature Publishing Group*. www.nature.com/news/four-ethical-priorities-for-neurotechnologies-and-ai-1.22960

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., & Weller, A. (2017). From parity to preference-based notions of fairness in classification. Paper presented at Advances in Neural Information Processing Systems.

Zarkadakis, G. (2015). *In Our Own Image: Will artificial intelligence save or destroy us?* Random House.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. Paper presented at the International Conference on Machine Learning.

Žliobaite, I. (2015). *A survey on measuring indirect discrimination in machine learning*. arXiv preprint arXiv:1511.00148.

Žliobaite, I., Kamiran, F., & Calders, T. (2011). Handling conditional discrimination. Paper presented at the Data Mining (ICDM), 2011 IEEE 11th International Conference on data mining.

Appendices

Appendix 1: Summary of literature reviews

This report draws on a series of literature reviews of how the ethical and societal implications of algorithms, data and AI have been discussed across a range of academic disciplines, in policy and civil society reports, and in popular science and the media. The reviews cover over 100 academic papers from disciplines including (but not limited to) computer science, ethics, human computer interaction, law, and philosophy, 20 policy documents (at least one from each of the seven continents), over 25 of the most commonly cited popular books, news and media articles, and several reports documenting public engagement research.[61]

Sections 1–3 in this appendix summarise the main observations from each of these literature reviews. Section 4 summarises a number of overlapping themes that emerged.

1. Academic literatures

1a. Computer science and machine learning

We covered papers published in 2017 in the most relevant conferences and journals in the areas of AI, machine learning, data science, and data mining.[62] In most of the venues covered, less than 1% of papers were directly related to ethical or societal impacts of the technology.

In general, there seems be a "culture of disengagement" among technical researchers and engineers, who generally do not see ethical and societal questions raised by technology as their responsibility.[63] However, the last 2-3 years have seen a growing interest, as illustrated, for example, by relevant workshops and symposia at major conferences, the FAT/ML conference and the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

Where technical research directly addresses ethical and societal issues, unsurprisingly it tends to focus on those that can be framed or simplified in technical terms: including how we might explain or interpret the decisions of 'black box' machine learning systems, how we can assess the reliability of different systems, issues of privacy and data protection, and how we can build important values like 'fairness' into the use of algorithms and AI systems.

We also covered several survey papers looking at the way AI and data science professionals are trained in ethics. While there are ethical codes being developed for data scientists, most AI professionals haven't had any training at all in ethical issues. Since this space is so fast-moving, and there is no 'standard' route to becoming an AI professional, it's not yet clear exactly what should be covered in ethics training, or who exactly should receive it.

1b. Philosophy and ethics

We surveyed a range of recent papers listed on PhilPapers.org, the main index of English language philosophy publications, under several headings related to ethics in AI.

Most of these papers focus on the moral significance of the kinds of advanced AI technologies that might exist in the future.[64] Questions explored include whether and how artificial agents might make moral decisions, at what point (if ever) intelligent machines might have moral status normally accorded to humans,[65] and the implications of 'superintelligent' machines for our concepts of autonomy and what it means to be human.[66]

Work on current technologies is not widely discussed in the ethics of AI literature. Instead, these are usually covered under the headings of 'Information Ethics' or 'Computer Ethics'.[67] Early papers in this field

[61] All of the sources covered are referenced in the bibliography, and for each area of literature, we highlight the key references used.

[62] Including papers from the following conferences: IJCAI, AAAI, NIPS, ICML, KDD, ICDM and the following journals: AIJ, JAIR, AIMag, AIRev, MLJ, JMLR, TKDD, TPAMI, TIST, IDA.

[63] Cech, (2014).

[64] https://philpapers.org/browse/ethics-of-artificial-intelligence

[65] See Boddington et al. (2017); Gunkel et al. (2014); Muller (2014); Wallach and Allen (2009); Allen, Varner and Zinser (2010); Anderson and Anderson (2007).

[66] Bostrom (2003).

[67] See for examples Dignum (2018) and Brynum (2015).

have highlighted the issues of accountability, bias and value-ladenness in data and algorithmic decision-making. More recent work has started to analyse how key concepts such as 'transparency', 'bias', 'fairness' or 'responsibility' apply to ADA-based technologies. This literature is usually published in specialist journals (e.g. *Ethics and Information Technology*, *Philosophy & Technology*, *Science and Engineering Ethics*), or often in proceedings of technical ADA fields.

However, there seems to be a relative lack of systematic work analysing the ethical implications of current ADA technologies from a philosophical perspective. What literature exists still seems to be relatively low profile within the discipline of philosophy.

1c. Law

The academic law literature we covered mostly discusses questions of how the law (both existing and potential) can be used to mitigate risks from ADA-based technologies, particularly given rapid advances in AI and machine learning.

Some of the key questions covered include: what existing regulation means in practice for the use of data and algorithms (e.g. to what extent does the GDPR mandate a 'right to explanation', and what type of explanation?);[68] whether such regulation can actually solve the problems it aims to (e.g. how much does a 'right to explanation' help with issues of privacy, personal autonomy, and trust?);[69] and how existing law can deal with liability and accountability issues that arise with increasing deployment of AI.[70]

As well as the question of what law is needed around the use of AI, data, and algorithms, there is also a question of how these technologies might impact the legal process itself – changing the way that testimony and evidence are given, for example.

More than other fields, the legal literature tries to pull apart different interpretations of ambiguous terms – like 'privacy' or 'fairness' – and to discuss the implications of different meanings.

1d. Human-machine interaction

Human-machine interaction (HMI) is an interdisciplinary field combining philosophy, psychology, cognitive science, computer science, engineering and design. We looked at papers from prominent HMI journals including *Human-Computer Interaction, Neuroethics*, and *AI and Society*.[71]

Recurring ethical and social issues discussed in this literature include: the psychological and emotional impacts of different human-computer interactions, as well as their impacts on different parts of wider society such as the economy and labour markets, and concerns about human agency, responsibility, autonomy, dignity, privacy, and responsibility.

One interesting aspect of this literature is some attention to how the consequences of human-computer interaction will differ depending on the types of people affected, types of technology used, and the contexts of the interaction. However, more attention could be given to the question of how each particular kind of technology might affect different demographics differently, and how the implications of these interactions may differ depending on the nature and context of the interaction.

1e. Political and social sciences

We surveyed the large and growing literature across politics, economics and social science that discusses how algorithms, AI and data will impact society more broadly.

The main issues covered in this literature include: how ADA will impact economic growth and disrupt the economy in general, the impact on jobs and the labour market more specifically,[72] and experts' attempts to predict how automation will affect jobs, how quickly, and what policy responses to technological unemployment will be needed (including training, education, and redistribution schemes such as universal basic income). Another key issue is the impact of ADA on global inequality and prosperity, with concerns being raised that technology many widen the gap between developed and developing countries.[73]

Finally, there is a discussion around how ADA will impact national and international politics: what politics, power, and democracy will look like

[68] See, for example, Goodman and Flaxman (2016); Wachter, Mittelstadt, and Floridi (2017); Selbst and Powles (2017).

[69] Edwards and Veale (2017).

[70] Kuner et al. (2017).

[71] Key papers include Becker (2006); Clark (2008); Jotterand and Dubljevic (2016); Menary (2007); Parens (1998); Schermer (2013); Sharkey (2014).

[72] Frey (2017); Kaplan (2015); Marcus (2012); Marien (2014).

[73] Eubanks (2018).

in an increasingly ADA-controlled society,[74] and how the use of autonomous weapons and the risk of an AI 'arms race' might threaten international stability.

1f. Other cross-disciplinary literature

Finally, we looked at how ethical issues around ADA are discussed at a cross-disciplinary level in publications spanning philosophy, law, and machine learning.[75]

There is a substantial amount of cross-citation across these different fields, and focus on a shared set of key concepts, particularly those of fairness, accountability, and transparency, but also bias, discrimination, explainability, privacy, and security.

However, these key terms are often used unreflectively, in different and inconsistent ways, and without much further analysis – for example proposing methods to increase 'transparency' without clarifying what this means or why it is important.

2. Popular science and media

We surveyed how ethical and social issues relating to algorithms, data and AI have been discussed in the media and in popular science literature, as these issues have received increasing public attention.

2a. Popular science books

Looking at a range of popular science books on artificial intelligence, we found that two topics arose particularly prominently: (a) the risks posed by potential 'superintelligence' and particularly the challenge of aligning advanced AI with human values, and (b) the future of work and potential disruption of automation. Other issues covered less prominently include whether a machine can be held responsible for its actions or given rights, and how to prevent the use of big data from increasing inequality and leading to abuses of power.[76]

2b. Popular media

Popular media articles have recently begun to focus much more on the risks posed by current uses of algorithms and AI, particularly the potential for bias against underrepresented communities, and threats to personal privacy from the use of large amounts of personal data.[77]

It is not easy to measure how relevant these concerns are becoming in terms of popularity, but some indicators can help us get a rough idea of the relative growth. For instance, figure 4 shows the percentage of documents from the 'AI topics' database related to ethics of AI or its impacts.

A similar perspective can be seen from the articles in the New York Times, as analysed by Fast and Horvitz (2017), who find that AI discussion has exploded since 2009, but levels of pessimism and optimism have remained balanced.

3. Policy reports and wider international landscape

Various organisations and government

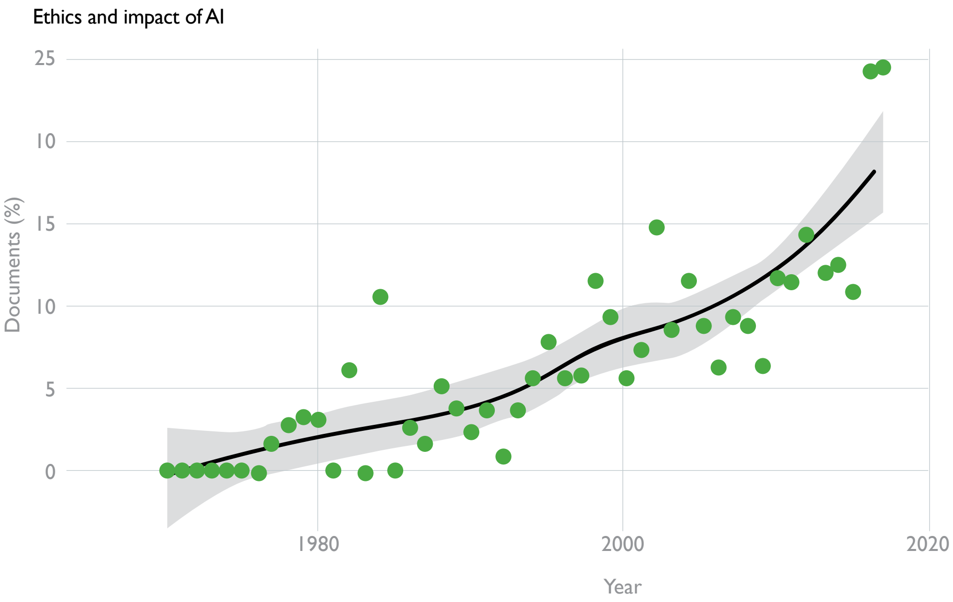


Figure 4. Percentage of articles (green dots) in the 'AI topics' having at least one keyword related to ethics and impact of AI. General tendencies (black line) and standard errors (grey band) are also shown. The methodology is explained in full in Martínez-Plumed et al. "The Facets of AI", IJCAI 2018.

[74] See for example Monbiot (2017) and Helbing et al. (2017).

[75] See for example Lipton (2017), Weller (2017), Binns (2018), Tene and Polonetsky (2017), and Selbst and Barocas (2016).

[76] Key books covered include Barrat (2013); Bostrom (2014); Brynjolfsson and McAfee (2014); Chace (2015); Collins (2108); Eubanks (2018); Harari (2015); Kurzweil (2012); McFarland (2008); Moravec (1988); Noble (2018); O'Connell (2017); O'Neil (2016); Shanahan (2015); Tegmark (2017); Wachter-Boettcher (2017); Wallach and Allen (2009); Walsh (2017); Zarkadakis (2015).

[77] See for example Crawford (2016); Li (2018); Shariff, Rahwan and Bonnefon (2016); Mukherjee (2017); Kleinberg, Ludwig and Mullainathan (2016).

institutions in the UK, US, and across the world have also begun mapping out the ethical and societal questions posed by the use of algorithms, data and AI, with a focus on identifying their practical implications for policy. We looked at how these policy-focused reports discuss the issues in this space, with a particular focus on any international differences.[78]

These policy-focused reports naturally look at a very wide range of issues, covering most of those we found in different parts of the academic literature. There was particular focus on the following issues: data management and use, fairness, bias, statistical stereotyping, transparency, interpretability, responsibility, accountability, the future of work, and economic impact.

However, reports from different parts of the world focused their attention on different issues. In developed countries, there is greater focus on the safe deployment and potential risks of technology than in other parts of the world. In developing countries, the conversation focuses more on building capacity and an ecosystem around technology and research.

4. Public engagement research

Public engagement is an attempt to involve members of different publics into agenda setting and decision-making on various matters of policy, science, medicine, and technology. They involve a variety of methods of polling, survey, consultation and citizen fora. So far several survey initiatives have mapped aspects of ADA understanding, acceptance, and concerns surrounding AI in the UK, including:

- Ipsos MORI and the Royal Society (2016/2017) carried out the first public dialogues on machine learning in the UK, looking at public attitudes in health, social care, marketing, transport, finance, policing, crime, education, and art. The initiative made use of dialogue discussions and surveys, and revealed that only 9% of the people surveyed had heard of machine learning before.[79]

- The RSA (2017) looked at ways to engage citizens in the ethical deployment of AI technologies, and found that very few people know the extent to which automated decision-making influences their lives.[80]

- The Cabinet Office (2016) investigated how the public weigh up risk around machine learning and AI when applied to administrative decisions, by means of a conjoint survey and a data game. They too found that public awareness of data science is limited.[81]

- The Wellcome Trust (2016) queried the public acceptability of commercial access to patient data, by means of a large public dialogue. This was followed up by a quantitative survey. Their work revealed that without a clear public benefit, the public are very concerned about the idea of commercial access to healthcare data.[82]

- The Health Research Authority and Human Tissue Authority facilitated three location dialogues with public and scientist stakeholders about consent to sharing data and tissue in a world of genomics. Participants expressed the worry that by giving consent now to future uses of their data, if legislation changes they might be unwittingly contributing to a two-tier society where people can be discriminated against based on their genome.[83]

- The Academy of Medical Sciences (forthcoming 2018) is engaged in public and stakeholder dialogue on the role of beneficial AI in healthcare.

- DeepMind (2018) organized a public and stakeholder engagement initiative to develop

[78] Including EU EDPS Advisory Group (2018); Future Advocacy and the Wellcome Trust (2018); Government Office for Science (2016); IEEE (2018); Omidyar Network and Upturn (2018); National Science and Technology Council (2016); Royal Society (2017); Royal Society and the British Academy (2017); Select Committee on Artificial Intelligence (2018).

[79] Ipsos MORI and the Royal Society, (2016); Ipsos MORI and the Royal Society. (2017).

[80] Royal Society for the encouragement of Arts, Manufactures and Commerce (RSA), (2018).

[81] Cabinet Office. Public dialogue on the ethics of data science in government. (2016). www.ipsos.com/sites/default/files/2017-05/data-science-ethics-in-government.pdf

[82] Wellcome Trust. Public attitudes to commercial access to patient data. (2016). www.ipsos.com/sites/default/files/publication/5200-03/sri-wellcome-trust-commercial-access-to-health-data.pdf

[83] www.hra.nhs.uk/about-us/what-we-do/how-involve-public-our-work/what-patients-and-public-think-about-health-research/

the principles and values which should drive its behaviour globally.[84]

5. Summary

At a general level, we identified the following commonly discussed issues emerging from these literatures:

- Ensuring that 'black box' algorithms and AI systems are **transparent / explainable / interpretable.**
- Ensuring that uses of ADA are **reliable** and **robust.**
- Maintaining individual **privacy** and **protection of personal data.**
- Ensuring algorithms and AI systems are used **fairly** and do not reflect historical **bias**, or lead to new forms of bias or **discrimination.**
- Ensuring algorithms and AI **reflect human values** more broadly.
- The question of whether AI systems can ever **make moral decisions.**
- The question of whether AI systems might ever **attain moral status.**
- Issues of **accountability**, **responsibility** and **liability** around the use of ADA.
- The role of **ethics statements** or **ethical training** in ensuring responsible use of ADA.
- The role of **law and regulation** in mitigating the risks and ensuring the benefits of AI.
- Building appropriate levels of **trust** among humans and machines algorithms.
- Implications of ADA for **human agency**, **autonomy**, and **dignity.**
- Impact of ADA on the **economy** and economic growth.
- Impact of ADA on **jobs and labour markets**, developing policies around **technological unemployment.**
- Impact of ADA on **global inequality.**
- Impact of ADA on **national politics**, **public opinion**, and **democracy** – including how unemployment might lead to disruptive changes in public opinion.
- How ADA changes **power** in a society.
- How ADA might be used to **direct our attention** or **manipulate** opinions (e.g. for political or commercial purposes).
- Impact of ADA on **international relations**, **conflict**, and **security** – including impact of **autonomous weapons** and risk of a **global arms race.**
- What new methods of **global governance** might be needed to deal with the challenges posed by increasingly powerful technologies.

[84] https://deepmind.com/applied/deepmind-health/transparency-independent-reviewers/developing-our-values/#image-27248

Appendix 2: Groupings and principles

Below we list some of the common ways that organisations have so far grouped and structured issues relating to ADA, as well as various sets of principles that have been proposed.

One dividing line is between those approaches that give fairly long lists of principles (Asilomar, Partnership on AI), and those that use as few as four categories (e.g. the AI Now 2017 Report). There are advantages and disadvantages to both: the shorter lists can be useful for providing a simple overview of a complex field, but risk conflating importantly different issues or leaving out important themes. They can only aim to be comprehensive by making the categories broad and unspecific. Longer lists, on the other hand, can aim to be more comprehensive and to capture more nuance, but risk losing a clear overview, and may include categories that overlap in non-perspicuous ways.

A strategy for trying to balance these two approaches is to draw on a broader analytical framework. For example, the EDPS Ethics Advisory Group propose to derive the most important issues from eight 'European' values, while Cowls and Floridi (2018) propose that all relevant issues can be captured as resulting from technologies being either overused, misused or underused relative to 'four fundamental points in the understanding of human dignity and flourishing: *who we can become* (autonomous self-realisation); *what we can do* (human agency); *what we can achieve* (societal capabilities); and *how we can interact with each other and the world* (societal cohesion).' (p.1).

While these frameworks can strike a balance between complexity and systematicity, they still carry the risk of leaving out or downplaying some issues. For instance, it is not immediately clear where issues of bias and discrimination should figure in Cowls and Floridi's list. Furthermore, systematic frameworks of this kind generally presuppose a judgment of what the fundamental values are that should structure the framework. This can be useful in contexts where there is a prior commitment to such a set of values (as one may be able to do with regards to the 'European values' in the context of the European Union), but agreement on such value judgments cannot be universally presumed.

Different ways of carving up the space of ADA ethics and societal impacts can serve different purposes – for example providing an overview, capturing all relevant issues, or providing practically relevant advice. The different frameworks surveyed below can all be useful for these purposes. It is doubtful that a single framework could capture the entire field and serve all purposes, and this is neither necessary nor sufficient for making constructive progress on these issues (although, of course, it might be useful for any given organisation or community to settle on such principles). Instead, efforts to map and organise the relevant issues should be understood as contextually useful tools for specific purposes.

1. Common ways of organising issues

The AI Now 2017 Report identifies four focus areas:

1. Labor and Automation
2. Bias and Inclusion
3. Rights and Liberties
4. Ethics and Society

DeepMind Ethics and Society split their work into six research themes:

1. Privacy, transparency, and fairness
2. Economic impact, inclusion, and equality
3. Governance and accountability
4. AI morality and values
5. Managing AI risk, misuse, and unintended consequences
6. AI and the world's complex challenges

The Partnership on AI uses a breakdown into six 'thematic pillars':

1. Safety-critical AI
2. Fair, transparent, and accountable AI
3. Collaborations between people and AI systems
4. AI, labor, and the economy
5. Social and societal influences of AI
6. AI and social good

The EDPS Ethics Advisory Group highlights seven key 'socio-cultural shifts of the digital age':

1. From the individual to the digital subject
2. From analogue to digital life
3. From governance by institutions to governmentality through data
4. From a risk society to scored society
5. From human autonomy to the convergence of humans and machines
6. From individual responsibility to distributed responsibility
7. From criminal justice to pre-emptive justice

And consider the impact of digital technologies on the following values:

1. Dignity
2. Freedom
3. Autonomy
4. Solidarity
5. Equality
6. Democracy
7. Justice
8. Trust

The Obama White House report, 'Preparing for the future of artificial intelligence' divides its discussion into the following sections:

1. Applications of AI for public good
2. AI in government
3. AI and regulation
4. Research and workforce
5. AI, automation, and the economy
6. Fairness, safety, and governance
7. Global considerations and security

The Royal Society and British Academy joint report (2017) uses the following categories:

1. Safety, security, prevention of harm
2. Human moral responsibility
3. Governance, regulation, monitoring, testing, certification
4. Democratic decision-making
5. Explainability and transparency

The European Group on Ethics in Science and New Technologies presents the following breakdown of issues:

1. Privacy and consent
2. Fairness and statistical stereotyping
3. Interpretability and transparency
4. Responsibility and accountability
5. Personalisation, bubbles, and manipulation
6. Power asymmetries and inequalities
7. Future of work and the economy
8. Human-machine interaction

2. Principles and codes[85]

The Asilomar AI Principles include the following 'ethics and values' principles:[86]

- Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
- Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.
- Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.
- Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
- Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
- Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
- Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
- Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
- Shared Benefit: AI technologies should benefit and empower as many people as possible.
- Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
- Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
- Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.
- AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Partnership on AI members 'believe in and endeavour to uphold the following tenets':

1. We will seek to ensure AI technologies benefit and empower as many people as possible.

2. We will educate and listen to the public and actively engage stakeholders to seek their feedback on our focus, inform them of our work, and address their questions.

3. We are committed to open research and dialogue on the ethical, social, economic and legal implications of AI.

4. We believe that AI research and development efforts need to be actively engaged with and accountable to a broad range of stakeholders.

5. We will engage with and have representation from stakeholders

[85] These summarise the key aspects of various principles and codes, but do not necessarily represent the principles in full – sometimes just using the title and not the full explanation of each principle, for example.

[86] The full Asilomar Principles include a further ten principles on 'research issues' and 'longer-term issues' which we do not include here.

in the business community to help ensure that domain-specific concerns and opportunities are understood and addressed.

6. We will work to maximise the benefits and address the potential challenges of AI technologies, by:

   a. Working to protect the privacy and security of individuals.

   b. Striving to understand and respect the interests of all parties that may be impacted by AI advances.

   c. Working to ensure that AI research and engineering communities remain socially responsible, sensitive, and engaged directly with the potential influences of AI technologies on wider society.

   d. Ensuring that AI research and technology is robust, reliable, trustworthy, and operates within secure constraints.

   e. Opposing development and use of AI technologies that would violate international conventions or human rights, and promoting safeguards and technologies that do no harm.

7. We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.

8. We strive to create a culture of cooperation, trust, and openness among AI scientists and engineers to help us all better achieve these goals.

The **Lords Select Committee on AI report** suggests five principles for a cross-sector AI code:

1. AI should be developed for the common good and benefit of humanity.

2. AI should operate on principles of intelligibility and fairness.

3. AI should not be used to diminish the data rights or privacy of individuals, families, or communities.

4. All citizens should have the right to be educated to enable them to flourish mentally and economically alongside artificial intelligence.

5. The autonomous power to hurt, destroy, or deceive human beings should never be vested in AI.

The IEEE Standards Association has also developed a set of general principles to guide ethical governance of 'autonomous and intelligent systems':

1. Human rights
2. Prioritising well-being
3. Accountability
4. Transparency
5. Technology misuse and awareness of it

The Association for Computing Machinery (ACM)'s 'Principles for Algorithmic Transparency and Accountability':[87]

1. Awareness
2. Access and redress
3. Accountability
4. Explanation
5. Data Provenance
6. Auditability
7. Validation and Testing

The **Japanese Society for Artificial Intelligence** (JSAI) Ethical Guidelines:[88]

1. Contribution to humanity
2. Abidance of laws and regulations
3. Respect for the privacy of others
4. Fairness
5. Security
6. Act with integrity
7. Accountability and Social Responsibility
8. Communication with society and self-development
9. Abidance of ethics guidelines by AI

The Future Society's Science, Law and Society Initiative – Principles for the Governance of AI:[89]

1. AI shall not impair, and, where possible, shall advance the **equality in rights**, **dignity,** and **freedom to flourish** of all humans.

2. AI shall be **transparent**.

3. Manufacturers and operators of AI shall be **accountable**.

4. AI's effectiveness shall be **measurable** in the real-world applications for which it is intended.

5. Operators of AI systems shall have appropriate competencies and **expertise**.

6. The norms of delegation of decisions to AI systems shall

[87] See the ASM US Public Policy Council's 'Statement on Algorithmic Transparency and Accountability' (2017)

[88] http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf

[89] www.thefuturesociety.org/science-law-society-sls-initiative/#1516790384127-3ea0ef44-2aae

be codified through thoughtful, **inclusive dialogue** with civil society.

UNI Global Union's 'Top 10 Principles for Ethical Artificial Intelligence':[90]

1. Demand that ai systems are transparent.
2. Equip ai systems with an 'ethical black box'.
3. Make AI serve people and planet.
4. Adopt a human-in-command approach.
5. Ensure a genderless, unbiased AI.
6. Share the benefits of AI systems.
7. Secure a just transition and ensuring support for fundamental freedoms and rights.
8. Establish global governance mechanisms.
9. Ban the attribution of responsibility to robots.
10. Ban AI arms race.

The Montréal Declaration for Responsible AI[91] consists of the following principles:

1. Well-being
2. Respect for autonomy
3. Protection of privacy and intimacy
4. Solidarity
5. Democratic participation
6. Equity
7. Diversity inclusion
8. Prudence
9. Responsibility
10. Sustainable development

[90] www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

[91] www.montrealdeclaration-responsibleai.com/the-declaration

Appendix 3: Different perspectives

The entire space of ethical and societal impacts of ADA is large, highly complex and unlikely to be captured within a single unified framework (cf. section 2 and appendix 2). This makes it difficult to understand individual issues without first zooming-in and filtering out some information. Conversely, it is easy for important aspects or dimensions of a given issue to be overlooked or get lost in the complexity of the space.

This appendix outlines a number of salient perspectives one can take on the space of ADA ethical and societal importance, with examples of the subdivisions one might make within each. These can be thought of as axes, along which one can zoom-in on different parts of the space to filter out information. These perspectives can be used singly or in combination to either restrict the range of issues considered, or to think through a single issue from several different perspectives to ensure that as many relevant aspects as possible are considered.

1. Which sectors or parts of society?

Societies consist of a number of different sectors or parts, as reflected in the ways governments divide their administrations into different departments. Using the US and UK government departments as a template, one might for example focus on how ADA impacts:

Agriculture
Business
Culture, media, and sports
Energy
Education
Environment
Development
Trade
International and institutional relations
Transport
Work and labour
Health and social care
Finance and the economy
Security and defence
Community and housing
Crime and justice

2. Which level of social organisation?

Human relations are structured into a number of different levels of social organisation, from the local community to global international relations. In addition to looking at issues by governmental sector, one can focus on issues that arise at different levels of social organisation (figure 5).

3. Which time-frame?

Issues relating to ADA may emerge at different time-scale. For instance, we may distinguish:

1. **Present challenges:** What are the challenges we are already aware of and already facing today?

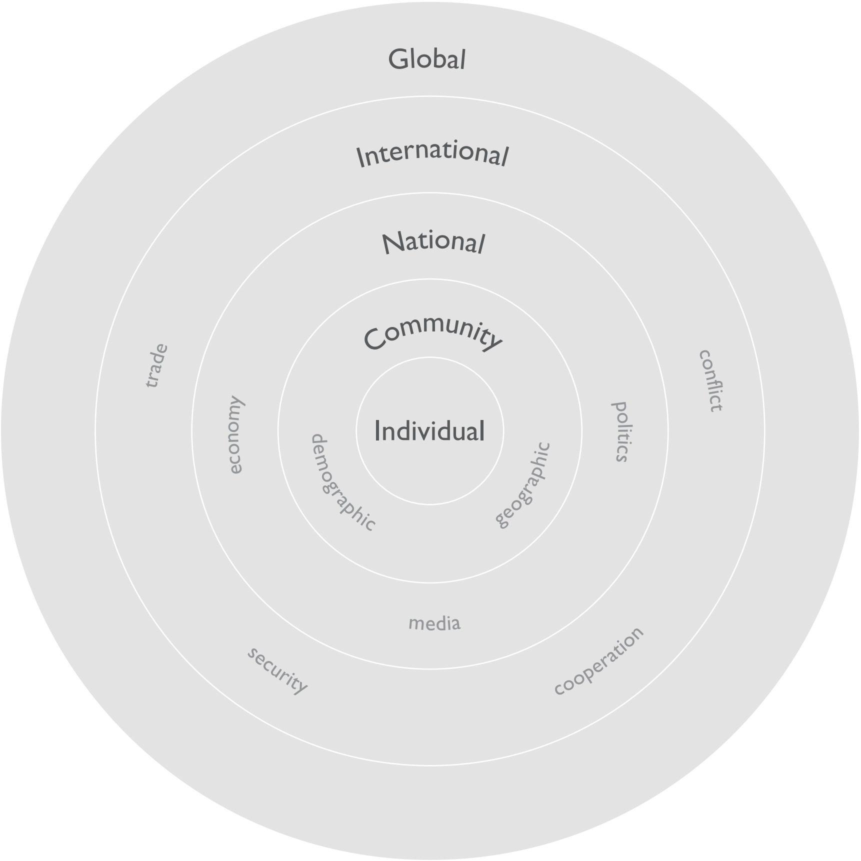


Figure 5. Different levels and parts of society which may be impacted by ADA-based technologies.

2. **Near-future challenges:** What challenges might we face in the near future, assuming current technology?

3. **Long-run challenges:** What challenges might we face in the longer-run, as technology becomes more advanced?

Thinking about challenges in the first category is the easiest, as they are ones that are currently in front of us and discussed: ensuring the privacy of personal data, for example. Challenges in the second category require more thought, to imagine how current technologies might pose new challenges in future. An example might be how current image synthesis techniques could be put to malicious use. Challenges in the third category are the most difficult to forecast, since they require thinking about the impact of future technological capabilities. Discussion of the potential impacts of superintelligent AI would fall into this category.

4. Which publics?

Different publics concern themselves with different problems and have different perspectives on the same issues. The following distinctions between publics and their corresponding questions about ADA technologies can inform how one organises inquiry into moral relevance of these technologies:

- **Designers and engineers**: What responsibilities do I have to ensure the technology I am developing is ethical? What ethical standards need to be met? How can demands of technology like privacy, fairness, and transparency be made technically precise?

- **Users/general public**: How does a given technology impact my day-to-day life? What new trade-offs does it introduce for me?

- **Marginalised groups:** Is this technology a threat or an opportunity given our precarious status? How can we use it to fight prejudice?

- **Organisations and corporate bodies**: What are the costs and benefits of automating a given service/task? What do we need to think about to ensure our use of technology is ethical/trustworthy?

- **Policymakers and regulators:** Where is policy or regulation needed? Where is there public pressure to act?

- **NGOs and civil society:** How can we ensure widespread public engagement on these issues? Whose interests might not be being represented?

- **Researchers:** What intellectually interesting questions do the use of new technologies raise? What issues require deeper intellectual thought?

- **Journalists and communicators:** What aspects of technology and their impacts on society most need to be communicated to different publics? How can these issues be communicated most effectively?

5. What type of challenge?

When a piece of technology goes wrong it can do so for different reasons. We might consider different types of challenges that can arise from:

- How technology is **designed** or **developed** (e.g. what biases might exist in the data used to train an algorithm, inability to examine how exactly an algorithm uses different features to make a decision).

- **Externalities** or **unintended consequences** of how technology is applied in society (e.g. impact of automation on the labour market, issues of liability as algorithms are used to make decisions in healthcare and other important areas).

- Consequences of **failure to perform** or **accidents** (e.g. self-driving car accidents).

- **Malicious use** of technology (e.g. for surveillance, manipulation or crime).

6. What type of solution?

We have many different mechanisms and methods at our disposal for addressing the various ethical and social challenges arising from the use of ADA. Different types of solution will be needed for different types of problem, and multiple different approaches will be needed to solve most challenges. Thinking more systematically about the different methods available for tackling these problems could help to identify new approaches and angles.

For example, we might break down different types of solution as follows:

- Law and regulation: national and international.
- Government policy.
- Public engagement and education.
- Activism.
- Different types of research:
  – Technical research
  – Philosophy/ethics/ humanities research
  – Social science research.

Published by the Nuffield Foundation, 28 Bedford Square, London WC1B 3JS

This report is also available to download at **www.nuffieldfoundation.org**

Designed by Soapbox
www.soapbox.co.uk