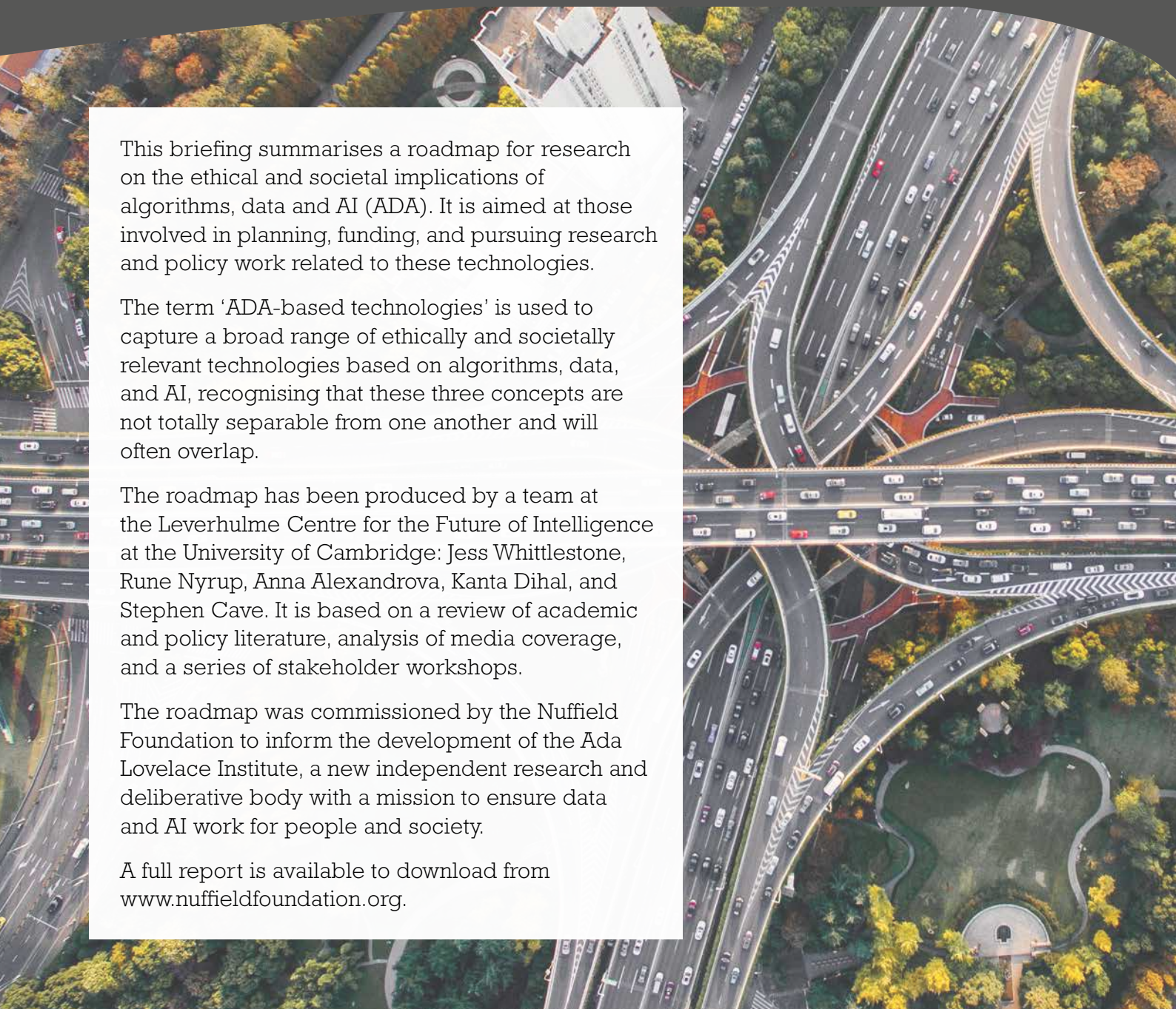
Nuffield Foundation

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

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

Summary report



This briefing summarises a roadmap for research on the ethical and societal implications of algorithms, data and AI (ADA). It is aimed at those involved in planning, funding, and pursuing research and policy work related to these technologies.

The term 'ADA-based technologies' is used 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.

The roadmap has been produced by a team at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge: Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave. It is based on a review of academic and policy literature, analysis of media coverage, and a series of stakeholder workshops.

The roadmap was commissioned by the Nuffield Foundation to inform the development of the Ada Lovelace Institute, a new independent research and deliberative body with a mission to ensure data and AI work for people and society.

A full report is available to download from www.nuffieldfoundation.org.

Three key tasks

The roadmap presents an exploration of current research and debates on ethical and societal impacts of algorithms, data, and AI, and identifies what has been achieved so far and what needs to be done next.

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) on which an ethics of algorithms, data, and AI should focus. 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, our work has identified three main gaps:

1. A lack of clarity or consensus around the meaning of central ethical concepts and how they apply in specific situations.

2. Insufficient attention given to tensions between ideals and values.

3. Insufficient evidence on both (a) key technological capabilities and impacts, and (b) the perspectives of different publics.

In order to address these gaps, and to move the discussion forward, we recommend prioritising three key tasks.

Task 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.

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

a. Exploring concrete instances of tensions central to current applications of ADA. We have identified four central tensions:

   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, disability, 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.

Task 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.

In this briefing, we summarise our suggested questions for research relevant to achieving each of these three tasks. These are by no means exhaustive, but they highlight areas where we believe there is strong potential for 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 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 four central tensions described in Task 2. This challenge requires 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 the 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 **versus** 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 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 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?

About the Nuffield Foundation

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. We fund research that informs social policy, primarily in Education, Welfare, and Justice. We also fund student programmes that provide opportunities for young people to develop their skills and confidence in quantitative and scientific methods.

We have 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.

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

Registered charity 206601

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

Published by the Nuffield Foundation, 28 Bedford Square, London WC1B 3JS

This briefing paper is also available to download at **www.nuffieldfoundation.org**

Designed by Soapbox
www.soapbox.co.uk