Ipsos MORI
Social Research Institute

April 2017

Public views of Machine Learning

Findings from public research and engagement conducted on behalf of the Royal Society

Ipsos MORI

THE ROYAL SOCIETY

Contents

List of figures

List of tables

1 Executive summary

1.1 Background and objectives

Machine learning is a way of programming a system to learn from data and self-improve. Traditionally, programmers set static instructions to tell a computer how to solve a problem, step by step. In contrast, machine learning algorithms can identify patterns in data and use this information to learn how to solve the problem at hand. Machine learning algorithms enable the analysis of much larger quantities of data than a human could work with, and, as a result, can identify complex patterns or relationships. The models built on the basis of this analysis can then be used to make predictions or decisions.

The Royal Society launched a project on machine learning in November 2015, which aims to increase awareness of this technology, demonstrate its potential, and highlight the opportunities and challenges it presents. The project's focus is on the current and near-term (5-10 years) applications of machine learning. The UK public is a key audience for this project, and public engagement is an integral part of the programme of work. The Royal Society therefore commissioned Ipsos MORI to carry out research into public knowledge of, and attitudes towards, machine learning.

Ipsos MORI's task for the research was to create an evidence base about public perceptions around the potential benefits and risks of the technology, to inform the Royal Society's policy project on machine learning. Exploring these issues required an approach involving depth, breadth and iterative engagement. As such, the methodology used in this research was designed to incorporate three elements: a quantitative survey, public dialogues, and an online community.

Between 22 January and 8 February 2016, 978 face-to-face interviews were conducted with members of the public across the UK. All interviews were carried out in-home, using computer-assisted personal interviewing (CAPI) on Ipsos MORI's weekly omnibus survey. Final data was weighted to ensure the individuals selected for interview were representative of the national population.[1] This was followed up by qualitative research, which involved two weekend-long dialogue events in Birmingham and London, along with two evening focus groups in Oxford and Huddersfield.

1.2 Reactions to machine learning

Most participants were not familiar with the term 'machine learning' and found it easier to engage with the idea through real-life examples. Most had come across at least some specific applications of machine learning in their day-to-day experiences. Indeed, the quantitative survey found that people were much more likely to have heard of at least one of the examples of machine learning applications than they were to have heard of the term 'machine learning'.

The workshop participants were introduced to machine learning through a series of examples, focusing on the areas of: health, social care, marketing, transport, finance, crime, education, and art. In general, they were not concerned with the detail of how machine learning works, focusing instead on how and why it could be used in different contexts.

[1] Please see appendix (A.2) for the quantitative sample breakdown

Four different spontaneous reactions to machine learning were observed among these participants:

- 'I can personally relate to this technology, because I can see where this could have an impact on my life, whether good or bad'
- 'This is an important emerging technology and it carries potential risks and benefits to society'
- 'I can't see how this would work – humans are too unique for machines to really understand us'
- 'I'm suspicious about the purpose of this technology'

All participants could relate personally to the technology, as they could see where it could have an impact on at least some areas of their life. Their views developed over time as participants explored different machine learning applications in more detail, but these spontaneous reactions continued to be important in shaping their more considered opinions.

1.3 Considering the risks and benefits of machine learning

Participants generally took a pragmatic approach to how machine learning should be applied. They discussed the intended purposes, perceived motivations of those using the technology, and the consequences for individuals and society.

Workshop participants used the following criteria for deciding whether they liked an application in principle:

- The perceived intention behind using the technology in a particular context → Participants typically wanted to understand who would be involved with the development of machine learning applications. They felt that the motives and intentions of those involved might shape the success, and direction, of the technology as it progresses.
- Who the beneficiaries would be → Participants were more positive about machine learning when they thought there would be worthwhile benefits for individuals, groups of people, or society as a whole. They were less positive when they could only see machine learning applications serving private interests.
- How necessary it was to use machine learning → Many participants struggled to see why machine learning was necessary in some contexts. This was particularly the case where humans were seen as being as good as or better than a machine at completing the task.
- How appropriate it was for machine learning to be used → Participants felt that machine learning was inappropriate in some circumstances, particularly when it involved the loss of valuable human-to-human contact.
- Whether or not a machine will need to make an autonomous decision → If an application would involve a machine making a decision, the seriousness of the potential consequences of that decision was key in assessing the application.

If they felt a particular use of machine learning was desirable and appropriate, based on these principles, participants then weighed up the detailed benefits and risks to decide whether they could support it or not. The quantitative survey found an even split between those who thought that overall the benefits of machine learning outweigh the risks, and those who thought the opposite.

Participants' assessments of the risks and benefits were often instinctive and nuanced, with their views on one of these criteria sometimes dominating. In other cases, they balanced different criteria, and did not all agree. Four main types of risk and four main types of benefit emerged throughout the discussions:

Types of risks associated with machine learning:

- 'This technology could harm me and others'
- 'This technology could replace me'
- 'This technology could depersonalise me and my experiences'; and,
- 'This technology could restrict me'.

Types of benefits associated with machine learning:

- 'This technology has a lot of potential to benefit individuals and society'
- 'This technology could save a lot of time; and,
- 'This technology could give me better choices'.

The risks were usually easier for participants to identify, and they spent considerable time discussing these in the context of the different examples considered. Despite their concerns, participants recognised the opportunities associated with machine learning and the potential for significant benefits for individuals and society.

1.4 Views of specific machine learning applications

Health: The use of machine learning in health was where participants could intuitively see the greatest potential for benefits to individuals and society. They felt that it could improve accuracy as machines would be able to consider more data when making diagnoses than humans. However, they stressed the need for human doctors to remain involved, to ensure personal contact continues where it is needed.

Social care: On the one hand, participants saw the potential of machine learning to help with resourcing issues in the sector. On the other hand, they feared an over-reliance on machines would lead to reduced human involvement and emotional contact. Participants tended to envisage a best-case scenario where machines would perform tasks that would enable human carers to spend more time with patients.

Marketing: Participants were not generally aware that machine learning is already used to tailoring marketing online. Concerns centred around manipulation and increased spending, and an invasion of privacy. The minority who were more positive felt that it was better to have relevant adverts and offers that people might want to take advantage of.

Transport: Driverless cars were seen as having benefits, by offering independence to those who are unable to drive, and by leading to more efficient travel through uniform driving. However, some participants had strong reservations about the ability of an algorithm to adapt to road conditions and to deal specifically with sudden changes. They wanted clear evidence that driverless vehicles would be safe.

Finance: Participants were universally supportive of algorithms being used to monitor potentially fraudulent activity. However, they were much more hesitant about the idea of algorithms warning individuals about spending based on past behaviour or current financial circumstances.

Crime: Participants tended to think that using machine learning to spot patterns in crime was a good idea in principle, but struggled to see how it might work accurately in practice. They saw it as a useful tool to aid with limited police resources, but were also concerned about the consequences of stereotyping individuals or groups.

Education: Some participants were concerned that tailored education based on machine learning would result in de-skilling and limiting people to certain career paths at too young an age. However, the majority felt that tailored learning was a positive. They saw the potential of machine learning to spot patterns in attainment, attendance and general attitude, to flag any issues for teachers to address.

Art: Participants failed to see the purpose of machine learning-written poetry. For all the other case studies, participants recognised that a machine might be able to do a better job than a human. However, they did not think this would be the case when creating art, as doing so was considered to be a fundamentally human activity that machines could only mimic at best.

1.5 Machine learning in practice – the policy context

Participants generally found it hard to discuss the ethics around and regulation of machine learning, other than recognising that it was important to ensure the risks of this technology were considered carefully. In part, this was because their focus had been on discussing the acceptability of a range of potential applications in varied contexts, rather than on the specifics of how machine learning technology itself works.

Discussing the ethical framework which should govern machine learning was also challenging, particularly when it came to the safeguards that would be needed if machines make important decisions independent of human involvement. It was difficult to imagine how a machine could behave 'ethically' because of the subjectivity they thought was involved in ethical judgments. As such, their views of the ethics of machine learning often returned to the extent to which humans would still be involved in the process.

While regulation was considered important, there was no clear consensus about what this should look like in practice. It was felt that the technology should not be allowed to advance without oversight, to ensure that it was not being abused or was not being portrayed as accurate, if this was not the case.

The breadth of possible machine learning applications made it hard for participants to come to a general view about regulation. Most expected some government involvement, but tended to prefer an independent regulator or regulators funded by – but ultimately separate to – government. They also highlighted broader regulatory issues related to machine learning, including where agencies or companies are passing data to one another.

Participants generally assumed that government would have a role in research around machine learning, but expected that the technology that drives it will mostly develop commercially. However, where possible, they felt the two sectors should work together in its development.

The survey found a similar mix of views. While most thought there should be a role for government in the development and regulation of machine learning, there was less consensus about what this should look like in practice. There were also different perspectives on who should be held responsible when machine learning goes wrong. The two most common answers were that the organisation the operator and machine work for should be to blame, followed by the manufacturer. Few would hold other individuals or organisations involved with machine learning responsible.

2 Introduction

2.1 About the Royal Society

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- promoting excellence in science;
- supporting international collaboration; and
- demonstrating the importance of science to everyone.

The Society provides expert, independent advice to policy-makers and the public, championing the contributions that science can make to economic prosperity, quality of life and environmental sustainability.

With the expertise of their Fellowship, the Royal Society uses high quality science to guide and develop policy studies, rapid reports and consultation responses, with the aim of informing policy developments on important topics like health and well-being, security and risk, and energy and environment.

The Society also provides a forum for debate, bringing together diverse audiences to discuss the impact of science on current and emerging policy issues.

2.2 Background to the project

Machine learning is a way of programming a computer system to learn from data and self-improve. Traditionally, programmers set static instructions to tell a computer how to solve a problem, step by step. In contrast, machine learning algorithms can identify patterns in data and use this information to learn how to solve the problem at hand. Machine learning algorithms enable the analysis of much larger quantities of data than a human could work with, and, as a result, can identify complex patterns or relationships. The models built on the basis of this analysis can then be used to make predictions or decisions.

The Royal Society launched a project on machine learning in November 2015, which aims to increase awareness of this technology, demonstrate its potential, and highlight the opportunities and challenges machine learning presents. The project's focus is on the current and near-term (5-10 years) applications of machine learning. The UK public is a key audience for this project, and public engagement is an integral part of its programme of work. The Royal Society therefore commissioned Ipsos MORI to carry out research into public knowledge of, and attitudes towards, machine learning.

2.3 Previous research on public attitudes to emerging technology

There has been very little exploration of the public's views of machine learning. Surveys and qualitative studies of public understanding of emerging, and often data driven, technologies have mainly focused on robotic technology and

autonomous systems. Many more studies have explored public perceptions of the collection, storage and use of personal data.

2.3.1 Public attitudes about emerging technologies

Broadly speaking, the public are supportive of science and scientific developments and want to know more about them. Four in five people believe that science makes people's lives easier (81%) and 55% think that the benefits outweigh any harmful effects. Clear majorities think that scientific research that advances knowledge should be government-funded, even if it brings no immediate benefits (79%). However, most people do not feel informed about new technologies (55%).[2]

When it comes to robotics, the public favours the use of robots in situations that might be dangerous to humans, but public support falls when the context changes to a more personal one. For example, 87% of people support the use of robotics in space exploration, 81% in manufacturing and 72% for military purposes. This falls to 18% of the public being in favour of robots being used to care for the elderly and 14% for robots being used to care for children. In these latter examples, fears over the loss of human-to-human contact are often cited.[3]

2.3.2 Public attitudes about data

Public awareness of the potential uses of large datasets is low – including how much data they generate (and how quickly) in their personal lives and how individual records could be aggregated and analysed to produce insights by government to improve public services, for example.[4] Despite this, three in five people do not mind how their personal data is used, as long as it is anonymised (61%)[5] and around three-quarters of the public are willing to share their anonymised medical records (77%), or their anonymised genetic information (75%), for the purposes of a medical research study.[6] However, support for the use of personal data appears to be conditional; people are more concerned about their data being used for commercial purposes, preferring uses that result in tangible public benefits, such as improvements in the health sector, transport and crime prevention.[7,8]

Despite broad support among the public for their anonymised data to be used to improve public services, awareness of how data science works in practice is very low. This often results in the public struggling to see the value of using new computer analytic techniques, as opposed to more traditional methods. Public concerns tend to centre on:

- low awareness of how datasets are collected and collated;
- doubt as to whether computers can make better decisions than humans;

[2] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf
[3] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf
[4] Public dialogue on the ethics of data science in government, Ipsos MORI, 2016, available at: https://www.ipsos-mori.com/Assets/Docs/Publications/data-science-ethics-in-government.pdf
[5] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf
[6] Wellcome Trust Monitor Report – Wave 3: Tracking public views on science and biomedical research, Ipsos MORI, 2016, available at: https://www.ipsos-mori.com/Assets/Docs/Publications/sri-wellcome-trust-monitor-wave-3-2016.pdf
[7] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf
[8] Public dialogue on the ethics of data science in government, Ipsos MORI, 2016, available at: https://www.ipsos-mori.com/Assets/Docs/Publications/data-science-ethics-in-government.pdf

- caution about techniques that cluster individuals or the use of correlations between datasets that initially appear unrelated; and
- ambiguity about the level of control and automation that can, or should, be given to a computer.[9]

Where they are concerned about emerging technologies, the public often employ 'slippery slope' arguments to express their doubts. This is the idea that if we allow one aspect of a new technology to happen this will in turn justify further developments that people might not have consented to. For instance, our 2014 *Dialogue on Data* discussed participants' views on the increased use of data linking for analytical purposes, amongst other things. Participants used slippery slope arguments to outline their fears over data linking being extended – some returned to the idea of building a 'super database' or were worried about the data being used for purposes they would not support.[10]

Attitudes towards government use of data science also differed based on individual experience – those who regularly interact with a number of different government services were often quicker to see the benefit of policy objectives and those who are more used to sharing data through digital interactions were often quicker to see value in the concept of data science.

2.3.3 Public engagement strategies

The public want to be involved with new technologies. Three in four people feel that the government should act in line with public concerns with regards to scientific developments (75%) and nine in ten think that regulators need to communicate with the public.[11] Overall, people respond favourably to attempts by the scientific community to engage with the public. [12,13]

In dialogues conducted for scientific research institutes, participants typically say they want to learn about 'the scientific approach' and to find out about the latest developments as they happen. Participants usually argue that they should be consulted, and in principle tend to like the idea of a 'two-way conversation' between the public and scientists – even if they would not personally want to be involved.[14,15] Reflecting this, the majority of the public are interested in hearing directly from scientists about their research, but tend to prefer to hear from them via passive means, such as television, radio, newspapers and websites, as opposed to direct interaction.[16]

2.4 Objectives

The Royal Society's overarching objectives for the machine learning project are:

[9] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

[10] Dialogue on Data, 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/DownloadPublication/1652_sri-dialogue-on-data-2014.pdf

[11] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

[12] Babraham Institute: Public Dialogue on Future Strategy, Ipsos MORI, 2015, available at: https://www.ipsos-mori.com/Assets/Docs/Publications/sri-centgov-public-dialogue-babraham-2015.pdf

[13] John Innes Centre: Public Dialogue to Inform Science Strategy, Ipsos MORI, 2015, available at: https://www.ipsos-mori.com/Assets/Docs/Publications/sri-john-innes-centre-public-dialogue-strategy-2015.pdf

[14] Babraham Institute: Public Dialogue on Future Strategy, Ipsos MORI, 2015, available at: https://www.ipsos-mori.com/Assets/Docs/Publications/sri-centgov-public-dialogue-babraham-2015.pdf

[15] John Innes Centre: Public Dialogue to Inform Science Strategy, Ipsos MORI, 2015, available at: https://www.ipsos-mori.com/Assets/Docs/Publications/sri-john-innes-centre-public-dialogue-strategy-2015.pdf

[16] Wellcome Trust Monitor Report – Wave 3: Tracking public views on science and biomedical research, Ipsos MORI, 2016, available at: https://www.ipsos-mori.com/Assets/Docs/Publications/sri-wellcome-trust-monitor-wave-3-2016.pdf

- to raise awareness of machine learning, and its opportunities and challenges, amongst the public, policymakers and business;
- to raise the level of public engagement and debate, through increased public awareness and understanding of the technology, its current uses and near-term applications;
- to help ensure that public views of the technology inform relevant policy development;
- to identify the key social, ethical and legal issues that machine learning raises and suggest how these can be addressed; and
- to identify how the social and economic opportunities provided by the technology can be developed to deliver wider benefit to the UK.

Ipsos MORI's task for the research was to create an evidence base about public perceptions of the potential and risks of machine learning, to inform the Royal Society and the Working Group. Engagement activities were used to:

- engage relevant public groups about the potential of the new technology, to find out what they thought about machine learning, both before understanding it fully, and after;
- explore which attitudinal or demographic segments within the public should be priorities for further engagement;
- engage the public through the lifetime of the project, by designing ways they can explore emerging findings and help develop hypotheses; and
- give the public a voice in debating the recommendations and next steps.

It is important to note that fields that might be considered to be related to aspects of machine learning were not covered in detail by the public engagement exercise. These include robotics and drones, employment issues and the use of the technology in automating decision-making within government. However, some of these topics emerged spontaneously from the engagement exercises conducted with the public and so will appear as examples throughout this report.

2.5 Methodology

Research for this project required depth, breadth and iterative engagement. As such, the methodology was designed to incorporate three elements: a quantitative survey, public dialogues, and an online community.

While quantitative and qualitative methodologies are inherently very different, the methods used in this project were designed to complement each other in answering the same research objectives. Workshop participants went on a much more substantive journey through the day, and their views were nuanced. In the qualitative work, there was more scope for sharing opinions about machine learning and to bring participants to a level of understanding sufficient to be able to discuss specific case studies in depth.

A quantitative survey provides less opportunity to give respondents background information or indeed for them to truly deliberate; however, the findings give a robust overview of overall spontaneous attitudes towards machine learning amongst the general public. The quantitative and qualitative findings have been described separately in this report, but

drawn together in the concluding section. Findings from the online community are included in a separate annex to this report, given that the objectives were different for this element of the study.

2.5.1 Quantitative survey

A quantitative survey was used to capture the views of a representative sample of the general public. The objective of this quantitative research was to uncover the public's baseline understanding of machine learning.

Between 22 January and 8 February 2016, 978 face-to-face interviews were conducted with members of the public across the UK. All interviews were carried out in-home, using computer-assisted personal interviewing (CAPI) on Ipsos MORI's weekly omnibus survey. Final data was weighted to ensure they were representative of the national population.[17]

The findings from this aspect of the research are discussed in Chapter 7 of this report.

2.5.2 Dialogue events and evening discussion groups

A series of public dialogue events and evening discussion groups were used to develop in-depth qualitative insight into public views of machine learning. Two dialogue events were held in London and Birmingham in March 2016, each consisting of a Friday evening and a day-long Saturday workshop with the same participants. Two shorter, evening discussion groups were held in Oxford and Huddersfield.

An exploratory public dialogue approach was taken due to the complex nature of machine learning and the anticipated low levels of awareness and understanding of the topic. A workshop is an open environment that gives people time and space to learn new information, ask questions, change their minds and develop their views with other people. Workshops also allow an opportunity to explore how views develop when participants are given more detail via case studies and other stimuli. This meant that participants were able to see the practical applications of machine learning that are currently in use and better deliberate on how they might be used in the future.

Participants were recruited on-street by specialist Ipsos MORI qualitative recruiters. Recruitment quotas were set to ensure that, overall, people of a range of ages and from a variety of ethnic and socio-economic backgrounds took part.[18]

The two evening discussion groups were designed to target specific sub-groups and were recruited with additional quotas. These covered 'technologically literate' participants (Oxford) and participants who tended to have a high reliance on core services, such as the health service and employment support (Huddersfield). It was felt that these two groups of people may provide different insight about the impact that machine learning technology could have on public services.

2.5.3 Online community

Following the quantitative survey and the qualitative discussion groups, an online community was run to further explore the public's views on machine learning and how best to engage the public about machine learning in the future.

[17] Please see appendix (A.2) for the quantitative sample breakdown

[18] Please see appendix (A.3) for qualitative sample breakdowns

In total, 244 people signed up to take part in the online community, run with Ipsos MORI's partners, CMNTY. Most of them were recruited using a specialist online recruitment company, with a small number made up of those who had taken part in the discussion groups (for a full sample breakdown, please see Appendix A.4). The community consisted of five weeks of activities, spread over three months. The findings from the community research are included in the Appendices.

2.5.4 A note on interpreting *qualitative* research findings

Qualitative research approaches (including dialogue workshops) are used to shed light on *why* people hold particular views, rather than *how many* people hold those views. These approaches are used to explore the nuances and diversity of views, the factors which shape or underlie them, and the ideas and situations in which views can change. The results are intended to be illustrative, rather than statistically reliable.

This report aims to provide detailed and exploratory findings that uncover the perceptions, thoughts and feelings of people about machine learning, rather than statistical evidence from a representative sample. It is not always possible in qualitative research to provide a precise or useful indication of the prevalence of a certain view, due to the relatively small number of participants generally involved (as compared with the larger respondent bases involved with quantitative surveys).

Sometimes, ideas can be mentioned a number of times in a discussion, and yet hide the true drivers of thoughts or behaviours; or a minority view can, in analysis, turn out to express and important emergent view or trend. The value of qualitative work is to identify the issues which bear future investigation. Therefore, we use different analysis techniques to identify how important an idea is. The qualitative report states the strength of feeling about a particular point, rather than the number of people who have expressed that thought.

However, it is sometimes useful to note which ideas were discussed most by participants, so we also favour phrases such as 'a few' or 'some' to reflect views which we mentioned infrequently and 'many' or 'most' when views are more frequently expressed. Any proportions used in our qualitative reporting should always be considered indicative, rather than exact.

Verbatim comments have been included in this report to illustrate and highlight key points, i.e. those views either expressing strong sentiment shared by the group as a whole, or reflecting the strong views of a smaller subset. Where verbatim quotes are used, they have been anonymised and attributed by location.

3 Reactions to machine learning

This chapter explores the public's understanding of machine learning and its applications, as discussed during the dialogue workshops. It sets out the different spontaneous and more considered reactions observed, highlighting people's perceptions of machine learning as a concept, and their take on the risks and benefits associated with this technology in different contexts.

Summary

Participants were not familiar with the term 'machine learning' and found it easier to engage with the idea through real-life examples. Most had come across at least some specific applications of machine learning in their day-to-day experiences. In general, they were not concerned with the detail of the mechanisms underpinning how machine learning works, focusing instead on how and why it could be used in different contexts.

All participants could relate personally to the technology, as they could see where it could have an impact on at least some areas of their life. Those who had a more positive outlook tended to see machine learning as an important emerging technology, whereas those who were more cautious tended to focus more on concerns about how the technology would work, and the purpose behind its use in different scenarios.

Several different spontaneous reactions to machine learning were observed among participants. These spontaneous reactions formed the basis for how participants engaged with machine learning as the discussions progressed, shaping how they viewed individual examples and their overall views of machine learning. These spontaneous reactions were:

- 'I can personally relate to this technology, because I can see where this could have an impact on my life, whether good or bad'
- 'This is an important emerging technology and it carries potential risks and benefits to society'
- 'I can't see how this would work – humans are too unique for machines to really understand us'
- 'I'm suspicious about the purpose of this technology'

Participants' spontaneous reactions represented both concerns about machine learning and positive views about the opportunities this technology could bring. Participants' views developed over time, but these spontaneous reactions remained the basis for their more considered opinions.

As participants considered machine learning in more detail, several key concerns and opportunities emerged. Concerns focused on depersonalisation, risk of individual and societal harm, restriction of choice, and people being replaced. The opportunities discussed included the potential to improve how services work, saving time, and enabling more meaningful choice for service users and consumers, as well as a more general sense that this technology has the potential to improve life and society in many different ways.

3.1 Spontaneous awareness and understanding of machine learning

During the qualitative workshops and discussion groups, participants were introduced to machine learning as technology that allows machines to learn from data and improve their own performance. From the early discussions it was clear that most participants knew little or nothing about maching learning before taking part.

> "Machine learning… are we talking about people learning by machines or machines actually learning?" (Birmingham)

Learning was seen a human activity by participants, and the idea that a machine could learn was not one that all found easy to grasp. Participants were more familiar with the idea that machines can be used to analyse large amounts of data to spot patterns. However, most participants took time to understand how a self-learning algorithm could work, and how it was disctinct from machines simply following a set of rules or instructions. In particular, the link to decision-making made the concept of maching learning difficult for many, at least initially. Many participants felt a machine would not be able to process the variety and nuance of factors that humans analyse when learning to make complex decisions.

> "There's no way a computer can learn to make its own decisions, no way on this earth … It can never be as big as the brain. It's making it sound like the computer's intelligent, like us, and it can't be…" (Birmingham)

The 'machine' aspect of machine learning also generated initial debate among participants. They often associated machine learning with robots that were able to learn, rather than algorithms or computer programmes. Other research tells us that people tend to find it difficult to envisage how robots might be used in everyday life. As a result, they default to thinking of a robot performing exactly the same task as a human, in the same way.[19] Many participants' first interpretations of the 'learning' part of machine learning were grounded in these very direct comparisons between how humans learn and whether they felt it was possible for machines to learn in the same way.

Those participants who felt comfortable more quickly with machine learning as a technology used different parallels between human and machine learning to try to explain the concept to others. They suggested that 'our brains are computers' and 'we analyse our memories and experiences like this'. These kinds of analogies were often helpful for those participants who struggled to understand the concept.

After further discussion, most participants were clear that machine learning is not equivalent to human learning, or the same as artificial intelligence. Machine learning requires human input, as a machine needs to be given data and programmed in the first place. Machines that learn can make decisions and predictions based on processing large amounts of data, and can improve how well they do this.

3.2 Awareness of current applications of machine learning

While machine learning was not a familiar term and took time for many participants to grasp, there was more awareness and understanding of the applications of machine learning. Despite not knowing the term 'machine learning', they did know about some of its uses.

[19] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

During the workshops, participants discussed machine learning through the use of practical examples, which are summarised in the table below:

Table 3.1: Summary of case studies used in the qualitative phases			
Qualitative case studies			
Art Machine learning used to generate poetry	Transport Driverless vehicles which can adapt to road and traffic conditions	Finance Spotting fradulent activity or warning before transactions if balances are low	Crime Analysing statistics and predicting crime patterns to allocate resources
Social care Robots that adapt to the home environment, for example helping to care for older people	Health Cancer screening technology; vocal analysis to detect Parkinson's or mental health issues	Education Online learning providing a tailored experience	Marketing Tailored online adverts and predicting products

The qualitative participants' levels of awareness of specific machine learning applications reflected the survey findings, discussed later in Chapter 7. They tended to have come across the same subset of applications, most commonly recommendation services on sites like Amazon or Netflix, or reward schemes such as supermarket loyalty cards and tailored vouchers.

> "I'd never heard it called that before, but I recognise a lot of these examples and now I realise I had heard of the concept." (London)

However, most participants understood machine learning applications only from a user perspective, with very few aware of the mechanics of how the technology works, even at a broad conceptual level. This meant participants used a combination of their own experiences of familiar machine learning applications alongside the information provided during the discussions to arrive at a broader understanding of machine learning as an idea, and the basics of how it works . Some experienced a moment of realisation when they made the link between their own interaction with a specific technology and the broader discussion about machine learning.

> "You go on Amazon, and it says 'you watched this, you'll like this', and it's something that you'd like. It's tailoring it to you." (Huddersfield)

3.3 Reactions to machine learning

After developing a basic understanding of the concept of maching learning, participants reacted in a number of different and often overlapping ways as they considered it further. These spontaneous reactions can broadly be divided into four types:

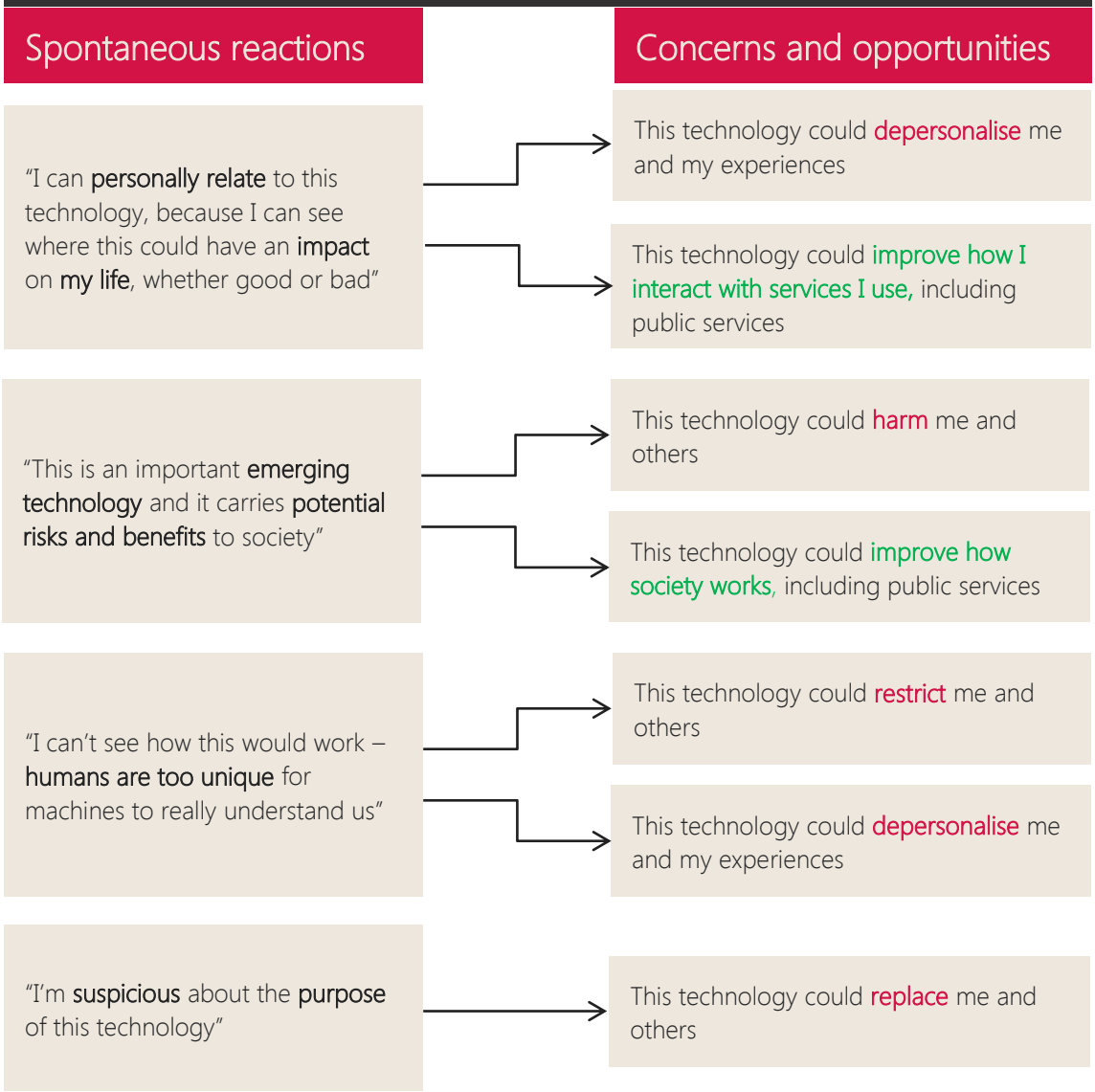
1. "I can personally relate to this technology, because I can see where this could have an impact on my life, whether good or bad"
2. "This is an important emerging technology and it carries potential risks and benefits to society"
3. "I can't see how this would work – humans are too unique for machines to really understand us"

4. "I'm suspicious about the purpose of this technology"

All participants could relate personally to machine learning, as they could see where it would or could have an impact on their life through the broad case study areas discussed. Those who were spontaneously more positive tended to see it as an important emerging technology which could bring benefits but was not without risks. Those who were more cautious about machine learning tended to be concerned about preserving the uniqueness of humans, and were often worried in a general sense about why this technology was being introduced.

All four of these reactions were common among participants, with almost all responding in more than one way. These reactions tended to shape their perceptions as they absorbed more information and discussed current and potential machine learning applications, as shown below in Figure 3.1.

Figure 3.1: Spontaneous reactions and approaches to machine learning



It is worth highlighting that most participants were not particularly interested in the complex nature of machine learning algorithms. Instead, they tended to be content with a basic conceptual understanding that machine learning algorithms require data, which they use to analyse and make predictions, before learning from feedback from those predictions. They assumed that if the algorithms did not work then they would not be used.

Furthermore, participants tended to think that the technology that underpins machine learning was neutral. They did not really consider how the algorithms worked. Rather, concerns were often based on how they expected the technology to be used, and in particular the motives of the people behind machine learning's specific applications.

Overall, participants typically took a pragmatic, balanced view of the potential risks and benefits of machine learning and its applications. Whilst the idea that machine learning could depersonalise people appears twice in Figure 2.1, this is not because that sentiment was necessarily stronger or more prevalent than any of the others. Rather, it emerged that participants with different spontaneous reactions shared this concern:

1. Some feared depersonalisation because they saw machine learning as altering how they enjoy experiences they value (for instance, driverless cars taking away the pleasure of driving).

2. Others were worried about depersonalisation because they did not believe that an algorithm would be able to accurately predict individuals' needs or behaviours, particularly in people-facing roles. Instead, they worried that machines would make broad generalisations about groups of people, rather than producing a tailored, individual analysis.

However, their discussions about the risks were often more nuanced and complex than those on the potential benefits of machine learning. Discussions surrounding the benefits were usually obvious to most participants, but seen in more general terms, as described in Section 3.5.

One of the main reasons participants gave for engaging with machine learning was because it was already present in their lives. Despite many being positive about the current uses of this technology, participants could also understand why they were being asked for their views about the acceptability of different uses of machine learning.

> "I am excited about finding out more about this subject [...] I've learnt a huge amount and am looking forward to discussing it further with friends and family." (Birmingham)

When making a poster about machine learning, one London participant's phrase was 'Machine learning is here to stay – get used to it'. Some, though not all, felt that it was inevitable that machine learning would play a greater role in their lives as individuals and to society as a whole. These participants also thought it was important for society to consider carefully how best to exploit this technology, in order to maximise the benefits while minimising the potential risks.

Participants' familiarity with current applications helped them to engage with potential future uses for machine learning. In addition, they were unaware of any significant negative consequences as a result of machine learning they had encountered thus far.[20] As a result, these participants felt better able to trust the safety and reliability of machine learning in what they considered to be more serious and sensitive contexts, such as healthcare or education.

3.4 Concerns about machine learning

This next section will consider each of the broad concerns expressed by different participants about maching learning, and how these shaped views throughout the dialogue discussions. As such, this is drawn from their spontaneous reactions and their feedback on case studies of specific machine learning technologies explored during the workshops.

[20] The only exception to this were some who were aware of an accident involving self-driving cars

3.4.1 'This technology could harm me and others'

As a new technology, some participants wanted reassurance that extensive testing would be done on all new applications that would rely on machine learning technology. This was especially the case when machine learning could be used in higher stakes situations that could shape people's lives.

For instance, participants typically wanted driverless cars to be tested under a range of conditions (such as icy roads, heavy rain, sudden objects appearing in their path) and to pass them all before they would want to see them integrated onto the road. Furthermore, for machine learning to be used in medical diagnoses, they usually wanted an experienced consultant to always be checking the findings. This nervousness about safety was less evident when discussing applications of machine learning that already exist, such as Netflix recommendation services and supermarket loyalty cards. These were perceived as being relatively harmless, because the consequences of something going wrong were not viewed as serious for the individuals involved.

At the very least, in some more sensitive scenarios, some participants wanted to know that there was good evidence that people would not be harmed by this technology. They also wanted humans to continue to be involved in some way to make final decisions and deal with any problems that occur.

Linked to this, the fear of the unknown was also a barrier to engaging with machine learning for many. This was particularly the case for machine learning applications that were seen as further from how this technology is currently used. For example, all road users having driverless cars, or robots assisting with social care duties in the home. One participant summarised the feeling expressed in his group:

> "Nowadays, there is a lot of technology coming in, but a lot of us don't understand what these machines do, so we think they're a threat. I think the danger is that the level of communication about what they do and how they do it… there's not enough information out there, not enough effort to communicate with the public." (Birmingham)

Participants observed that a great deal of machine learning had 'already happened' without anyone really knowing about it. Some of the more sceptical members of the group felt that this lack of communication made the process seem secretive and thought that the potential risks of machine learning were being deliberately hidden from the public.

Some participants pointed out that, as a self-improving technology, machine learning would have to involve mistakes during the learning process. This concerned the more cautious participants. One spoke of her frustration at the autocorrect software on her phone making incorrect spelling or wording suggestions.

> "If it can't get something that simple right at this stage, I have no faith in it." (London)

There was some feeling that machine learning should really be better than humans to be worthwhile. If humans were able to carry out a particular task adequately, participants could see little value in switching to machine learning. They struggled to see the point of developing machine learning approaches where humans were perceived as performing well.

Some participants pointed to their other concerns about the potential risks of machine learning, and argued that continuing with a human approach was preferable unless there were clear benefits in terms of accuracy. Increased speed was seen as potentially beneficial, provided accuracy was maintained or preferably increased. While a minority view, for

these participants their concerns persisted throughout the discussions and they felt that only strong, consistent evidence of machine learning being accurate and safe would convince them of its merits.

3.4.2 'This technology could replace me'

Some participants were concerned more specifically that machine learning technology would become so sophisticated that it would eventually replace large numbers of roles currently carried out by humans. This concern was twofold:

1. On the one hand, some participants felt that machine learning could be developed to the extent that it could replace an array of skilled and manual labour jobs
2. On the other, some participants were worried that advances in machine learning would contribute to general de-skilling and over-reliance on technology that was seen as a negative characteristic of modern life

The potentially negative consequences of machine learning for jobs and employment were a repeated concern. Parallels were drawn between machine learning and advances in automation that have historically caused large-scale redundancies in production-line reliant industries, such as car manufacturing.

Participants noted that automation had often resulted in the loss of low-skilled jobs. They felt that machine learning was much more versatile, as its applications span many sectors and industries. The primary concern was that machine learning could cause unemployment on a mass scale, as opposed to unemployment in certain sectors.

> "I'm just thinking about where I could be replaced … probably most things I do!" (London)

In Oxford, one participant brought up mortgage advice as an example of machines and humans working in tandem. While a customer might consult with a human advisor, the advisor would simply input personal information into an algorithm that would then return a list of all the viable mortgage options. Participants only needed to make a small conceptual leap to see how these developments could lead to changes, and potentially unemployment, in many skilled, professional jobs.

> "I think machines are actually taking over quite a lot of jobs and that. You go into Tesco's and there's a self-checkout and automation, like your Oyster cards, etc. There's more people now, but less people to do more people's jobs, and I think with this advancement in technology it's just going to get worse." (Oxford)

Some participants also noted that successful advances in technology inevitably result in a corresponding increase in reliance on that technology. They were able to see more and more areas of daily life where machines could replace human roles, and this concerned them. As well as considering unemployment on a mass scale, some participants would also refer to their own jobs, suggesting that they feared replacement on both a societal and personal level. This was evident across all ages, socio-economic backgrounds and those with a variety of work histories.

As participants felt there was a significant risk of people being replaced, there was a real desire to know what benefits machine learning might have for individuals to offset this. Participants were generally eager to be convinced that there would be no detrimental effect – or that there could even be improvements – to their wellbeing, particularly through the impact on the jobs market.

> "Everybody here is thinking, 'well… I'm going to lose my job!'. That's what worries me, what's the purpose of it? No technology has made us any freer – unless you are the inventor who's sitting at the top of the tree." (London)

These concerns were particularly evident in a small but vocal element present in all groups, who were suspicious of machine learning because they thought it was intended to replace humans. These participants wanted to know what the exact purpose of machine learning was, when a human performing such jobs was often sufficient. This minority felt that government and businesses wanted to control people more, and saw machine learning as a way for this to happen – by replacing them with 'obedient' machines.

> "Machines do what they're told and human beings don't always do what they're told. So if you have the choice between a disobedient human and an obedient robot… Before we know it, anyone who speaks out against the system will be sent off by a machine." (London)

The second broad area of concern was that machine learning replacing people would lead to an overall de-skilling in society. It was felt that over-reliance on modern technology was already happening. Participants gave examples of skills that are being lost, such as reading maps or memorising phone numbers, because smartphones can do this for us.

Participants discussed how reliance on technology had made people 'lazy' and that machine learning would continue the existing trend of 'de-skilling'. Participants drew on examples such as doctors and pharmacists relying on computers for diagnoses and prescriptions, but also their own use of online tools such as maps and calculators. In these and other cases, responsibility for analysing data in order to make informed decisions was seen as being transferred from humans to machines.

There was a concern that this greater reliance on technology would result in people believing the first thing they came across, for example – choosing to believe the first item on an internet search return. Some participants observed that this over-reliance and lack of considered judgement was already commonplace. They feared that the ability to interrogate information and arrive at one's own conclusions would be permanently lost, having widespread, detrimental effects – especially if the technology (upon which they felt we heavily rely) malfunctioned.

> "All these subjects lead to 'we're giving our responsibilities away to machines'. It will lead to less skilled people and higher unemployment." (London)

3.4.3 'This technology could depersonalise me and my experiences'

Many participants could relate to the potential applications of machine learning on a personal level. They could see where it could have an impact on experiences that they value and that made them feel human. Concerns around depersonalisation were apparent in two main ways, as mentioned previously:

1. Some feared depersonalisation because they saw machine learning as altering how they enjoy experiences they value (for instance, driverless cars taking away the pleasure of driving).

2. Others were worried about depersonalisation because they did not believe that an algorithm would be able to accurately predict individuals' needs or behaviours, particularly in people-facing roles. Instead, they worried that machines would make broad generalisations about groups of people, rather than producing a tailored, individual analysis.

Many participants had concerns about machine learning devaluing experiences that were important to their sense of enjoyment. The specific applications that troubled individuals varied from person to person, based on the experiences they most valued. Those who were concerned about their experiences being depersonalised tended to be reacting to a specific example of machine learning under discussion. These perceived challenges to individual expression and personal fulfilment elicited some of the strongest concerns among participants. Below are two examples of the types of activities individual participants cited when describing their concerns about depersonalisation.

Personal experience: Poetry

"I told complete strangers that I lost my house through poetry [...] poetry is about spirit and soul. It's about the essence of your life, put out there for other people to say they don't like, or they do. Poetry is my life, it saved me. It's absolute heart and soul to me and to other people." (Birmingham)

Personal experience: Driving

"Driving cars isn't simply about going from one place to another. Could you have a driverless Formula One race? It's a very individual thing, it's a personal activity. Machine learning means you have to re-evaluate how you live and what you want from things like driving." (London)

While these individuals were particularly passionate, depersonalisation was also a broad concern shared by many others, with their focus on the activities and experiences that mattered most to them. Machine learning was seen as having great potential to improve life and society in some areas. Even so, some participants thought that this technology could also reduce personal fulfilment and enjoyment in other areas, with a risk that it could undermine important aspects of what it means to be human.

The more cautious participants were worried that machine learning would be introduced and 'sold' to the public on the basis that it would improve their personal experiences, but that there would then be a 'slippery slope' in future. They felt that as the technology became more accepted and commonplace, it would be introduced in more contexts, until it was a feature in most areas of everyday life. These participants were concerned that the technology might try to improve their experiences in a way that they didn't want, eventually taking over the activity altogether – until enjoyable activities such as driving or reading poetry were effectively lost to those who value them.

The second concern around depersonalisation was those who felt that machine learning would replace jobs that they felt humans should continue to do, resulting in more impersonal services and experiences. For example, in one discussion, participants speculated that a visit to their local GP might become a visit to a room with a computer, into which they would, for example, list their symptoms and a diagnosis would be determined. There was a great deal of concern that meaningful human interaction – which was considered to be especially important in sectors such as health, social care, and education – could be lost with the development of machine learning. Genuine engagement and empathy were considered vital in many services – and desirable in all human interactions – and participants felt that a machine could never be capable of connecting with a human on an emotional level.

> "I think it could save a lot of money, but I don't want it to be used as a money saving tool if they're losing the human element behind it. People will lose jobs, but they're the ones at the front, seeing what's going on. That's an important role." (Huddersfield)

This was linked to wider concerns about the loss of human interaction in society more generally. Participants typically felt that an increasing reliance on technology had already resulted in reduced face-to-face, personal interactions, referring to 'worrying' habits, like parents occupying their children with iPads or televisions.

> "My neighbour's daughter sounds like a computer game, because she's been exposed to so much technology. They'll have such a limited range of expressions… if we all learn from robots, it's taking away our humanity a little bit." (London)

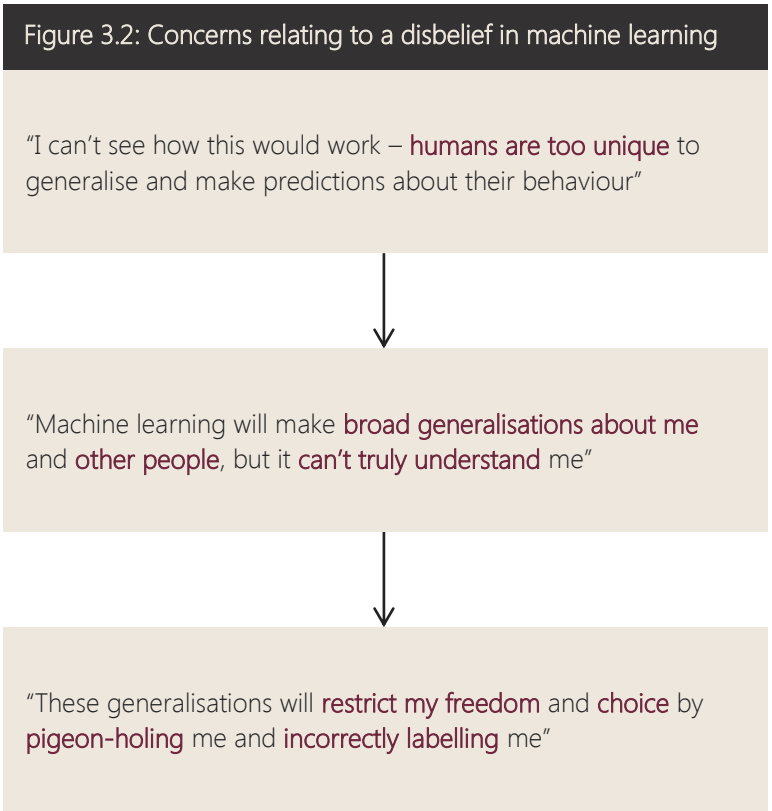
3.4.4 'This technology could restrict me'

Most participants understood the predictive power of machine learning algorithms, at least in principle. However, they did not believe that these predictions could or should be used in all contexts. In situations that required nuanced interpretation of data or that involved understanding the experiences or behaviour of individuals, many struggled to believe that machine learning could work effectively. Those who felt this way were concerned that a machine would make a 'best attempt' at predicting, and that this would amount to a broad generalisation, rather than an accurate prediction of how an individual would act.

As a result, these participants feared that the use of machine learning could ultimately lead to restrictions being placed on individual freedom and choice as a result of predictions and decisions made by algorithms. They were worried that they might fall victim to 'being pigeon-holed' or 'labelled incorrectly'.

The link between participants being unconvinced about the efficacy of machine learning and fears over restricted choice is shown in Figure 3.2, below.

Figure 3.2: Concerns relating to a disbelief in machine learning



Participants were able to accept that predictions could be accurately made where only observation and objectivity were needed. For instance, they understood that in terms of diagnosis, machines could spot patterns in physical health conditions better than human doctors, as they were capable of analysing much more data. Indeed, even those with

concerns about the reliability of machine learning often recognised how the technology could be helpful in reducing mistakes made in certain scenarios (particularly in a health setting).

However, if the task currently being carried out by a human was subjective and required interpretation, some could not accept that a machine would be able to make accurate predictions – these participants felt that human behaviour was often random, or unpredictable, and could not be learnt by machines. They argued that using machine learning in these contexts would introduce a different kind of error.

> "Human behaviour is different, you can't predict what's going to happen next. Humans might change their mind, but a computer won't know that you're going to change your mind." (Huddersfield)

This developed from a belief that humans were unique and that, whilst their behaviour might be predictable to a certain extent, a machine could never capture everything about them and their decisions because people are not always logical. These participants concluded that a machine would not be able to analyse such random behaviour well enough to always make an accurate prediction. Their concern was that the machine would make broad generalisations, based on data it held on them and people similar to them. As a result, they felt that computer algorithms could never encapsulate the full range of human experiences and preferences.

A further example that participants highlighted focused on understanding the complexity of human language. When discussing machine learning and poetry, participants were told that machines that analyse language are not cognisant of what individual words mean; they just look for patterns in their use. Some participants questioned the ability of machine learning algorithms to correctly interpret language, if they were not able to understand their meaning.

> "You can say something, say a certain word, but it can mean different things. It'll translate it to something totally different." (Huddersfield)

Another area where some participants were nervous about machine learning was mental health. Participants discussed how even human doctors struggled to understand and diagnose mental health issues, and did not believe that this could be done as effectively through machine learning. Because mental health manifests itself in so many different ways and participants argued each individual and their condition are unique, it was felt that a machine would not be able to make a tailored analysis, but would generalise when informing a diagnosis. Their concern was that this generalisation would lead to an inaccurate diagnosis, which could have far-reaching repercussions.

> "You start putting a label on people, because the machine says that they've got depression … that's going to affect your whole life. That affects your career. That affects your driving license, everything." (Oxford)

Participants speculated that the consequences of these generalisations would often be negative for individuals. They feared that freedom would be reduced because the options available to them would be restricted. For example, some participants were wary of Amazon or Netflix's recommendation services. They saw these algorithms, that suggest books or films you might like, as limiting choice. They were concerned that people would only read books that were recommended to them and never broaden their horizons. The result would be that they were confined by their previous interest in certain genres, or the interests of other people like them.

However, there were other areas where this potential restriction of freedom of choice was considered much more problematic because of the importance of the issues at stake. For example, some participants were worried about the consequences of using machine learning to provide more personalised education.

> "They'd allocate you a course, designed by the data they've got on you. They'd decide your future career pattern, you'd go to a centre and you're allocated your space and that's your education. It's promising us the earth, but what are they taking away from us?" (London)

While a minority view, there were concerns that children would no longer receive a well-rounded education as they would be pointed towards a specific career at an early age. They feared that they would miss out on other learning other life-skills and so be unable to deviate from their career plan.

3.5 Opportunities for machine learning

Running alongside the discussions about the risks of machine learning was one about the opportunities it brings, and the potential benefits to individuals and society. Despite some concerns, participants could envisage a future where humans and machines worked in tandem and were often optimistic about many of the ways machine learning could be used.

> "There's loads of benefits, it's a fantastic resource, but it's how a resource like this is used that's most important." (Birmingham)

> "Whatever can help us to progress and expand our minds, it can only be beneficial to humans." (Huddersfield)

There were three broad reasons given by participants in favour of using machine learning, with each discussed in turn below. Participants saw the main benefit of machine learning being the ability to process much larger data sets than humans ever could on their own. Because of this feature, machine learning could be a powerful way of augmenting human ability. They often came back to this point as a counterargument when the discussions turned to the potential risks of machine learning.

3.5.1 'This technology has a lot of potential to benefit society and individuals'

A substantial minority of participants were broadly optimistic about machine learning from the outset. They argued that this was a new technology with real potential to benefit both society and individuals. These participants alluded to how wide-ranging the positive influence of machine learning could be. Machine learning was thought to facilitate spotting patterns in data that would otherwise have gone unnoticed. As the discussions became more sophisticated, they pointed out that machine learning could be applied in any context where data is collected and decisions need to be made, and this often made them more positive about the technology.

Many participants saw machines as being detached and efficient and thought that they could remove many sources of human error. Some spontaneously mentioned the problems with human decision-making. They pointed out that humans could look at data in biased ways, or even that they could be deliberately prejudiced when they made their decisions. These participants argued that this doubt could be removed if the decision was being made objectively by an algorithm, provided it could be shown to be accurate and to work effectively.

> "People could let machines read and make diagnoses, I think a lot of mistakes are made when people are emotional and tired." (London)

Another important consideration for those who were positive about machine learning was the potential for increased accuracy and analytical power in the context of scarce resources in the public sector. Many felt that professionals in core services such as health and social care, education, and policing were under a lot of strain, trying to meet the demands of service users with reduced budgets and staffing levels. They felt that machine learning would be able to help by identifying patterns and allocating resources more effectively, reducing pressure on core services.

> "If it says, 'oh, look, this is our busiest time and this seems to be across the week, then we need to put more staff in this area at that time' [...] this benign data could actually be really helpful." (Oxford)

> "A lot of police are hired on a 9-5 basis... But if you could use this to spot when the most likely times are, if you had them at the right times, right places ... it would be better to ensure allocation of resources." (Birmingham)

For example, they discussed machine learning algorithms being used to identify students who were at risk of becoming NEETs (Not in Education, Employment or Training). Many were optimistic that machine learning would be able to spot patterns and identify students who seemed to be more likely to be at risk of becoming NEET in a way that humans never could on their own, flagging this to teachers who would then be able to focus more on these students.

> "You've got to be careful what you say as a teacher – parents can get shirty. But if it's coming from data, it's not coming from the teacher." (London)

Participants could also see significant benefits for businesses from machine learning. They cited examples of tailored marketing and offers, and discussed the data collected by retailers about their customers. Most had no problem with businesses using these technologies provided they had obtained consent to use data in this way (see Section 5.3 for more information). They also wanted businesses to use machine learning improve their service to customers, and not just to cut costs or target products at people.

Participants felt that machine learning would be able to help address current problems, such as the efficient allocation of public sector resources, and future, as yet unknown, problems. Participants were particularly hopeful that machine learning would be able to help identify and manage the consequences of an ever-increasing population, or improving our response to big challenges like climate change.

> "The whole population thing is quite interesting, the theory that the population will grow so large we won't be able to sustain it. The difference has been technology – we found more efficient ways to feed more people. Machine learning gives us a way to look at these big questions and try to solve them, so the fear of population growth could, in theory, be resolved." (Birmingham)

3.5.2 'This technology could save a lot of time'

There was also a more specific view that machine learning could save a lot of time, due to its greater efficiency at performing the same tasks as humans. Participants generally felt that this would free up humans' time, so that they could concentrate on other things. This was seen as the positive side of the expectation that machine learning could replace humans in many roles.

> "When it's managed correctly, we've got more time to get back to life, enjoying things, rather than working 70, 80 hours a week." (Huddersfield)

Participants thought that, if used appropriately, machine learning could do a lot of the background, administrative work humans do not want to do, or should not be doing. In the public sector, this would free up more time for healthcare and educational professionals to focus on their patients and students. While there were concerns about how the private sector would use machine learning, participants could also see the benefits this technology could bring to businesses in terms of efficiency.

Many participants also saw the potential of machine learning to make every day life easier. Those who took this view felt that the technology could help them do some of the things they do anyway, but do so much more quickly.

3.5.3 'This technology could give me better choices'

Some participants felt that recommendation services, such as Amazon and Netflix, may make suggestions to them that they never would have thought of themselves. They would then benefit by trying something different, possibly discovering that they enjoyed something new. As consumers, they saw this as increasing their choice and freedom, rather than restricting it, provided machine learning predictions were not forcing them to buy certain products or act in certain ways.

> "I'm quite interested in architecture books. I find recommendations really interesting – it's just the same as standing in the shop and looking at the shelves. It's not like it's telling me what I 'should' be reading." (Birmingham)

Participants also liked the idea of being treated as an individual, rather than the 'one size fits all' approach that they felt was prevalent, particularly in areas such as education. These participants were interested in the idea of receiving a more personalised service. For example, many felt that they were currently bombarded by unsolicited and irrelevant adverts. Some felt machine learning could improve tailoring, meaning they would be targeted with adverts that were more relevant to them.

> "The good side is that people try to guess what you like, what environment you like, what will make it a more pleasurable experience. There's a huge amount of money spent on marketing – machine learning could make it more refined." (Birmingham)

With education, many participants pointed out that people are different in terms of their strengths and weaknesses and so will learn in different ways. They saw the education example as having great potential to inspire children to learn, by delivering education in a way that children could identify with as individuals. Some participants, many drawing on personal experience, felt that the current education system alienates some pupils, who feel inferior if they cannot understand how they are being taught.

> "I think it's good, I thought that was what machine learning was going to be about. My bugbear is teachers teaching students to be teachers, they focus on those that are good at academics rather than practical things, and then they drill through a syllabus." (Huddersfield)

How participants viewed machine learning and choice often reflected and informed their overall outlook on the use of this technology, particularly from a consumer perspective. They were broadly split between two groups:

1. Some felt machine learning would restrict choice, in that it would mean they would not see all the options available to them because an algorithm had decided that only certain choices were relevant.

2. Others viewed machine learning as enabling meaningful choice in the context of an otherwise overwhelming range of options, provided they could and select options not recommended by machine learning

Those who had a more positive outlook felt that by analysing individual human needs and preferences, people could benefit from genuinely personalised services, and therefore more meaningful or useful choice. They assumed that machine learning algorithms could be flexible, taking into account their individual requirements and routines, when tailoring suggestions or programmes to them. There was also an assumption that they would be able to continue to make choices without reference to the algorithm's suggestions.

Taking book recommendations as an example, those who were positive about machine learning saw this technology as a way to help them navigate the otherwise overwhelming choice and to make a selection they were likely to enjoy. The more pessimistic participants, however, saw machine learning as restrictive, disliking the idea that they might not see all the available options and be restricted to books similar to those they had read before.

4 Considering the risks and benefits of machine learning

Participants were asked to consider different case studies and weigh up the potential risks and benefits of each one. For the case studies, information was presented about how this technology might be used, and the extent to which machine learning is involved in helping it to work.

This chapter presents a thematic analysis of how participants evaluated different machine learning applications. It describes the key stages they went through in considering each, and the ways they sought to weigh up the risks and benefits of machine learning in different contexts.

Summary

Participants took a pragmatic approach to how machine learning should be applied. They discussed the intended purposes, perceived motivations of those using the technology, and the consequences for individuals and society. If they felt a particular use of machine learning was desirable and appropriate in principle, they then weighed up the detailed benefits and risks to decide whether they could support it or not.

Their overlapping criteria for deciding whether they liked an application in principle were:

- The perceived intention behind using the technology in a particular context;
- Who the beneficiaries would be;
- How necessary it was to use machine learning;
- How appropriate they felt it was for machine learning to be used; and
- Whether or not a machine will need to make an autonomous decision.

In some cases, their views on one of these criteria sometimes dominated. In other cases, participants balanced a number of different criteria when giving their views, and did not all agree.

Participants' assessments of the risks and benefits were often instinctive and nuanced. Definitions of risk varied between different case studies and amongst participants. Some focused on risks around machine learning providing results that are inaccurate, or subject to contradiction. Others tended to focus on the broader risks of machine learning applications to society, including relinquishing control to machines, or being replaced by them in some way. For others, their main priority was managing potential harm to individuals, particularly those that resulted from automated technology.

The risks were usually easier for participants to identify, and they spent considerable time discussing these in the context of the different case studies considered. Despite their concerns, participants recognised the opportunities associated with machine learning and the potential for significant benefits for individuals and society.

4.1 How participants assessed machine learning applications

4.1.1 Criteria used before considering risks and benefits

Participants used a number of overlapping criteria to evaluate potential machine learning applications, before they would engage with the specific risks and benefits for a particular example. The most common were:

1. The perceived intention behind using the technology in a particular context. Linked with this was a desire to understand who would be involved with the development and delivery of the technology. Participants generally felt that the roles and responsibilities of individuals or organisations were important. They felt that the motives of those involved with the development might shape the success, and direction, of the technology as it progresses.

2. Who the beneficiaries would be. Where the benefit was felt to be more universal, such as with personalised learning across the educational sector, views were more positive than they were with self-driving cars, which may initially only be available to the few. For example, if the sole purpose was making money for companies, then the application was considered less worthwhile.

3. How necessary it was to use machine learning. In certain contexts, some participants struggled to see why machine learning was necessary. This was particularly the case where humans were seen as being as good or better than a machine at completing the task. The clearest example of this was creating art; many participants could not see the point of a machine doing so.

4. How appropriate it was for machine learning to be used. This shaped the discussion around some of the more tangible applications such as robots in the home, but the discussion was also relevant when the outputs were more abstract, such as personalised learning. Many of these concerns centred around the loss of human-to-human contact.

5. Whether or not a machine will need to make an autonomous decision. If the example necessitated making a decision, the importance of getting that decision right was a key factor in the public's assessment of the risks of machine learning.

These criteria overlap, and the way they were applied and weighted depended on participants' views of the specific type of machine learning in question.

4.1.2 Weighing up the risks and benefits

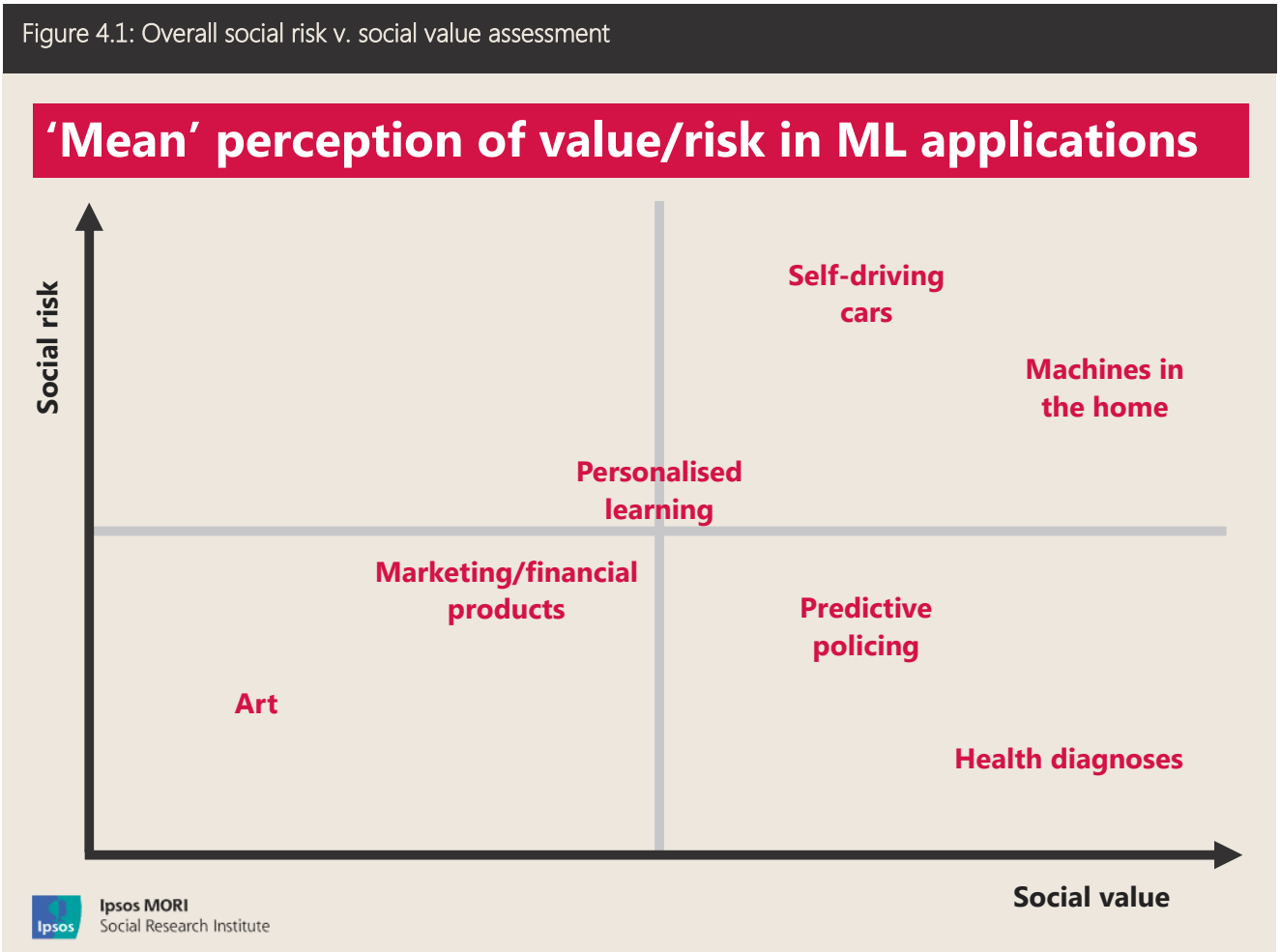
Within this wider framework then followed a discussion about risks versus benefits. Across the qualitative discussions, there were felt to be risks associated with machine learning generally, as well as in relation to specific applications. The actual process of machine learning – the computation and 'data crunching' – was not seen as being particularly problematic, although there were concerns about how accurate predictions would be, and about the consequences of any mistakes.

At first, participants discussed machine learning with reference to their own experiences and the impact it could have upon them as individuals, and on other people. They generally found this much easier than discussing machine learning and society more broadly. They tended to focus on the decisions made and actions taken as a result of machine learning as being the main sources of risk.

Participants then began 'weighing up' the benefits and their concerns about machine learning for society. This sparked a lot of debate and some participants found it easier than others to disentangle how machine learning could affect them personally from how they perceived machine learning's potential impact on society as a whole. That being said, the societal impacts of machine learning were typically framed under the five criteria outlined above.

Participants were then asked to place each type of machine learning on a quadrant that captured the perceived social risks and social value for each application. While there were some differences between groups, a broad consensus did emerge. Figure 4.1 shows where the different case studies were typically placed by participants.[21]

Figure 4.1: Overall social risk v. social value assessment



Machine learning being used in health diagnoses was the most supported example and the one where participants could see a clear role for machine learning in improving how things are done currently. Participants' feeling that the benefits outweighed the risks was also driven strongly by the fact that the intention was to improve diagnoses, the public would benefit and it was therefore seen as a necessary development.

They were also reassured that this application of machine learning had already been shown to work. Therefore, it was deemed as having low societal risk because participants were confident that misdiagnoses would not occur on such a

[21] Discussions around societal risks and benefits and the individual quadrants from across the dialogue workshops were reviewed in order to produce this summary chart.

scale as to cause societal harm. They were confident that experienced doctors would still have an important role to play – both acting as second check and maintaining the personal touch when delivering news.

> "It's like all these things, they're not being created to take over what humans should do. A machine can help with, 'these are all the possibilities', and then the doctor can go on and use that tool, communicate with the person. It's not the machine saying, 'you've got brain cancer', it's being used as a tool." (Birmingham)

Conversely, self-driving cars were seen as having the greatest risk to society. Participants were easily able to imagine the impact of inaccuracies in the algorithms used in self-driving cars, with physical harm being a key risk. They therefore concluded that whilst self-driving cars could improve travel and traffic conditions on the roads, as well as offering independence to those who are unable to drive, the risk of something going wrong was also substantial.

> "If you have driverless cars…you've got to programme them with algorithms, but surely it must learn through accidents etc. for algorithms it doesn't have. How will it learn without having the accident?" (London)

The definition of risk used differed as participants considered different case studies. Risks around machine learning providing results that are inaccurate, or subject to contradiction were highly relevant for some participants. Others tended to focus on the broader risks of machine learning applications to society, including relinquishing control to machines, or being replaced by them in some way. For others, their main priority was managing potential harm to individuals, particularly those that resulted from automated technology.

> "The more that human life is at risk the more human input there needs to be. For things like bus apps, they help you on a day-to-day basis or the Premier League scores that's useful on a day-to-day basis. But with self-driving cars that needs to have more of a human input because if it makes a mistake there's more at risk." (London)

Participants were most worried when they saw the consequences of inaccuracies or being replaced as resulting in physical harm. As such, the applications that involved embodiment, or automation by machines, tended to appear as having more risk associated with them, due to concerns about physical harm. Conversely, the applications that were more 'virtual', such as personalised learning or tailored marketing, had less risk of physical harm associated with them, and tended to be viewed as having lower social risk. The exception was machine learning being used in medical diagnoses – where the consequences of error could be serious. However, as mentioned above, participants were happy that this had been proven to be an effective tool and would have sufficient oversight from experienced consultants, as not to be a risk.

Applications that were felt to have social value tended to be those that could improve public services or lead to broad public benefit, such as in the health, crime, or social care sectors. Participants generally assessed societal benefits based on whether they felt the technology could save time, improve efficiency or improve a service.

In some instances, participants had to make a trade-off between the benefits of improving efficiency and delivering professional services, in order to decide on the overall benefit of a given application. This was particularly the case with areas that required interaction on a human level. Social care, seen by most participants as a more emotional and intimate form of medical support than the health case study, was particularly subject to this kind of analysis.

> "Yes, but it also has loads of potential benefits – you have to weigh them up. It's sort of in the middle. It can be used for some things and not others. Like it's good for lifting people but not for anything emotional." (Birmingham)

Participants recognised that machine learning being used in the home could help with resourcing issues in the social care sector, and ultimately help to provide a better service. However, some were reluctant to consider this example as being as socially beneficial as others, because of the importance of emotional support in a social care setting.

The importance of the personal touch in weighing up the societal benefits can notably be seen in the art example. This was deemed to have the least social risk, but also the least social value. Whilst some participants felt that they would enjoy reading machine learning-written poetry as an individual, most agreed that this new art form would not add much to society as a whole. Participants felt that being able to connect on an emotional level was an important aspect of enjoying art, and as such, consistently placed this in a low social value position on the quadrant.

When assessing the benefits and risks of machine learning to society, participants were typically lead by their concerns. The discussions often began with participants outlining the potential negative impacts to society, and then attempting to see whether any of the perceived benefits could override these concerns.

5 Views of specific machine learning case studies

Participants discussed machine learning through a series of case studies that described some of its potential applications in eight core areas. Each area will be considered in turn in this chapter.

Summary

Participants discussed the use of machine learning in each of the following applications to various degrees. How acceptable they found each was driven by the perceived intention of using the technology, who they thought would benefit, how necessary and appropriate they thought it was, and whether the machine was making an autonomous decision. As such, their focus was on why machine learning was being used and the impact they felt it would have on individuals and society. Participants were much less concerned about the technical aspects of how machine learning applications worked, other than wanting reassurance that they would be accurate.

- Health → The use of machine learning in health was where participants could intuitively see the greatest potential for benefits to individuals and society. They felt that it could improve accuracy and also allow more variables to be considered when assessing physical health conditions than was currently possible with human doctors. However, participants strongly stressed the need for doctors to remain involved throughout the process – acting both as another pair of eyes and to retain the personal interaction they felt was important when receiving news about personal health matters.

- Social care → Participants tended to be conflicted about how to consider the risks and benefits of machine learning in this context. On the one hand, they saw it as an exciting prospect to 'plug the gaps' in a much under-resourced sector. On the other, they cited concerns that an increased reliance on machine learning might detract from human relationships and emotional interaction that they considered important in this area. The best case scenario tended to be some combination of the two – where machines could aid with everyday tasks, to allow humans more time to care for those they were responsible for looking after.

- Marketing → Many participants were very familiar with tailored marketing, but few knew that machine learning was a part of it. Many of the negative comments centred around invasion of privacy, or manipulation into spending more money than usual, and machine learning was seen as playing a role in this. However, some of their concerns were linked to a more general dislike of intrusive marketing, as opposed to tailoring based on machine learning algorithms specifically. Those who were positive about this application felt that it was much better to have tailored advertisements or vouchers that were actually relevant – if they must exist at all.

- Transport → Participants were able to identify the positives of self-driving cars. They recognised that this could give independence to those who were unable to drive, and also that uniformity could result in more efficient travel. However, participants also recognised that the consequences of errors in the algorithms could result in serious accidents and this would raise complex questions about who was to blame. Safety remained the overriding concern and participants generally wanted to see clear evidence that driverless vehicles would be as safe as humans, if not safer, before they could fully recognise the benefits of this technology to society.

- Finance → Participants were universally supportive of machine learning being used to monitor their transactions in order to identify patterns and spot any unusual activity that may be fraudulent. However, only a few were positive about the idea of machine learning providing an advisory service – warning people against spending money when their balance was low, or stopping the transaction altogether. This was seen as potentially intrusive, and participants generally wanted this to be something individuals would choose to use, rather than it being imposed on them by banks.

- Crime → Participants tended to think that machine learning being used to predict future crime spots was a good idea in principle, but would not work in practice. They felt that machine learning would be useful to 'plug the gaps' in an under-resourced police force, but they also feared that machine learning might result in statistical stereotyping.

- Education → Participants were broadly positive about the idea of being taught as an individual, based on tailored recommendations from machine learning algorithms. Concerns centred on 'pigeon-holing' and people becoming specialist at too young an age, to the detriment of general skills. Participants generally felt that machine learning could be used to spot patterns in attendance and grades, and flag any concerning issues to the teacher. They felt that delivery was still the teacher's job and were keen to maintain this inter-personal element.

- Art → In general, participants could not see the point of machine learning-written poetry. With all the other case studies, they could see scenarios where a machine might be able to do a better job than a human – but they did not think that this was the case with creating art. Those who saw creating art, or consuming it, as a form of personal expression were particularly concerned about the idea of machines performing these tasks.

5.1 Health

Participants discussed the potential for increasing the use of machine learning in the health sector. This included improving cancer diagnosis, and analysing patterns in language and voice tone to detect conditions like Parkinson's disease, and to assess mental health issues.

Two themes emerged during the qualitative discussions as participants considered using machine learning in health. Firstly, views were shaped by a discussion about the appropriate role for machine learning within a patient's treatment pathway. Secondly, it was unanimously agreed that mental and physical health were different and therefore merited separate discussion. It was felt that the observable and quantifiable nature of physical health ailments lend themselves to analysis via machine in a way that the subjective and varying characteristics of mental health do not.

Participants were given the following example of how machine learning can be used in health diagnoses[22]:

[22] 'Stanford team trains computer to evaluate breast cancer', Stanford Medicine News Centre, November 2011, available at: https://med.stanford.edu/news/all-news/2011/11/stanford-team-trains-computer-to-evaluate-breast-cancer.html, accessed 10.6.16

Machine learning in action: Breast cancer diagnosis

In the past, to find out someone's prognosis, three specific features of breast cancer were evaluated, by a human looking at images through a microscope. Researchers at Stanford used a machine learning-based model to measure 6,642 features in the cancer and tissue around. The model performed better than humans in analysing images, but also came up with new, previously unknown features which worked better to predict the outcome for the patient. d

This example was crucial for many to accept that machine learning could actually work in practice. Participants could see how this was an improvement, because there was empirical proof that algorithms were able to analyse many more variables than a human and recognise previously unseen patterns in an important context. The use of machine learning in diagnosing physical conditions was seen very positively, and made intuitive sense to participants. Participants could see the benefits of having this as an aid for doctors when making diagnoses and thought it should be pursued in order to improve accuracy. Participants understood that a machine was capable of processing much more information than a human doctor – a machine could base its decision on thousands of different examples (e.g. of what breast cancer looks like) and could also evaluate many more different factors in arriving at its decision than a human doctor. They understood that this processing superiority meant that the machine's analysis would be more thorough, as well as being far quicker than a human's attempting to do the same type of analysis.

Participants were highly positive about machine learning's potential in the health sector. Their sole concerns related to fears about the loss of human interaction and 'the personal touch'. For participants, there was a clear red line: the final diagnosis and any treatment plan must be reviewed, decided on and communicated by a human doctor. They wanted human involvement in health diagnosis and treatment. In particular, they felt that the personal nature of health meant that any communication had to be done on a personal level.

> "If the machine contacts the doctor, that's fine, you want the human being talking to you to tell you the actual results." (Huddersfield)

Generally speaking, participants were very happy with the idea of machines and doctors working in tandem to provide a better service. However, in one group, participants tested the extremes of a depersonalised health service. They spontaneously brought up the idea of a future scenario where you would step into a full body scanner at your local GP surgery and a machine would tell you everything that was wrong with you, with no human involvement. These participants feared that giving machines a more active role in diagnosis would lead to 'watered down' care in the health sector.

Some participants were not keen on receiving texts with the results of scans or blood tests, or messages suggesting they come in for a check-up. Again, this was because they disliked the idea of loss of human interaction. Participants wanted the reassurance of human oversight and the comfort of a human talking to them in person, even those who thought machine learning had a lot of potential in this area.

> "The early diagnosis side is great – otherwise Parkinson's would only be picked up when more than half the brain cells are dead and it's way too advanced. It's great that the machine is doing the 'grunt work', but I'd still want a human to clarify and confirm it – also to have the personal touch." (London)

Whilst the concerns over the loss of human contact were strong, support for the use of machine learning in health diagnosis was just as, if not stronger.

Participants tended to be less supportive of machine learning being used to support mental health diagnosis. They typically struggled to believe that a machine could accurately diagnose a mental health condition. They found it hard to accept that there would be physical manifestations present in a consistent enough way that a machine could analyse these in order to make diagnostic predictions and recommendations.

> "Mental health concerns me, I work in that field. So it's something I've gone through personally, I can't see how a computer can recognise that… professionals can't even diagnose properly, so how can you trust a computer?" (Birmingham)

These participants thought that the voice recognition technology described in the case study would be flawed and potentially simplistic, only taking into account whether people were using certain words or not. They also felt that a machine would not be able to consider context – for instance, it might be the anniversary of the death of a loved one, on the day that the machine was analysing your speech patterns. Participants pointed to other limitations they saw with the idea, such as an inability to understand different accents, or being unable to use other senses upon which humans rely. There was a sense that there were too many variables to consider in relation to mental health and that a computer would not be able to take account of them all.

> "A lot of these things, the computer may notice that speech is slurred, but the computer won't notice that the person may smell like brandy when they come in the room." (Huddersfield)

However, a small number of participants turned this logic around, to acknowledge a computer's greater processing power. One who lived with a mental health condition had struggled to get a proper diagnosis. She felt that a computer would be better equipped to diagnose mental health, because of its greater capacity for processing data.

> "I had to have 3 or 4 different psychologists before I got a diagnosis. But a computer could do it a lot quicker." (Birmingham)

Underpinning the worries over accuracy was a concern about misdiagnosis as a result of computer error, and the potential consequences for individuals who receive an incorrect diagnosis and treatment. Participants felt that mental health issues were very personal and particular to each individual, and that humans should be closely involved at all stages of the process. For most participants, it was thought to be impersonal and uncaring for people with mental health issues to be assessed by a computer.

> "Mental health is very sensitive so people want to have to deal with people. It's a very lonely place sometimes, so getting a diagnosis from a computer wouldn't be very comforting." (Birmingham)

The use of machine learning in health was where participants could best see the greatest potential for benefits to individuals and society. They felt that it could improve accuracy in prognoses and could allow more variables to be considered when assessing physical health conditions. However, this was also a context where participants strongly stressed the need for human involvement. They felt that machines should be used as an aid in the diagnosis process, but should not be involved in final decisions, or in communicating these to patients. The importance of face-to-face, empathetic interaction was felt to be incredibly important in a health environment and participants did not think that a machine could effectively replicate this.

5.2 Social care

Participants discussed whether machine learning could play a role in social care. They discussed machine learning taking a more passive role – carrying out background tasks, which would allow carers more time to spend with their patients. They also considered machine learning taking a more active role – performing some more intimate tasks, such as lifting patients in and out of bed, or helping them to wash. Participants also debated whether it would be better to have a less engaged human carer, or a kind and attentive robot carer.

Many participants in the qualitative workshops were against the idea of machine learning being used to provide social care. They cited fears over the loss of human-to-human contact, consistent with the findings of previous research. For instance, only 18% of the public are in favour of using robots to care for the elderly and 14% to use them to care for children.[23]

> "It's the word 'care', sometimes it's not about the physical aspect, it's the human aspect." (Huddersfield)

Participants felt that social care should be about an emotional relationship and human interaction – with an emphasis on the *care* element. It was argued by many that a robot would never be able to replicate this role. For some, it seemed undignified to consider a robot helping an elderly or disabled person with things such as bathing or going to the toilet. Those who had had positive social care experiences often could not see a robot being able to perform similar tasks to the same standard.

> "Someone physically there for my Mum is highly beneficial – moral support, TLC … especially with a terminal illness. You need someone to show them care – a computer can't do that." (London)

But some participants considered this case study in a different way, pointing to the perceived deficiencies in the quality of care received by many older people. They discussed the following choice: is it better for an older person, who only sees their carer a few times a week, to have a robot carer on-hand to meet their needs, or limited, perhaps low quality human care? Some argued that if machine learning could allow robots to provide a good standard of care at reduced cost, then it would be immoral not to offer this as an option to those who might benefit.

> "I'd like that – I can't cope and I'm not getting any help at the moment […] Patterns in your voice, moods, the temperature of your skin … it could all be translated and could notify a nurse or a doctor. The more disabled you are, well, people would be able to be involved in the world more than ever before." (Birmingham)

Despite these differences in opinion, there was consensus that machine learning could play a supporting role in social care. Participants felt that machine learning could be 'an aid not a replacement', with many suggestions that a machine and a human social care worker could work side by side to provide a better service. This was seen as a highly positive solution, so long as this would free up care workers to take the lead on providing genuine support and meaningful human interaction to people who need social care.

> "Mundane tasks could be done by a robot, so the human could give more 'quality time'. A person being cared for should not feel uncared for." (London)

[23] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

5.3 Marketing

Participants were asked to consider the role of machine learning in tailoring marketing based on previous behaviour, and drawing on the preferences of other people who have behaved in similar ways. They also discussed the potential for call centres to employ voice recognition analysis in the future, to determine people's moods in order to improve customer service.

Participants were very familiar with the idea of tailored marketing strategies, but were often unaware that machine learning was the technology behind these approaches. Most had seen advertisements online, based on things they had previously been searching for, or knew of supermarket loyalty schemes that predicted future shopping habits. Some thought that tailored adverts and recommendations saved them time when they were shopping, and stressed the convenience of this type of service. A few argued that this could save them money in the long run, as they could take advantage of tailored, relevant deals that they may not have found (or been offered) otherwise.

> "By just going on the website it says, 'right – this is the type of thing you like'. It makes it faster. It's about the business catering for you." (Birmingham)

However, most participants saw little benefit to consumers from machine learning being used to inform marketing practices. They pointed out that some of the other applications discussed had more obvious benefits for individuals or society, such as improving health or education. Increasing the use of personalised marketing was not thought to fill an unmet need – even if it could make their purchasing experiences more convenient.

Many of these participants had a problem with marketing per se, and not just marketing that uses machine learning. As a result, they were not keen on tailored recommendations, because they felt like they were being told what to buy. Reasons for opposition to marketing techniques were similar to those opposing the financial advisory service, discussed below. There was concern amongst some that machine learning could be used to influence people's behaviour and encourage them to buy unnecessary things.

Our Global Trends data shows that almost two-fifths of people globally have been irritated by an unrequested online recommendation from an online retailer or service. Furthermore, only 30% of people feel comfortable with companies using information provided automatically when they go online (such as location and browsing history) in order to make recommendations.[24] Therefore, the negative reactions towards tailored marketing observed in the groups may be due in part to an aversion to marketing generally, rather than a specific attitude towards machine learning.

> "It's annoying, it drives me nuts. It's an invasion and it's irritating and it infuriates me. It's got bigger in the last few years… I'm a complete shopaholic, but I can make my own decision without a computer helping me." (London)

These participants also tended to dislike the fact that they had not willingly given consent or opted in to this service, but felt they had to give up their personal information in order to access services. Most did acknowledge, however, that adverts are worth it for the free, online content that they fund.

[24] Ipsos Global Trends Survey, available at: http://www.ipsosglobaltrends.com/index.html

This illustrates how participants' acceptance of specific machine learning applications was strongly driven by how they saw the reasons the technology was being used, and the perceived motivations of the organisation responsible. Whether or not participants were acknowledging the machine learning behind the application when they made this decision was secondary. Participants did not distinguish between their views of the application and their views of the machine learning element when they arrived at their conclusions. In short, they did not tend to 'notice' that machine learning was taking place, unless prompted. If applications were seen as socially beneficial, they were far more likely to be supportive. If, however, machine learning applications were seen as primarily profit-driven, then many were against the idea.

> "That supermarket thing … where they recommend products to you, that's like Big Brother. Sainsbury's think, 'we see that you buy cat food, wine' so we're going to try to sell you more of it. This machine learning is actually also Big Brother stuff, isn't it?" (Oxford)

However, there was also inconsistency in views. Supermarket loyalty cards were seen as having benefits for individuals, as they give customers money off items they generally want, even if they were also seen as a way to encourage consumers to spend more money than they usually would.

> "There's benefits like getting money off, but I've found… let's say you pay £50 a week, it's only giving me money off if I spend £60." (London)

Participants generally argued that machine learning-based marketing techniques posed quite a low risk to society, and they also acknowledged the prevalence of these approaches. Indeed, some welcomed the idea that machine learning would improve the adverts they see and in the future might actually make them more relevant. But they also felt that tailored marketing could be a high risk for some more vulnerable individuals. Participants disliked the idea of companies taking advantage of people in this way.

> "What about if you're vulnerable and they manipulate stuff to try to make you buy something? Maybe you shouldn't be spending money, but you are low and vulnerable." (Birmingham)

5.4 Transport

Participants discussed a future where driverless cars could understand their driving choices, and learn from traffic and weather patterns. They discussed the benefits and concerns over cars being able to predict conditions and override human controls, based on these predictions.

During the workshop discussions, it was clear that participants were more familiar with the automation involved in driverless vehicles than they were with the machine learning aspects.

Some participants questioned why using machine learning technology was necessary to help with driving. These participants were usually those who enjoyed driving and were concerned that introducing driverless vehicles would reduce their freedom to carry out an activity they took pleasure in. On the other hand, those who were unable to drive due to ill health, financial difficulties or those who had never learned said that access to a self-driving car could be liberating for them.

Reactions to driverless cars were not limited to personal preferences. Participants also discussed both efficiency and safety considerations.

The more technologically engaged participants argued that machine learning could analyse traffic patterns and make driving and traffic flows more efficient. They recognised that driverless cars would all be programmed to drive in the same way, whereas humans are not – and that this could result in more consistent driving speeds on motorways, for example. They felt that the exact nature of the programme would ensure that traffic could move in a more uniform and controlled manner, thereby resulting in greater efficiency on the roads.

> "I say bring it on! There is evidence that shows that if everyone drove at 20mph there would be no traffic jams. Driverless cars could be programmed to do that." (Birmingham)

Participants also discussed how much technology now helps with aspects of driving in modern cars, from anti-lock braking systems (ABS) to automatic parking. The merits of moving from computer-assisted driving to fully driverless cars caused more debate among participants. Having considered the idea further, most supported the idea of driverless vehicles, if they could be shown to be safer than human drivers.

Participants tended to expect higher standards from machines than humans because they could not otherwise see the point of making the switch to driverless vehicles. They wanted to be assured that the applications would be safe, and this was particularly the case when automation and the machine having greater responsibility were involved. For instance, participants wanted driverless cars to be tested under a range of conditions (such as icy roads, heavy rain, sudden objects appearing in their path) and to pass them all before they would want to see them integrated onto the road.

As such, they generally wanted proof that driverless vehicles were considerably safer than the current alternative before they could fully accept them. The Birmingham participants considered the following hypothetical scenarios:

Scenario A: 'The current state of play'	Scenario B: 'A driverless future'
Cars continue to be driven by humans, as they currently are. In any given year, there will be 1000 deaths as a result of road traffic incidents. These are directly the fault of human drivers.	Everyone is travelling in driverless cars. In any given year, there will be 500 deaths as a result of road traffic incidents. These are directly the fault of machines, malfunctioning.

For many, there would have to be an assurance that driverless vehicles would not cause any accidents at all to be considered worth pursuing, even amongst those who were initially supportive.

> "If I could look at evidence that showed there would be no accidents or fatalities, then I would use it because I quite like the sound of it." (Birmingham)

The second strand to the safety debate was one of practicality. In Birmingham, one participant was particularly interested in driverless cars. He introduced the example of driverless car testing in Los Angeles, where they were concerned about driverless vehicles interacting with 'real' drivers on the roads.[25] The participant was concerned that there would be more accidents if the two types of drivers were mixed.

> "There would be twice as many accidents because driverless cars would follow the Highway Code and drivers don't. The transition period would be really dangerous – we'd have to give everyone driverless cars all at once. Have them everywhere, or don't have them at all." (Birmingham)

[25] 'Google's driverless cars run into problem: Cars with drivers', NY Times.com, 2015, available at: http://www.nytimes.com/2015/09/02/technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html?_r=1

The idea of all cars being driverless was perhaps the greatest conceptual leap for the participants across all the workshops and groups. However as a result of this discussion on the interaction between driverless and driven cars, participants in this Birmingham group were notably less willing to accept driverless cars than the other groups, as they could not see how the tranisition to automated vehicles could be made in practice.

Safety remained the overriding concern. For those who felt that machine learning could never be good enough to predict situations or analyse patterns correctly, driverless cars could never be as safe as a human. They would need to see clear evidence that this was the case, and even then some would have concerns because they would not feel in control if something unexpected happened. References were made to the unpredictability of the roads; participants often asked what would happen if a dog jumped in front of a car, or a strong gust of wind suddenly blew. For these participants, a machine would be unable to predict such events, but more than that, they would not be able to react in sufficient time to prevent an accident.

5.5 Finance

Participants discussed machine learning in finance in two different contexts. The first was for banks to monitor spending patterns to detect fraudulent activity. The second was an automated advice service – warning against purchases when bank balances were low, when large bills were yet to be paid, or at times of the week or month where overspending was an issue for an individual based on their previous spending habits.

Many participants had direct experience of their banks stopping fraudulent transactions. Participants were universally supportive of this idea: partly because this function already exists, but mainly there was also a sense that banks ought to be offering this service. Algorithms were also seen as being far more efficient at monitoring for unusual financial transactions than humans.

> "It used to be done manually by people sat for hours looking at things, but now the hard work is done by the computers, and the human just calls you up to tell you about it." (Huddersfield)

However, the second finance application was more controversial. A few participants said they would consider using the service if it helped them control their spending, assuming it offered advice rather than removing their freedom to choose how they spent their money.

But there were strong concerns about providing automated advice to customers. Participants generally did not like the idea of a computer acting paternalistically and telling them what they should and should not spend their money on. There was a great deal of discomfort at the thought of their day-to-day finances being scrutinised, even by a machine. For some, this feeling of being watched would inspire them to deliberately ignore the advice and make the purchase anyway, as an act of rebellion.

> "I feel like I'd want to buy the shoes just to spite it." (Oxford)

Participants also tended to dislike the advisory approach because they feared that machine learning could ultimately take away their free choice to spend money as they wished. Participants argued that whilst it may begin as an advisory service, banks would eventually justify enforcing the service to control people's spending to stop them getting into debt. These participants saw it as their right as an individual to spend how they chose.

5.6 Crime

Data can be used to predict who is likely to engage in criminal activity, or where this might take place. Machine learning is not currently widely used as a tool by police; there have been trials of this technology, which could be used to analyse patterns of crime in order to predict where future crime might occur. These predictions could then be used to allocate police resources more effectively.

Broadly speaking, participants felt that this machine learning application was a good idea in principle, but they could see considerable challenges as they deliberated further.

They felt that it would be a way for the authorities to gain the advantage on criminal groups, in an era of stretched police resources. However, they were generally mistrustful of the integrity of the predictions. Whilst they liked the idea in principle, they did not think it would work in practice.

Participants were introduced to the idea of statistical stereotyping. Participants were told how machines could cut out prejudices that people might have (such as making decisions on the best candidate for a job, or reducing sex discrimination). However, they were also told how algorithms are only as good as the data fed into them. Using an image-tagging example, they were told how if the majority of photos uploaded by users were of white people, then the machine learning algorithm would erroneously assume that white skin was the 'default'. The result would be that non-white faces would be inaccurately recognised, or mislabelled. Participants understood that the machine was not racist, or prejudiced. They could see that what was at fault was the quality of the data initially fed into the machine to train it to make predictions.

As discussed in Section 3.4.4, participants were spontaneously concerned that machine learning could result in people being wrongly labelled. In terms of this case study, labelling was thought to be a problem in two ways:

- Areas would gain reputations as 'crime hotspots' which would result in people moving away and the area falling into disrepute, potentially driving up crime rates in the longer term.
- Machine algorithms would make generalisations about certain groups in society. This would give credibility to profiling and statements such as 'group X are statistically more likely to commit a crime'. In turn, this would help police justify targeting individuals or groups who have not done anything wrong. Some participants worried that this would result in an invasion of privacy that would disproportionately affect some groups more than others.

  "You're walking the line of racial profiling, which is a really distasteful topic. It's a small step towards isolating certain sectors of society and saying that they're more likely to commit a crime." (Oxford)

Participants' concerns about machine learning being used in a crime setting were similar to those employed in a mental health setting and an education setting. In all cases, participants feared a machine making a generalisation, the result being an individual incorrectly labelled, pigeon-holed or having their freedom restricted.

As with education and careers advice, it was felt that using historic crime data would only serve to reinforce racial stereotypes – in much the same way that historic employment data might reinforce traditional gender roles. There was a key concern for the need to get the data right that was being fed into the algorithm, to ensure it could work accurately and effectively.

The consequences of a machine relying on historic data to spot crime patterns were much the same for participants as those outlined in Section 3.4.4 on machine learning restricting people. Some participants feared that a reliance on machine learning could cause people's horizons to narrow – for instance, if they only ever read books recommended to them based on a machine learning algorithm, they might miss out on a whole other range of genres. With crime, some participants feared that the machine, using historic crime data, would only look for the profiles of people that matched, rather than considering perpetrators from a range of backgrounds and demographics.

A small number of participants had broader concerns that the use of this technology would result in a 'slippery slope' towards a police state. These participants felt that predictive policing would open the door to excessive police monitoring, which they saw as an unnecessary infringement on their rights.

> "A lot of these things can be used excessively and against privacy. The police could use it to listen to you … they're trying to keep watch of everything which they don't necessarily have to." (Birmingham)

However, generally speaking, participants did not fear personal loss of privacy as a result of using machine learning in this way. Their main worry was that certain groups would be disproportionately and unfairly targeted as a result of machine predicutions. This was a risk they balanced against wanting to give the police the tools they needed to tackle crime, with participants reaching different conclusions on that point.

5.7 Education

Participants discussed the potential of machine learning in an educational setting through the idea of a 'personalised learning experience'. The case study focused on online courses, where data collected on test scores, which tasks were completed and demographic data could be used to tailor the learning on offer to the individual. Participants also discussed whether this could be applied to secondary education.

Spontaneous reactions to this case study were positive. Participants, regardless of their own educational experience, could see the potential for machine learning in education. Participants warmed to the idea of being taught as an individual, rather than in large classroom settings.

> "It's teaching a person as an individual rather than just mass. You do need to assess that people are, kids, whatever age, are totally different at using information […] if they can take that further and then use it to assess that person in the right way as to what the qualities of that individual are, then I think that's really positive." (Oxford)

The education example seemed to elicit a completely different response to the mental health example. Participants broadly believed that machine learning would be able to differentiate between individuals and produce positive effects, in a way that they could not believe would be true with mental health. For education, participants accepted the potential of machine learning to tailor services by identifying differences between individuals.

On the one hand, this might be because participants saw the current education system as generalising, already. People are taught in classes, rather than as individuals, and so machine learning's ability to spot patterns in the way people learn was perceived as a good thing, even if generalisations were being made about how various groups of students learn, this was still a more tailored experience than assuming that a classroom of thirty children will learn in the same way.

On the other hand, the concerns around machine learning's inability to accurately predict whether someone is suffering from a mental health condition are quite likely to be borne out of a general misunderstanding of mental health. Despite greater awareness of mental health issues, there is still a considerable stigma attached to the condition – as evidenced by participants' fears over being 'labelled' as suffering from mental health.

However, there were some concerns that tailoring in education might be taken too far and could result in children losing core skills by simply "learning how to learn in one way". Participants expressed some concerns that children would be learning in a personalised way from too young an age and this would restrict their horizons by only focusing on certain things.

> "It might make your choice, you don't even have a choice. If you're being tailored and tailored into this direction, you won't even be aware of what else is out there that might pique your interest." (Oxford)

Participants were clear that machine learning should be a tool used by human teachers, and not used as an alternative way of educating people. Some were concerned that children in particular would lose out on interactions with their teachers and each other. This was similar to the discussion about the appropriate roles for machine learning and humans in other contexts, particularly health.

Some participants recognised that socialisation in particular was an important aspect of education because children learn from each other and work together to solve problems. Other participants highlighted the responsibility of teachers as role models and the importance of being charismatic and inspiring children to learn. They felt that a machine would not be capable of this.

> "The teacher is the person to motivate, and unless the person is engaged by the programme, they won't do it." (Huddersfield)

Other participants recognised the strain on resources in teaching. They felt that teachers were sometimes too busy to identify problems with students. As such, there was strong support for machine learning in education if it could play a supporting role that enabled teachers to spend more time with children. Many participants suggested that machine learning could be used to spot patterns in pupils' knowledge gaps by analysing test results, for example. This would help teachers to plan content in future lessons.

> "It's an alternative. At the moment, teachers are too busy, classrooms are too large and issues with students go unidentified. This could help students to realise their full potential. There's nothing more sad than not realising your potential. I'd be willing to give it a chance." (London)

Others, however, did not feel that this would be feasible. As with the health example, these participants did not think that a machine could be equipped to pick up on context. They felt that it would not be able to provide support in a more pastoral capacity.

> "Why are those students at risk? What are the kids' barriers to learning? Can a computer tell that a child can't learn because their parents are violent and abusive? The children that have issues learning also often have problems at home – a computer couldn't know that." (Birmingham)

There was more widespread support for machine learning in the context of adult education. Participants felt that tailored, individual learning was better suited to adult learners, who had already developed core skills and were aware of their own strengths and weaknesses. Participants stressed the benefits of the courses being free and accessible and the flexibility to learn in your own time and in your own way.

"It's better further on in life where you've found your weaknesses and positives and can learn from them" (London)

5.8 Creating art

Participants were asked to consider machine learning in art, and specifically algorithms that can generate poetry. An algorithm is given examples of poetry and it analyses them to spot patterns in structure and language. The computer learns from these patterns to produce a unique work of poetry, but does not understand the meaning of the individual words. Participants were shown a video that includes examples of a poem written by a machine and one written by Gertrude Stein, without being told which was which. Stein's poem was deliberately abstract to seem 'less human', whereas the algorithm's poem was more conventional and used more emotive language.

Most of the participants believed that the machine learning poem had been written by a human, because of the language used. As a result of considering this example and discussing machine learning in art more generally, two different views emerged among participants:

Group A: 'Reflection'

Those with this mind-set felt that the machine-created work was not really art. They felt that creating art was an essentially human endeavour, as it is an individual expression of personal, human experience. A machine, that could never have human emotions or experiences, could never produce true 'art'.

"The writer or the poet is relating to something that they're going through, have been through, so there's a person's emotions involved." (Oxford)

Group B: 'Reaction'

Those with this mind-set cared more about the effect the poem had on them, and not how it had been written. The machine-written poem gave them more as a reader than the human poem. They argued that they would always prefer a poem that they could relate to, regardless of whether it was written by a machine or a human.

"Art is personal. You see some weird pictures, but it's art to someone out there. It comes down to: 'do you enjoy it'?" (Huddersfield)

Despite their differing mindsets, participants agreed that machine learning poetry was not particularly risky to society, but they also felt that it had very low social value. Much of this was due to the fact that humans can write poetry already, and have done for centuries. Some of the the other case studies emphasised machine learning's superiority over humans when it comes to analysing data. Participants did not think that machine-written poetry would be an improvement on the status quo.

"The examples you initially gave were about things that would take so long, that it isn't feasible for us as humans to be able to ascertain that information. We're now talking about something that we can do and we've been doing for... this is just taking away the last few things we've got. I don't see why it's important." (Oxford)

Some of those who saw artisitic creativity as important to their personal identity – or who were passionate about culture – had concerns about undermining the cultural value of art by using machines to generate large numbers of poems or paintings. However, participants generally did not see machine-generated art causing any harm in the future, but also struggled to see any benefit. They understood that machine learning had the potential to improve the way we analyse data; they did not see this as relevant in the field of art.

6 Machine learning in practice – the policy context

This chapter explores public perceptions of how machine learning should work in practice, including the development of the technology, the ethical considerations around its use, and views of how it should be regulated. It draws on discussions that took place at the end of the qualitative events.

Summary

Participants generally found it difficult to discuss the ethics around and regulation of machine learning, other than recognising that it was important to ensure the risks of this technology were considered carefully. Participants had discussed a range of applications in varied contexts, and developed ideas on how the risks and benefits of machine learning played out differently in these different contexts. They were consequently used to discussing machine learning as anchored around – and specific to – particular case studies. Coming to a general view on ethics and regulation was therefore challenging, as participants could see how these questions might be framed differently in each application area.

Participants therefore found it challenging to discuss the ethical framework which should govern machine learning, particularly the safeguards that would be needed if machines make important decisions independent of human involvement. They found it hard to imagine how a machine could behave 'ethically' because of the subjectivity they thought was involved in ethical judgments. As such, their views of the ethics of machine learning often returned to the extent to which humans would still be involved in the process.

While regulation was considered important, there was no clear consensus about what this should look like in practice. It was felt that the technology should not be allowed to advance without oversight, to ensure that it was not being abused or was not being portrayed as accurate, if this was not the case.

The breadth of possible machine learning applications made it hard for participants to come to a general view about regulation. Most expected some government involvement, but tended to prefer an independent regulator or regulators funded by – but ultimately separate to – government. They also highlighted broader regulatory issues related to machine learning, including where agencies or companies are passing data to one another.

Participants assumed that government would have a role in research around machine learning, but expected that the technology that drives it will mostly develop commercially. However, where possible, participants felt the two sectors should work together in its development.

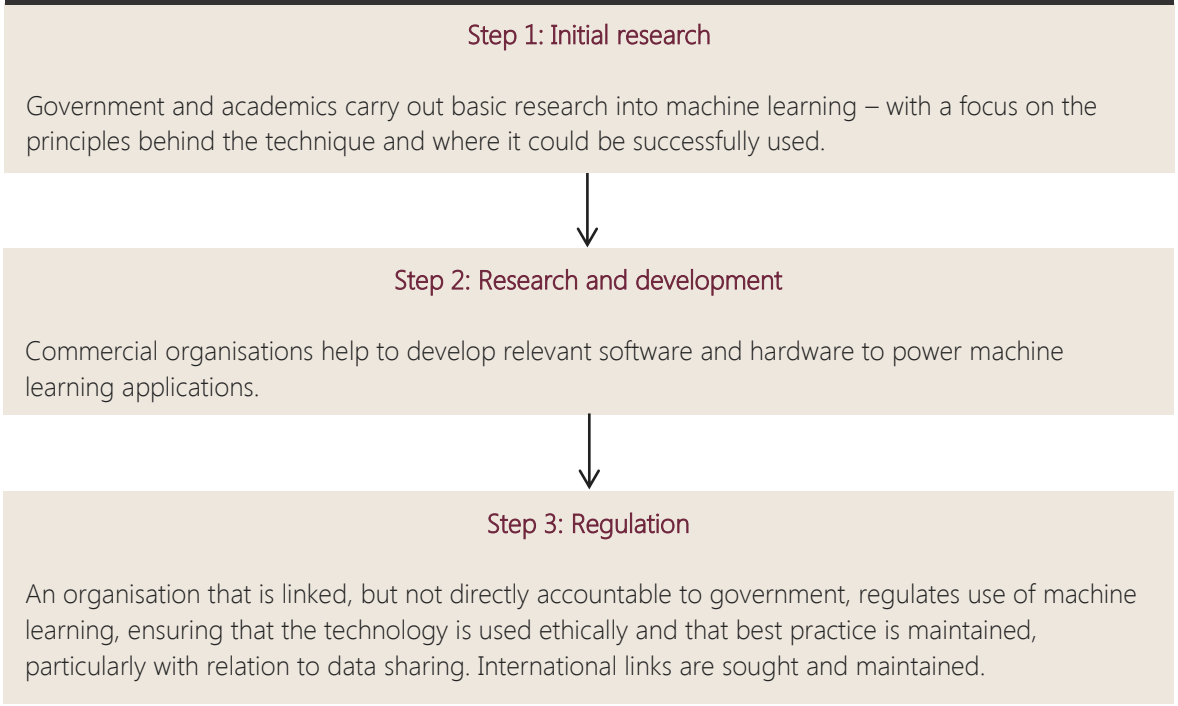
6.1 Taking machine learning forwards

Across the discussions, participants expected that the continued development of machine learning, at least in some form, was inevitable. This was largely because participants became increasingly aware that it was already present in their lives, in the everyday examples, such as how products are recommended to them, or how decisions are made about their finances. However, this was not to say that the development of machine learning for every possible context was

necessarily seen as a fait accompli, particularly when there were moral, legal, or privacy issues associated with some of its uses.

Participants understood why they were being asked to consider machine learning applications. They could see that this technology would have significant implications for individuals and society, and that it was not always straightforward to decide which specific applications were appropriate, and took account of the risks and benefits in a considered way. Figure 6.1 shows how participants' views on machine learning progressed.

Figure 6.1: Participants' views on the development of machine learning regulation



Participants felt that the initial developmental stage of machine learning would be supported by public sector funding. Indeed, it was felt that academics would be the driving force behind it in its embryonic and early stages, until commercial or socially beneficial applications could be clearly identified. These early stages were perceived to be the development of the principles and protocols that will define exactly how machine learning can take place, namely the learning algorithms and the programming that will structure it.

> "There could be a system where organisations could use the info for academic research, and companies like Tesco could use it if they donated to the research – so they give something back even though they're using it for their own good." (London)

Following this initial stage, it was felt that the specific applications of machine learning will be developed by the private sector. This was because of the clear potential commercial uses of machine learning, both in everyday consumer activities (such as in the retail and financial sectors) but also with the applications for the future, such as self-driving cars or 'smart' homes. Many participants assumed that the development of machine learning had already reached this second stage, even if academics would continue to be involved.

Participants drew parallels with the development of communications technology in the private sector, which built on early academic research. It was felt that large multinational corporations in particular had the sufficient resources in their research and development wings to fund more experimental technology, particularly where there is the opportunity to

patent the technology for commercial gain. Their role was seen to be in developing the specific applications powered by machine learning.

However, participants felt that the private and public sectors should work together to develop machine learning. Indeed, some in the groups were not happy that it was the private sector that was developing the technology, as they are not subject to the same scrutiny and accountability as the public sector. There was a sense that this symbiosis would work best with the private sector providing the funding for machine learning to flourish, and the government performing checks on the commercial sector and ensuring it behaves responsibly and appropriately.

> "It's got to have some kind of control in there...and what about people also making mass amounts of money out of it for private profit....Big corporations or private individuals who have the intel and know how, getting in first and making serious amounts of money." (Oxford)

> "The usual thing might happen where public bodies like universities develop something. Then private companies take it, adjust it and sell it back to us. It's not fair." (Birmignham)

6.2 Ethical considerations

Generally speaking, though there was some variation across groups, participants did not come to a clear conclusion about ethics in machine learning.

In considering this technology, participants had been less engaged in the mechanics of machine learning and were more interested in the reasons for using this technology and the consequences of doing so. Having discussed a variety of different applications, participants found it difficult to develop an overall ethical framework for machine learning that was applicable across these different cases. As such, devising a consistent ethical framework that operated in isolation from the technological applications was challenging.

It was also difficult for some participants to grasp exactly what ethics referred to in the context of machine learning; it was used variably to refer to the ethical use of automation, the ethics of research, and ethics relating to the role of machines in society. For example, discussions about the ethical use of automated technologies tended to be framed around who would be most likely to benefit from the technology, particularly where the end user was in receipt of a public service. This tied in with the discussion about the replacement of the personal that we have already touched on – whether or not there is a 'net benefit' in replacing services delivered in person in the desire to provide convenience or save money. It also was couched in a much wider discussion about ensuring that the end user's needs were put first, an important principle that guided many of the discussions, and one that participants often felt the need to anchor their discussion back to.

A separate, and much more challenging discussion was around the ethical behaviour exhibited by machines that learn. Much of this stemmed from the difficulty in understanding how a machine might make decisions independently of human involvement, and, indeed, act on those decisions. During such discussions, participants tended not to see the machine itself as operating either ethically or unethically. Their understanding was that a machine operates based on logic and rules, rather than an ethical code. Participants did not believe that these rules could constitute a framework for ethical decisions – they perceived that this required more than simple logic.

> "I don't see how we can give a robot free rein when there are still ethical dilemmas that we don't agree on universally. I don't see how a machine can make an ethical decision when even we can't make them." (London)

In this context, it was too much of a leap to believe that a machine would be able to *learn* how to behave ethically, irrespective of how much data it is given, and even if it could be programmed with relevant examples of correct 'ethical' behaviour. Context and mitigating factors were seen as too important in such ethical decisions. There was also a concern that we might not understand the algorithms a machine uses and therefore we would not understand the ethical framework it employs.

> "If you are teaching it ethics and it teaches itself then might it not transform the ethics into something you didn't expect?" (Birmingham)

The idea of being able to *judge* machines that learn for the decisions they make was inherently problematic – participants did not accept that machines could be equipped to make decisions to the extent that allows them to be held responsible for those decisions. There was, however, a belief that the person that programmes the machine can be judged based on the information that they 'fed into' the machine.

> "Bearing in mind people have different ethics, the person has agreed on an ethical rule and given it to the machine." (London)

With this in mind, the ethics around machine learning are wrapped up in the application being considered, and the extent to which humans still have involvement. As discussed, participants generally preferred to envisage scenarios in which machines did not act completely independently, and, as such, they wanted there to be enough human input to be able to feel that the machine is being checked – therefore any mistake can be seen as a human mistake for which an individual (or an organisation) can be held accountable.

A good example of this was with assigning credit ratings, an activity which is already based on a complicated algorithm that is not always understood by financial advisors. For such an important issue, affecting individuals' ability to, for example, buy mortgages, participants felt it is crucial that such an application is frequently quality checked.

6.3 Monitoring and regulation – rules and accountability

The quantitative findings, discussed in the next chapter, suggested a lack of consensus about where accountability should ultimately lie, and this was certainly reflected in the groups. In one sense, the participants were unified – in their thinking that some form of regulation and accountability should be in place, in case of error or malfunction.

> "The reason you need a human is you need someone to blame when it goes wrong." (London)

Participants were keen that the government should be involved, but the extent of this involvement was not always clear. However, there was a sense that regulation was critical, and the technology should not be allowed to advance without oversight. This oversight would ensure that the technology was not being abused and would guard against it being portrayed as accurate, if this wasn't the case. Participants felt it was important that this oversight be impartial and independent, and driven by a concern about the welfare of society and individuals, rather than government or corporate agendas. It is worth noting that these discussions were quite abstract for participants. They had no strong views about the specific way regulation would work in practice but seemed to assume a range of regulatory approaches across the different examples of machine learning discussed.

> "I think there needs to be a regulatory body, it's getting bigger and bigger, you have the Data Protection Act, but is there a body that's just looking at this sort of thing?" (Huddersfield)

> "I think that I worry that you can regulate it, but how will you know when to regulate it, you can only regulate things when there's a problem." (London)

Balancing this, many participants thought it was important that government should not impose too much regulation. As such, organisations associated with, but not part of government were seen by some as the best approach.

> "It has to be an institution that's strong enough to stand up to the government and is used to being at arm's length. In an ideal world, institutions would be independent from government even if they were funded by government but we don't live in an ideal world." (London)

There was also an understanding that machine learning would be developed internationally, particularly in the US. Participants assumed that the technological research would be carried out by the large corporations who own the patents, who may have research and development departments across the globe, including in countries where regulation was much more light touch.

> "It's not a UK issue, it's a world issue, the data may be held elsewhere, it needs to be an international body that oversees and regulates. This world isn't this room, it's this world, anything you do is around the world in seconds." (Huddersfield)

This coloured perceptions of the regulation of machine learning, as, should a technology not be allowed to be developed in the UK, then participants speculated that it *would* be in less regulated markets. The fact that machine learning will develop irrespective of borders led for participants to call for greater monitoring by worldwide agencies and better cooperation between nations and their governments to ensure that machine learning develops as a force for good.

> "A country like North Korea is going to be far less concerned about peoples' privacy." (London)

Given the difficulties many participants had in developing an ethical framework for machine learning, concerns often focused on privacy, consent, data security and transparency – these worries were more immediate and personal for many.

6.3.1 Privacy and consent

One aspect of machine learning and algorithms that participants discussed was the use of data, and related concerns around issues like privacy and consent. While this was not the focus of the research, it was something that they returned to throughout the qualitative events.

Participants often discussed data as being something relatively new – a product of the digital age and increased use of online services. Most understood that it could be personally identifiable or anonymous and that many different people's data could be aggregated for analytical purposes. Some grasped the potential financial value of their personal data.

> "Information is a currency in itself nowadays [...] Just because the amateur people, random people off the street don't know the value of their information, it doesn't mean it doesn't have a value. The people who collect and sell it on have the value. When we get things for free, they're not for free – we give them information to use the service." (Birmingham)

However, many of the participants had not considered how their personal data could be used, or how much of it was publically available, prior to the workshops. Some were concerned that they did not know how much of their data was 'out there' or what it was being used for.

Participants found it difficult to come to any conclusions about who owns individuals' data. The two main suggestions were the individual to whom the data relates and the company that has collected or holds the data. Despite wanting to assert some sense of ownership, participants acknowledged that they did not own their data in any practical sense; the companies that hold personal data have the ability to use it and do so. Others referred to the fact that we generate data online, and speculated that we no longer have the right to privacy if we share things in a public forum.

> "We all own our own data – where we were born, our birthday." (Birmingham)

> "I don't believe in privacy – whatever you put out there is out there. I think that everything you put in there is kept somewhere." (London)

Those who were more comfortable with sharing their personal data, particularly when accessing services from companies, tended to mind less about privacy and consent. In their view you could not opt-out from sharing your personal data because it was a part of enjoying the benefits of the service.

> "I think it's a good thing. My account sets aside all my bill money, and whatever I'm left with I'm left with. That suits me cos I'm prone to over spending. It's a safety guard for me." (Birmingham)

However, for others consent was a key issue – one that wasn't currently being addressed by private companies. These participants stressed the need to be better informed about what their data would be used for, so that they could accurately and confidently give consent for it to be used.

> "If it offers you a choice and makes you aware and you have the decision to accept that choice or opt-out, that's alright. If it makes that choice [...] that I haven't chosen to have, I think that's where it takes it a step too far." (Huddersfield)

Whilst not all participants were convinced that machine learning algorithms could make accurate predictions based on personal data, there was a general sense that data science more generally was something that could be used effectively to improve services.

> "You've reinforced what I thought about big data and how it can really change the lives of everyone in the UK and beyond – extend your life and improve the quality of life. There is a very strong case for throwing resources at these things." (London)

6.3.2 Data sharing

Most of participants' concerns about what their data would be used for centred on data sharing. They understood that machine learning needed data in order to make predictions. Participants therefore assumed that organisations would want to access more data in order to improve the machine learning they carry out.

For example, though the data a supermarket collects about retail purchases could seem relatively harmless, there were concerns about would happen if it was sold on, for example, to healthcare providers. Alternatively, there were concerns that data collected by health providers might be passed on to insurance providers, and cover withheld as a result.

> "Yes, the Tesco card, if they send that to my doctor, and subsequently they restrict my insurance company, that's a problem. It's who sells what data to who." (Huddersfield)

> "What if she isn't allowed to have a liver transplant because her fridge says she always has alcohol? Or her house insurance goes up. It could have negative effects on her." (Birmingham)

There was strong aversion to companies passing or selling on their data to third parties – participants were keen that their data should be used for its original purpose, and that this purpose should be communicated to them at the point of data collection. This concern was prevalent throughout the groups. The only exception to this rule seemed to be data used by not-for-profit organisations to analyse the data for educational or research purposes – such as universities or charities.

> "It shouldn't be passed on to others for them to make assumptions about you. [...] If I sign up for a loyalty card, I've consented to giving this information – the company can use it however they want. But it's immoral for it to be used by other companies." (London)

Participants felt that having strict rules on data sharing would be a way to ensure data was used for the purposes that it was collected to achieve. For example, shopping data, that might reveal unhealthy eating habits should not be passed on to insurance companies.

> "You need the protection of a law, and you have that in the Data Protection Act where anything which is held in electronic or written format about you, you have the right to request. [...] the DPA exists to protect you from exactly your data being thrown at people and disseminated widely." (Oxford)

6.3.3 Transparency

Transparency also emerged as a related issue when discussing the regulation of the machine learning industry. Generally, people wanted to be informed about all aspects of machine learning – how the data was gathered, how the process would work (at a level they could understand) and what would happen with the results. There was consensus on the idea that the process should be made public knowledge. This is consistent with other research that shows that the public want to know more about scientific developments – over half of people do not feel informed about science (55%) and a similar proportion think that they see or hear too little (51%)[26].

> "I think it should be available if people want it to be – should be public knowledge." (Birmingham)

Participants felt that there was not currently sufficient transparency when it came to finding out who had access to their personal data for any purpose, including data science and machine learning. They found it difficult to access relevant information, with most stating that it is hard to find, and, when it was found, it was often complex to understand.

> "If you go on Google, it's half an hour just looking things up, looking for plain English versions. The question is, 'are people aware of the fact they're not informed? Are they aware of the world around them?'. Then, I suppose it's 'is the onus on us to educate ourselves through legislation, or what kids are taught at school', or would we instead approach it from 'it's our personal responsibility to learn about algorithms just like we learn to cross the road'?" (Birmingham)

When discussing whose responsibility it was to teach people about uses of data, data science generally, and machine learning specifically, participants saw two approaches: the individual actively seeking information, or the companies themselves making this information more readily available and accessible. Participants who had tried to find out what data was being held on them often described it as a laborious process, as they had to go to each individual company. As a result of this difficulty in navigating who has personal data, there were some suggestions that there should be one central place that people could go to find out what information was held on them.

[26] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

> "You'd want to be able to access your own records [...] I should be able to log on and get the information myself, but currently there's no guarantee you're seeing everything that you need to see." (Huddersfield)

On machine learning specifically, participants generally trusted that the algorithms would work, although their trust varied depending on the context as outlined in previous sections. Participants were a long way from understanding the intricacies of machine learning algorithms, but they trusted that there were experts who did understand this. They assumed that if machine learning did not work then it would no longer be used. Instead, safeguards should focus on the ways it is used – rather than on regulating machine learning overall.

> "I don't know how to fly an aeroplane, but I trust the pilot." (Birmingham)

Other research tells us that there are high levels of trust in scientists to thoroughly consider the risks of new technologies before they are used – seven in ten people agree with this statement (69%). However, trust in scientists to follow rules and regulations varies considerably depending on where they work. Whilst nine in ten people trust scientists working in universities, this falls to just six in ten in relation to scientists working for private companies[27].

The other side to transparency and machine learning, was that participants wanted the steps taken by the algorithm to arrive at a decision to be 'unpicked'. For example, the Oxford group used the example of applying for a mortgage. This group accepted that a machine learning algorithm would be able to analyse vast amounts of personal data and consider many different variables when deciding whether someone was eligible for a mortgage or not. However, they also then wanted to be told why they had failed – which criteria they had not met.

> "If they don't tell me why I couldn't get a mortgage, then I can't change anything. It would be so frustrating!" (Oxford)

Another group in Birmingham shared the same desire to be able to unpick a machine learning decision. This group were concerned that there might be an error in the algorithm, and if the decision could not be understood completely by humans, then this error would go unnoticed and continue to unfairly affect people.

> "I think it should tell you exactly where on that algorithm you've gone wrong. It's unacceptable if it can't tell you at what point it's gone wrong; some data could've gone in wrong – maybe lots of people are being denied, based on a glitch." (Birmingham)

[27] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

7 Quantitative survey findings

Between 22nd January and 8th February 2016, a representative sample of 978 adults aged 15 and over across Great Britain were interviewed on the subject of machine learning. Respondents were asked first about their overall awareness of machine learning, then for their opinions on how machine learning should be regulated, and what they thought about the risks and benefits associated with machine learning and its individual applications.

A quantitative survey can complement, or inform, qualitative research. While it does not allow space for respondents to develop a detailed understanding of machine learning, or deeply considered views about how the technology should be used, it generates data about a higher number of participants and allows analysis of how views on machine learning relate to demographics.

In the context of very low awareness of machine learning, in many cases the views given therefore reflect respondents' initial reactions to the technology; it was the subsequent qualitative dialogue that created space for deliberation about its the implications. Questions about responsibility or regulation were also framed around the technology in general terms, in contrast to the application-specific, case study-led approach of the qualitative dialogues.

Summary

Awareness of machine learning among the public is low. Just 9% of people have heard of the term 'machine learning', and only 3% feel they know either a great deal or fair amount about it. The public are more familiar with machine learning's applications – a majority have heard of at least one of the eight examples given (89%). People are most likely to have heard of computers that can recognise speech and answer questions (76%). They are less likely to have heard of computers which can make investments in the stock market by adapting to the financial market (30%).

Men, and the more affluent, were more likely to say they recognised the term than women and those aged 65 and over. This is also the general pattern for the individual applications, with the exception of the social care example – 34% of 15-24 year olds have heard of this, compared with 44% of those aged 65 and above. There is an even split between views of the overall benefits and risks of machine learning, but men, and the more affluent again tended to be more positive about the benefits outweighing the risks.

While most think there is a role for government in the development and regulation of machine learning, there is less consensus about what this should look like in practice. A similar proportion feel that the government should regulate machine learning but not provide funding (37%) as think the government should provide funding (34%). Overall, though, this means that 71% of people think that the government should play 'some' role in the development of machine learning – either through regulation or funding.

There are also mixed views on who should be held responsible when machine learning goes wrong. The two most common answers are the one in three people (32%) who say that the organisation the operator and machine work for should be to blame, followed by one in five people (20%) who think this should be the manufacturer. Few would hold other individuals or organisations involved with machine learning responsible.

7.1 Awareness and understanding of machine learning

The term 'machine learning' is not a familiar one among the public. Just 9% say they know the term and only 3% overall feel they know a great deal or fair amount about it.

This low awareness of machine learning is fairly consistent across age groups, although there is a slight drop off in awareness for those aged 65 and over. Men and the more affluent are more likely to say they recognise the term, and to claim awareness of individual machine learning applications. One in seven men (14%) say they have heard the term 'machine learning', compared to 4% of women. Table 7.1, below, outlines the case studies used in the quantitative research.

Table 7.1: Summary of case studies used in the quantitative survey

Quantitative case studies			
General Computers that can recognise speech and answer questions	Transport Driverless vehicles which can adapt to road and traffic conditions	Crime Facial recognition computers which can learn identities through CCTV video to catch criminals	Marketing Computer programmes which show you websites or advertisements based on your web-browsing habits
Health Computers which analyse medical records to help diagnose patients	Military Robots which can make their own decisions and can be used by the armed forces	Social care Robots that can adapt to the home environment, for example helping to care for older people	Finance Computers which can make investments in the stock market by adapting to the financial market

Familiarity with machine learning applications is far higher than awareness of the term itself. Whilst just one in eleven people (9%) have heard of the term, almost nine in ten (89%) have heard of at least one of the examples given of machine learning applications. Figure 7.1 below shows public awareness of the eight examples presented in the survey. As well as being twice as likely to have heard of the term, men are consistently more likely than women to say that they have heard of specific machine learning applications. The examples where awareness is most similar between men and women are social care (44% to 38%) and medical diagnosis (49% to 45%). Greater claimed knowledge among men is not uncommon when asking the public about emerging technologies.[28] These results may, to a degree, reflect this trend. Those aged 65 or over tend to claim less knowledge about the applications than younger respondents. The exception to this is the social care example, where knowledge generally increases with age (34% of 15-24 year olds had heard of machine learning being used in a social care context, compared with 44% of those aged 65 or older).

[28] For example, in *Public Attitudes to Science 2014*, men were found to be more likely to claim knowledge of the risks and benefits of emerging technology. See the following report for more details: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

Figure 7.1: Awareness of machine learning applications

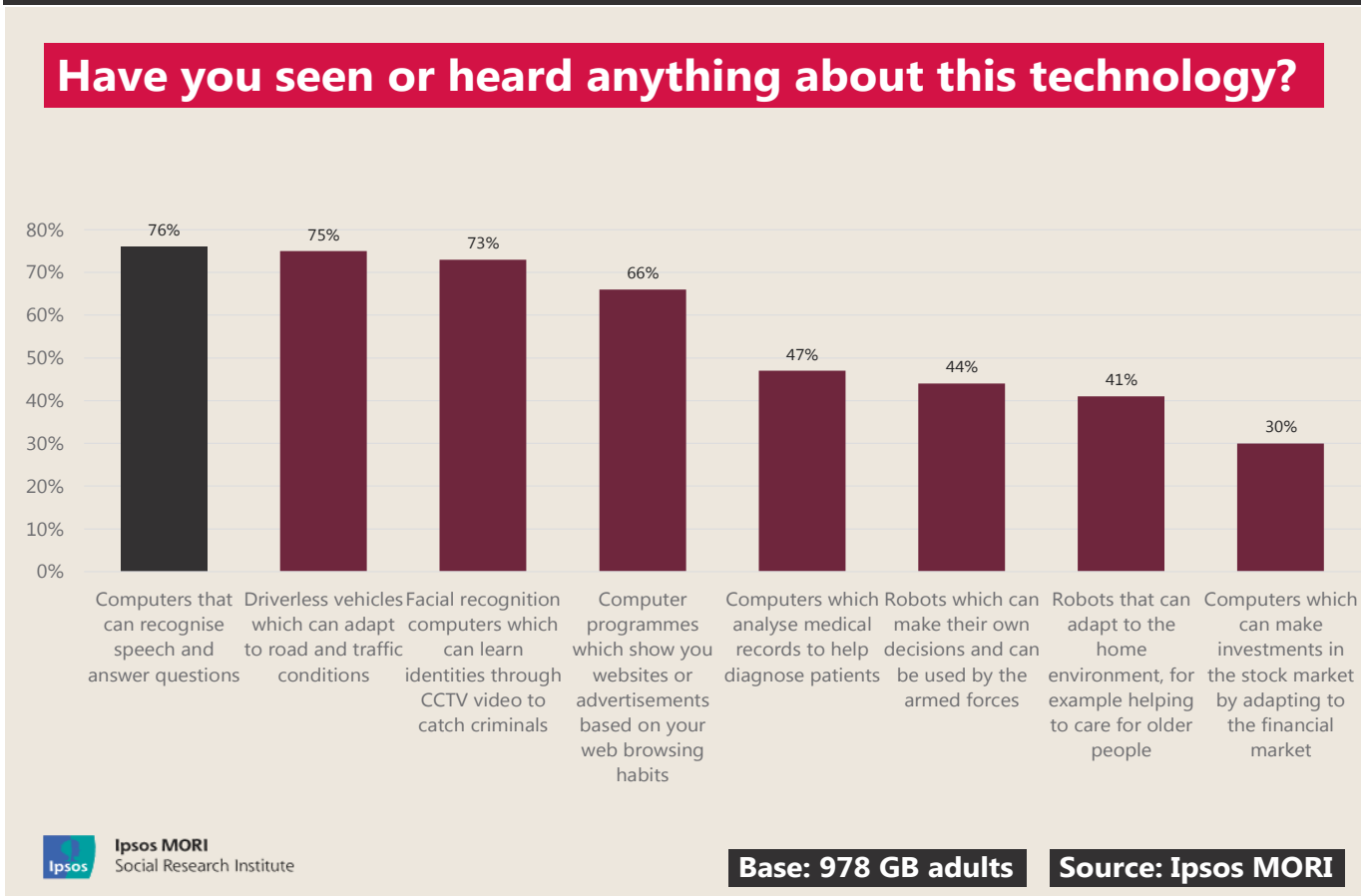
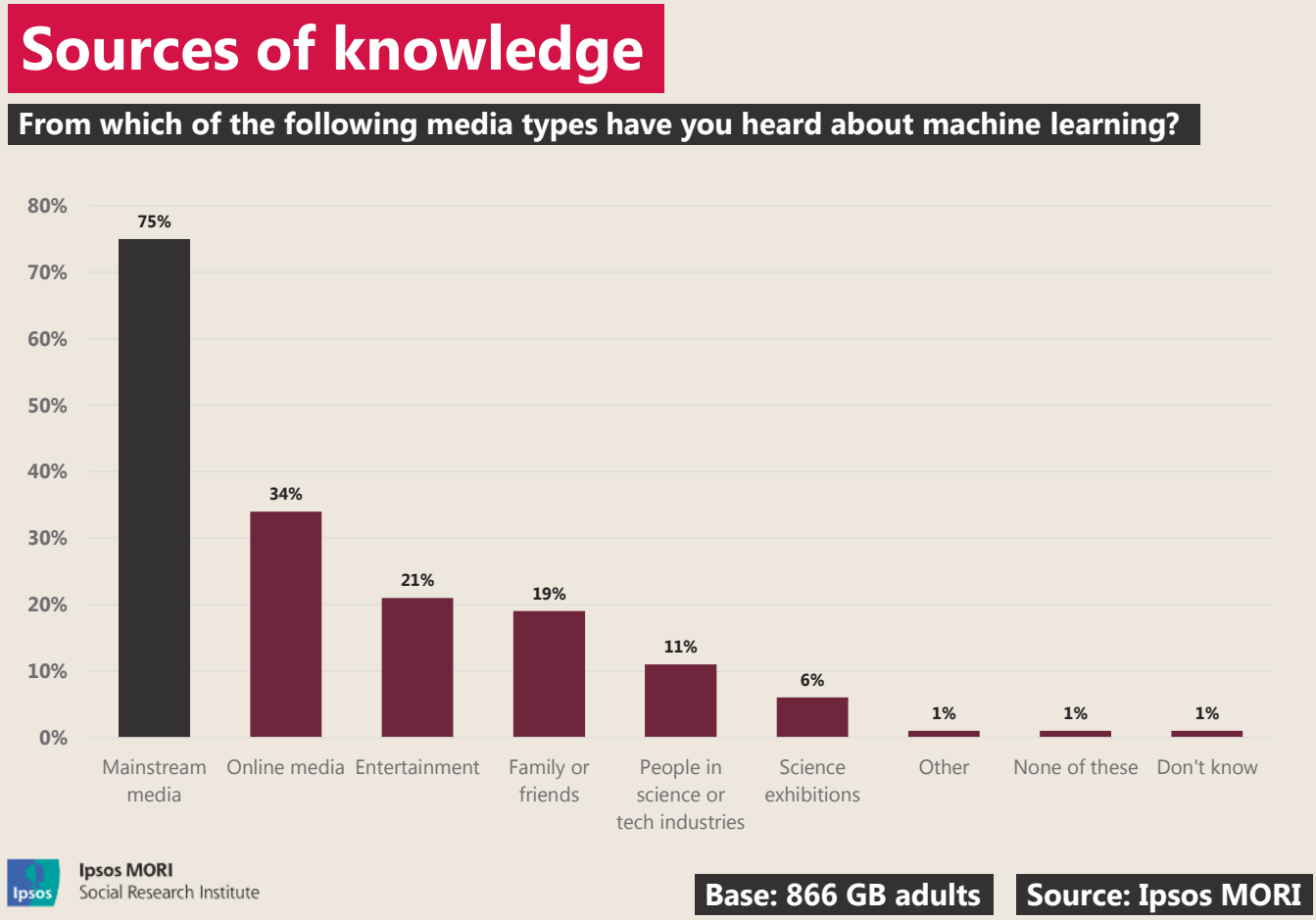


Figure 7.2 shows where participants said they had heard about machine learning from. Three quarters (75%) of those who have heard about at least one example of machine learning say this was from mainstream media (68% of the population overall) and 21% say their source of information was entertainment (19% of the public as a whole). There are differences by age in terms of sources of information – older participants are more likely to have heard of machine learning through mainstream media, whereas younger people are more likely to have heard about the technology online. Around one in five (19%) have heard about machine learning from friends or family.

Figure 7.2: Sources of knowledge about machine learning's applications



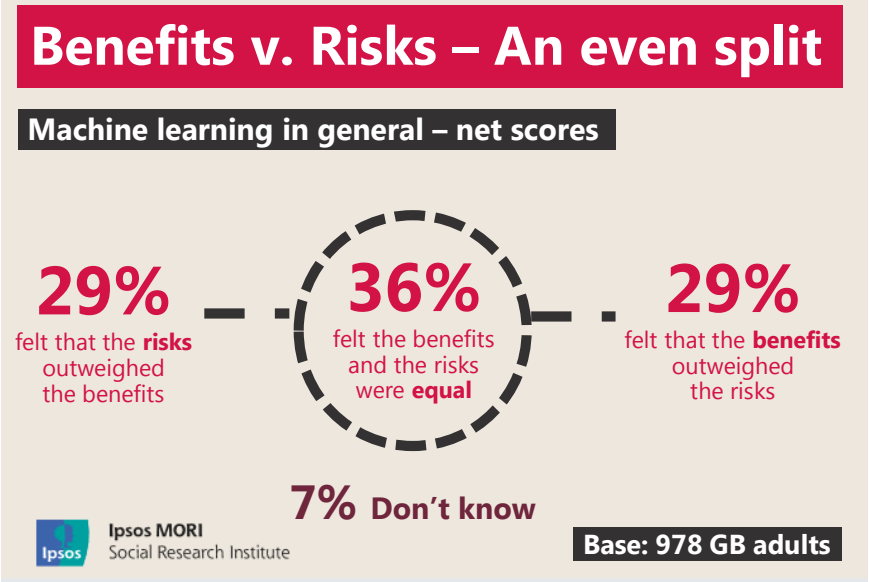
7.2 Considering the risks and benefits of machine learning

As awareness of machine learning was expected to be low, respondents were asked to give their opinions on the risks and benefits of the technology based on the following introduction:

> 'Some suggest that machine learning can benefit society by allowing computers to add to what people can already do, such as diagnosing diseases or making public transport more efficient. Others say there are risks because the learning process of a computer is not always perfect which can present possible dangers if a computer makes a decision rather than a human. Which of the following is closest to your view about the balance of risks and benefits?'

There is an even split when it comes to views on the overall benefits and risks of machine learning. Almost four in ten feel that the benefits and risks are equal (36%). As Figure 7.3 shows, the same proportions take a view that either the risks or the benefits outweighed the other (both 29%).

Figure 7.3: Respondents' views on risks and benefits of machine learning

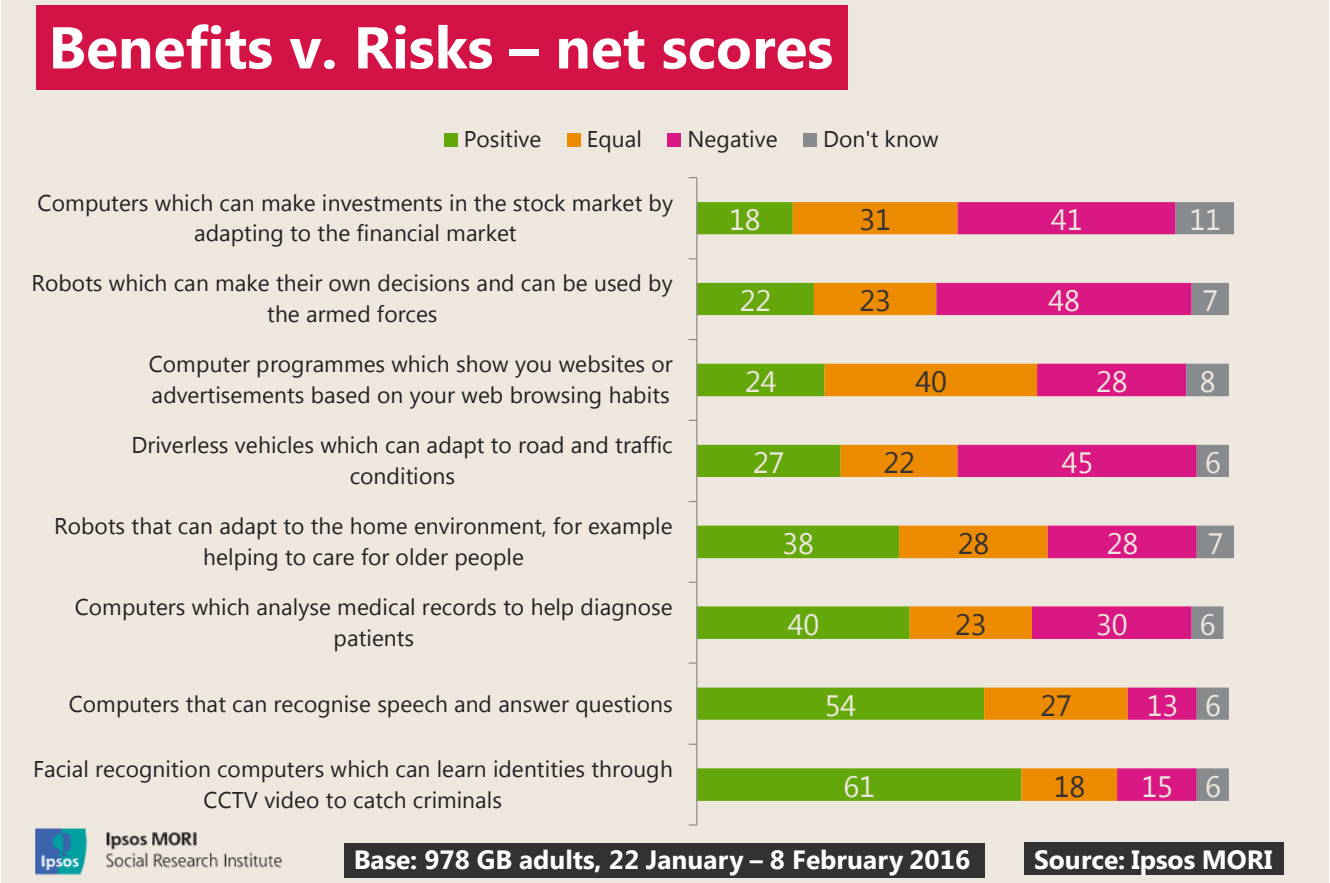


The public's initial opinions on the risks and benefits of machine learning are different to those they hold for science in general. Over half of people think that the benefits of scientific developments outweigh any of the harmful effects (55%)[29] – almost twice as many who feel the same about machine learning.

Again, men are more positive about machine learning than women (36% and 22% respectively think the benefits outweigh the risks) as are those with degrees (37% compared with 19% among those with no formal qualifications) and those in the highest social grades (40% of AB compared with 21% of DEs).

During the quantitative survey, respondents were asked whether they felt that the benefits of specific machine learning applications were greater than the risks, or vice versa. The findings illustrate how the specific machine learning application can make a large difference to assessments of risks and benefits. This emphasises the importance of the purpose for which the technology is used in shaping views, rather than the detail of how the machine learning actually works.

Figure 7.4: The public's initial views on the balance of risks and benefits for individual applications



For several applications, the overall view is that the benefits outweigh the risks. Three fifths (61%) think that the benefits of **facial recognition computers which can learn identities through CCTV video to catch criminals** are greater than the risks, compared with 15% who feel the opposite. Over half (54%) are positive on balance about the benefits of **computers that can recognise speech and answer questions**, compared with 13% who see this as having more risk attached. Two fifths think that the benefits of both **computers which can analyse medical records to help diagnose patients** and **robots that can adapt to the home environment, for example helping to care for older people**, outweigh the risks (40% and 38% respectively compared to 30% and 28%.

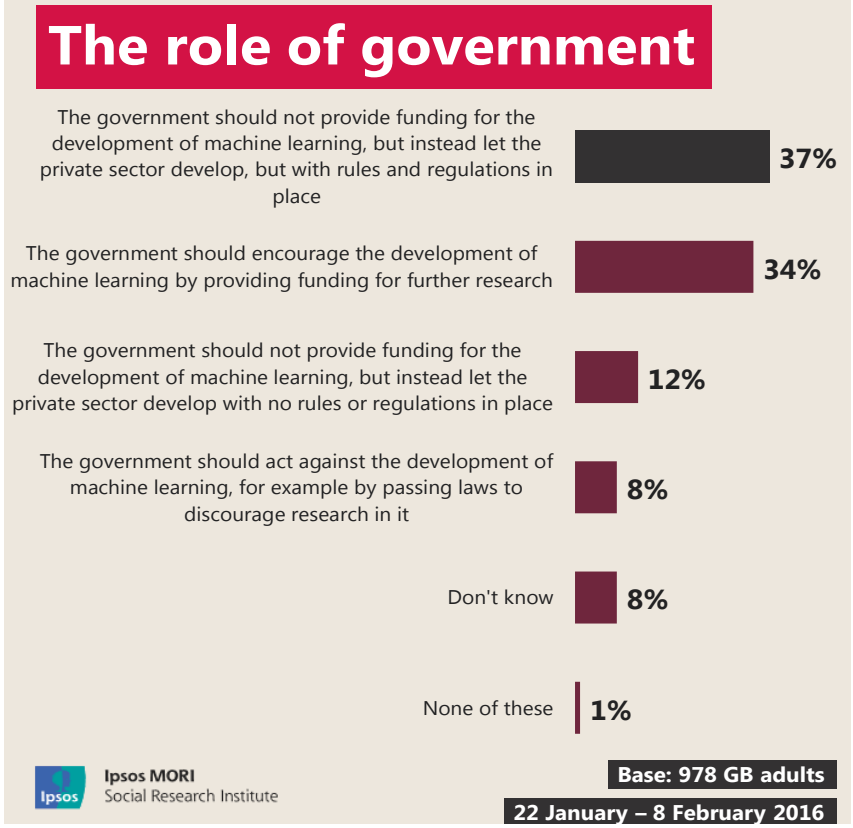
The risks are felt to be greater than the benefits for the remaining applications – 28% feel that **computer programmes which show you websites or advertisements based on your web browsing habits** have more risks than benefits, along with 41% who feel the same way about **computers which can make investments in the stock market by adapting to the financial market**, **driverless vehicles which can adapt to road and traffic conditions** (45%) and **robots which can make their own decisions and can be used by the armed forces** (48%).

For each example, it is always the case that men and people with degrees are more likely to say that the benefits outweigh the risks, compared with women, and those with no formal qualifications respectively. Generally, the more affluent are more likely to feel that the applications are on balance beneficial, rather than harmful. The only exception to this is of **robots which can make their own decisions and can be used by the armed forces**, where 56% of those in social grade AB feel that the risks are greater than the benefits, compared with 37% of DEs.

7.3 The development of machine learning

The findings from the survey show that there is no clear consensus about exactly what role government should play in the future development of this technology.

Figure 7.5: Public views on the role of government in developing



Half of the public (49%) feel that the government should not provide funding for the development of machine learning. This 49% is made up of 37% of who believe that the government should impose rules on the private sector, and a further 12% who feel it should not do so.

However, the results show that 71% of people feel that the government should play some role in the development of machine learning. This is comprised of those who feel the government should regulate the technology (37%) and those who feel the government should fund further developments into machine learning (34%).

The survey findings on the regulation of machine learning are broadly in line with views about the regulation of science

generally. Overall, 71% of people think that the government should play 'some' role in the development of machine learning – either through regulation (37%) or funding (34%). This broadly reflects the high proportion of people who associate the funding or regulation of scientific developments with government (70% and 51%, respectively).[30]

However, support for the government to have a specific role in the development of machine learning is lower than science in general – 34% of people think that the government should provide funding for machine learning, whereas 79% of people think that the government should fund scientific research generally.

Whilst this might reflect lower support for the government to play a role in the development of machine learning, than in science in general, it may also be down to a research effect. In *Public Attitudes to Science*, support for the use of robots dropped when more specific examples were given. For example, seven in ten people supported the use of robots for general military or security purposes (72%). However, support fell to just five in ten when the specific example of unmanned planes being used in military operations was given (53%). These figures also need to be looked at in light of low awareness of machine learning, as people are unlikely to favour public funding of something they know little about.

An interesting picture emerges when these findings are looked at in light of perceived risk. Those who feel that machine learning carries a net risk overall are more likely to call for regulated private sector development (46% of those in this group feel this way, compared with 37% of the public overall). However, those who feel that it carries a net benefit are more likely to feel that the government should provide the funding (43% compared with 34% overall).

Without any other information, or definition of 'risk', these findings suggest that where machine learning is perceived to be a risk, the public are more likely to think that the private sector should take the risk, rather than the public sector.

Public Attitudes to Science

Our 2014 report on *Public Attitudes to Science* demonstrates that people seem to know very little about science funding – 70% of people associate the funding of scientific research with government, and 53% give a sole suggestion for funding sources, indicating an ignorance of which types of bodies are likely to provide funding for new research.

This research found that four in five people think scientific research should be funded by government (79%), even when funding may not bring immediate benefits. Qualitative data from the project explains that:

- Government financial support is seen as a more long-term approach, compared to support offered from profit-driven, private companies. The public tend to think that this will lead to greater benefits in the long run; and
- The public think government funding requires a more transparent process than that of private companies.

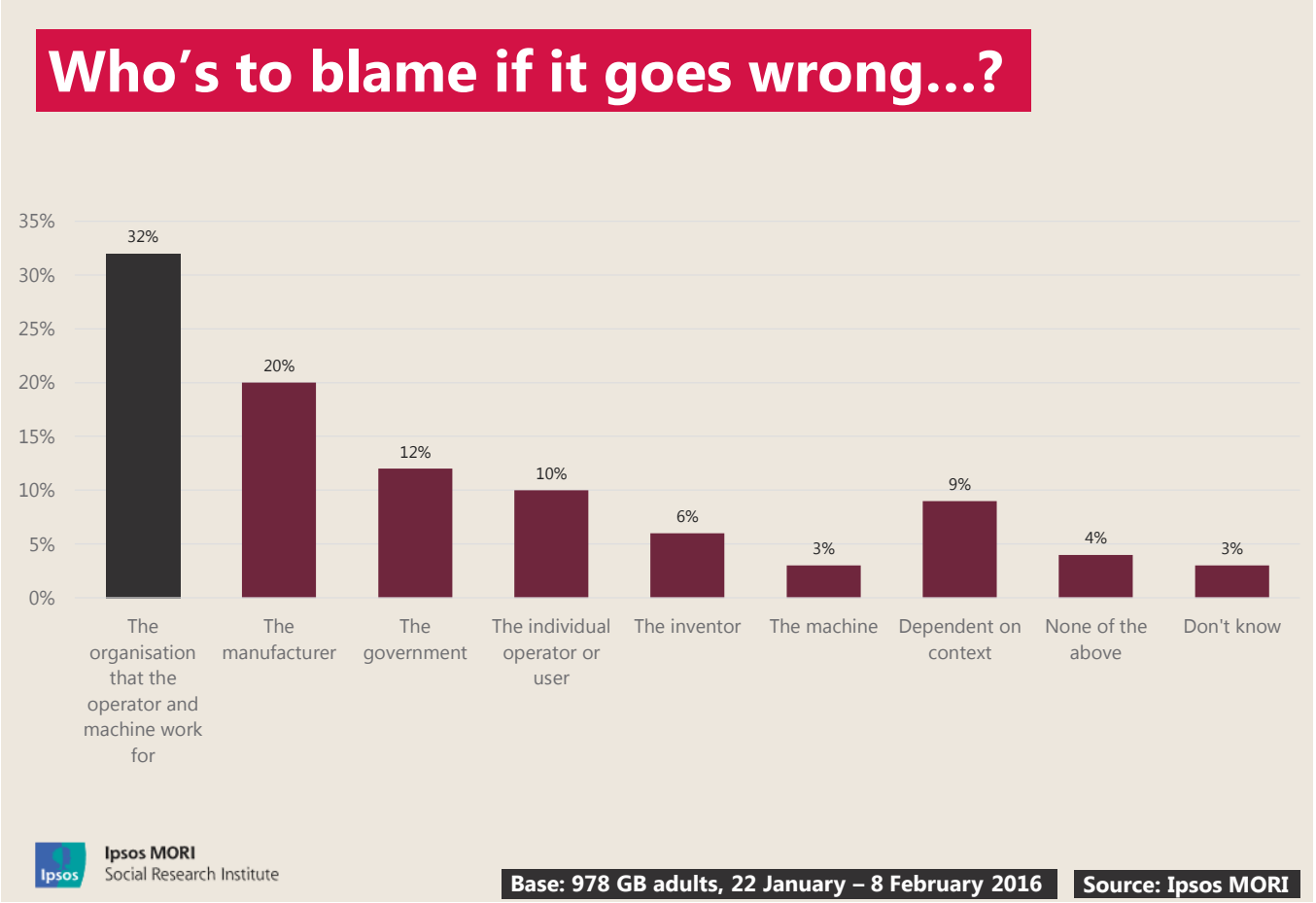
Half of people also think that the government should regulate the science industry (51%), with a further two in five people thinking that regulation should be undertaken by the industry itself – comprised of 25% who say mention scientists and 13% who mention professional bodies.

[30] Public Attitudes to Science 2014, Ipsos MORI, available at: https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

7.4 Monitoring and regulation – rules and accountability

As part of the survey, respondents were asked about who should be held responsible when machine learning goes wrong. The most commonly cited response, given by 32%, was **the organisation the operator and machine work for**. A further 20% would hold **the manufacturer** responsible, and 10% mention **the individual user or operator**. Just 12% feel that this responsibility lies with **the government**, and 3% that **the machine** itself can be blamed when machine learning technology goes wrong. These findings suggest a lack of consensus about where blame should lie – a finding that was also reflected in the qualitative groups.

Figure 7.6: Respondents' views of accountability and responsibility for machine learning errors or malfunctions



8 Key findings

8.1 Initial views of machine learning

Across the research, it was clear that most participants began with little or no knowledge of machine learning and took time to understand the concept. The dialogue workshops were useful in being able to examine public perceptions around machine learning in more detail. They also gave an opportunity to explore the extent to which perceptions of the specific applications and their associated benefits and risks shaped overall views about the potential for machine learning.

There were different spontaneous reactions to the idea of machine learning. As is often the case when considering new technologies, some participants were suspicious about why this technology was being introduced and who would really benefit, and concerned about what might happen to individuals and society as a result. Those who were more technologically engaged were often more open to the potential benefits of machine learning, and less concerned on the potential risks.

Participants had little interest in the mechanics of how machine learning works, beyond wanting to achieve a broad, conceptual understanding. They did not see the underlying algorithms that make machine learning possible as particularly problematic per se. Instead, they were viewed as tools that could be used in different ways, and with different levels of accuracy. Participants drew comparisons with how other new technologies are applied. They expected that machine learning applications would be judged based on whether they work for the intended purpose.

The decisions made and actions taken as a result of applying machine learning were considered much more important. As such, participants often connected machine learning to wider but related issues, including the role of automation and the application of data science more generally.

8.2 Weighing the risks and benefits of machine learning

Overall, participants took a pragmatic approach to how machine learning could and should be applied, discussing the intended purposes, perceived motivations of those using the technology, and the consequences for individuals and society. If they felt a particular use of machine learning was desirable and appropriate in principle, they then weighed up the detailed benefits and risks to decide whether they could support it or not.

Workshop participants used the following criteria for deciding whether they liked an application in principle:

- The perceived intention behind using the technology in a particular context → Participants generally wanted to understand who would be involved with the development of machine learning applications. They felt that the motives and intentions of those involved might shape the success, and direction, of the technology as it progresses.
- Who the beneficiaries would be → Participants were more positive about machine learning when they thought there would be worthwhile benefits for individuals, groups of people, or society as a whole.
- How necessary it was to use machine learning → Some participants struggled to see why machine learning was necessary in some contexts. This was particularly the case where humans were seen as being as good as or better than a machine at completing the task.

- How appropriate it was for machine learning to be used → Participants felt that machine learning was inappropriate in some circumstances, particularly when it involved the loss of valuable human-to-human contact.
- Whether or not a machine will need to make an autonomous decision → If an application would involve a machine making a decision, the seriousness of the potential consequences of that decision was also key.

8.2.1 Considering the risks

The risks were usually easier for participants to identify, and they spent considerable time discussing these in the context of the different case studies considered. For many, the most important risks were around potential physical harm to individuals. There was concern about the possible physical danger presented by automated technology that uses machine learning.

Another broad concern was that machine learning technologies might be developing in a way that undermines or removes valuable personal experiences. This was seen across the case studies in different ways, and was often quite a nuanced concern specific to the things individuals were most personally engaged with. It stemmed from a wider anxiety about the appropriate role for automation, and the potential consequences for human identity. For example, with art created by machine learning, the ability for the artist to enjoy the creation of the art is removed. Or, with self-driving cars, the fun of driving for leisure is also taken away.

Participants grasped that the main opportunity offered by machine learning is the ability to process large amounts of data, and to learn to make better recommendations and predictions by doing so. But some participants found it hard to accept that machine learning could be as accurate and effective as humans in contexts that require more than objective assessments. Participants seemed to feel that humans and machines are so fundamentally different that the algorithms would not be able to capture and interpret a broad enough range of data to reach the right conclusions. This was seen as a particular issue when there was a wider subjective context to consider, which some participants felt could not be approached by simply using logic.

There were also concerns about the wider consequences of relying more on machine learning. While we may be able to understand more about society through machine learning based on large data sets, some participants argued that we could understand less about individuals. They emphasised the importance of not treating people homogenously, or making inferences with significant consequences for individuals based on population-level data.

Participants discussed their unease that machine learning would be used to replace roles that should be done by humans, particularly when it comes to the provision of public services, or in caring professions. Their view was that it would be acceptable – and even desirable – to augment existing services, particularly if this allows professionals to provide better services. But there were much stronger concerns about removing people from processes where their expertise was seen as vital. For example, while machine learning was thought to be an effective tool for helping to diagnose health conditions, this should not mean that it replaces the role of a GP altogether. Similarly, while machine learning could have real benefits in being able to assist teachers in understanding the needs of their pupils, participants felt that it should not be used to replace the teachers themselves.

This led on to a wider issue, which was around machine learning applications being used for service providers in the public and private sectors to reduce costs. Even when efficiencies need to be made, the concern was that machine learning might be used to provide a cheaper, and, by its nature, a less effective, service. And from this stemmed a concern that machine learning could threaten jobs and deskill the workforce generally.

It was also important for participants to consider how appropriate it was for machine learning to play different roles in particular contexts. Instinctive concerns were raised when machines would be making decisions autonomously. This was seen as acceptable in certain low risk situations (e.g. product recommendations or dispensing general financial advice). However, it was felt to be much more problematic in more important contexts, such as actually giving a health diagnosis to a patient or making decisions about people's money, where the outcome is uncertain.

8.2.2 Considering the benefits

Despite these concerns, participants recognised the opportunities associated with machine learning, and the potential for significant benefits for individuals and society.

The power of machine learning to improve the way we analyse data was the most appealing benefit for participants. They fully understood that machine learning could process and analyse much larger datasets than a human ever could, and could do so quickly, even in real-time. Even the more apprehensive participants could see that this processing power could bring real benefits, by being used in conjunction with human judgement. Participants thought that machine learning could make a real difference to our understanding of the world, provided they agreed with the purpose for which it was being used. They discussed applications that received near-universal support, such as better medical diagnosis, improved planning of public services, or tackling some of the big challenges facing society, like population growth and climate change.

Another important perceived benefit was being able to provide more tailored services that take better account of the needs of individuals across both the public and private sectors. Recommendations and adverts could be personalised and made more appropriate to consumers (although there were broader concerns about existing marketing techniques). In addition, consumer technologies that use machine learning could provide a more convenient service. Importantly, this technology was already present in some participants' lives – most recognised services such as Amazon recommendations, and some used functions such as Siri to organise their days. But there was greater acceptance of machine learning technology where participants were able to extend this to more socially beneficial applications such as personalised learning.

Many felt that there was a sense of inevitability around machine learning, largely because of the clear benefits they could see from making better use of this technology. Participants were prepared to accept that machine learning would continue to grow as a technique, and none rejected it absolutely. They did not necessarily see this as a problem, but rather wanted to ensure that individuals and wider society think carefully about how this technology is applied. Participants were keen that its development is encouraged in a way that ensures there are social as well as commercial benefits.

8.3 The future for machine learning

There were a number of 'known unknowns' about exactly how the use of machine learning will grow and embed itself into our everyday life and the policy and commercial landscapes. Many assumed that advances in this technology would be driven by the private sector seeking commercial applications for machine learning. But participants also expected that there would be more academic and exploratory research, and that this would be carried out by academics funded by government.

It was difficult for participants to describe their views about the possible ethical framework that might govern machine learning, in part because the applications were so varied, and participants were less engaged by the mechanics of how the technology works. As machine learning was seen as relying on logic rather than emotion, the main role participants could

see for this technology was in providing objective analysis, with continuing human involvement to provide checks and balances. Participants tended to be more focused on whether or not the applications of machine learning could be shown to be based on good data and to work effectively.

As a result, participants found it challenging to see how machines could make decisions that would have an ethical dimension in the first place. Furthermore, they did not think that machines could be *judged* for any decisions they did make. This is because they were thought to operate within the confines of a logical framework, even if this does develop as the algorithm learns to make better recommendations and predictions. While there was no consensus, participants tended to hold responsible those people or organisations that had either developed or were using a specific machine learning application. As such, the ethics around machine learning seem to be entangled in the extent to which humans are still involved in key stages of the process.

While regulation was considered important, there was no clear consensus about what this should look like in practice. The participants generally wanted to ensure that the right balance was struck between benefits and risks, but the breadth of possible machine learning applications made it hard for participants to come to a general view about regulation. The picture was further complicated by the perception that international organisations with large research and development budgets will be responsible for the growth of machine learning technology. This was seen as making it harder for UK regulation to be effective in isolation. However, participants wanted to know that there would be some oversight, particularly of the applications viewed of as having greater social risk. Most participants expected some government involvement, although not all wanted direct involvement, preferring an independent regulator funded by, but not part of, government. They also highlighted broader regulatory issues related to machine learning, including where agencies or companies are passing data to one another.

Appendices

Online community findings

Following the quantitative survey and the qualitative discussion groups, an online community was run to further explore the public's views on machine learning and to give insight on how best to engage and educate the public about machine learning in the future. This community was largely used to explore concerns in more detail, and its results therefore reflect this focus, rather than the more balanced discussion of benefits and risks which took place in the dialogue process.

About the online community

In total, 244 people signed up to take part in the online community, run with Ipsos MORI's partners, CMNTY. Most of them were recruited using a specialist online recruitment company, with a small number made up of those who had taken part in the discussion groups (for a full sample breakdown, please see Appendix A.4). The community consisted of five weeks of activities, spread over three months.

Due to the nature of recruitment for online communities, all members had self-selected to take part in the research, and will have taken part in other studies in the past. Therefore, it is very likely that the participants were more engaged with the research and with machine learning than the general population. However, they were not experts on machine learning, and their feedback as an engaged public is very useful to the Royal Society and others when considering public understanding of, and reactions to, machine learning.

The overall aim of the online community was to develop and iterate ideas. This included ideas for engagement activities, and space to further explore topics that emerged from earlier stages of the research.

Throughout the five waves of activity, Ipsos MORI tested prototype engagement materials, providing feedback on their design and content to the Royal Society, in order to help shape their planned public engagement exercise. In addition to this, areas of the Royal Society's particular interests were explored qualitatively. The online community involved a range of different activities, using three main methods:

- Discussion forums: The forums involved a moderator posting a topic and participants responding with their thoughts, and engaging with other participants. Conversations were moderated to ensure they keep on track and to probe on particularly interesting or unusual perceptions.
- Stepboards: The stepboard involves a moderator posting a topic or a question, to which participants post their thoughts. They cannot see anyone else's comments until they have posted their own. They then click through to the next 'step' which will be another topic or question. This method was primarily used to test the Royal Society's prototype engagement materials.
- Surveys: These are a useful way to explore top of mind views and see how perceptions change. However, the findings presented here should be seen as indicative, rather than statistically robust. This is because the sample of participants was not designed to be 'representative' of the population as a whole.

This chapter summarises the activities designed to explore the community participants' perceptions of machine learning in general and their overall feedback on communication materials (rather than exploring their detailed views of the draft materials).

A.1 Testing communication materials: broad learnings on engaging the public with machine learning

Throughout the community, participants were presented with several animation scripts and infographics, explaining the basics of machine learning. The Royal Society developed these prototypes with the aim of using them in future engagement exercises, to help explain machine learning to the public and to encourage discussion of and interest in the topic. The community members' feedback was used to develop these materials. Participants' interest in the topic was shown by the early questions that the infographics raised. In particular, participants suggested the following could be useful:

- General reassurance about machine learning (some found it scary or intimidating, or felt that others might);
- Information on what would happen if machine learning technology went wrong; and,
- Any future engagement materials to remain 'jargon' free, to aid comprehension.

Participants responded positively to the use of real-life examples as this helped them to understand how machine learning worked. They also found it interesting to see where machine learning was currently being used. The prototype materials discussed machine learning technology in driverless cars and in product recommendation services:

- Participants found the subject driverless cars engaging, raising a number of questions, including how a driverless car would overcome external influences and factors; wanting more information on the specifics of how they work and the benefits and risks of the technology; and the timescale towards 'a driverless future' – when might driverless cars be considered 'the norm'?
- Participants also found it easy to engage with the recommendation services example, as most had either used this technology personally, or were aware of the idea. Those who had not come across it before were easily able to learn what it entailed. Many participants drew on examples of how they had been recommended films or products that were inappropriate. These past experiences affected their trust of the algorithm behind the recommendation. As with the dialogue findings, some of these reactions were borne out of a feeling that human action couldn't be 'predicted' by a machine.

> "I agree that film preferences are very personal [...] and it certainly doesn't understand that things bought at Christmas don't represent a change in my likes/dislikes."

A.2 Exploring perceptions of machine learning through case studies

Along with the communications testing, the community was used to further explore participants' perceptions of the risks and benefits of machine learning. As with the qualitative and quantitative research, this was done through the use of several case studies. Nine case studies were presented across two waves of the research; on occasion, the overall topics overlapped from week to week, but when this was the case, the case study was presented in a slightly different way. This community was largely used to explore concerns in more detail, and its results therefore reflect this focus, rather than the more balanced discussion of benefits and risks which took place in the dialogue process.

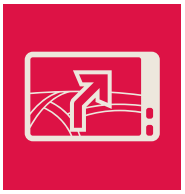
A.2.1 Case studies used in the online community

Personal assistants: People speak into can speak to virtual personal assistants on their smart phones, which respond, or asks for the user to repeat what they said, if it was unclear. The technology works best when the topic of the question is known in advance. It also struggles when people say 'um' or 'ah'.

Social media: Image recognition helps users 'tag' photos of their friends on social media. The machine is trained using images of the same and different people, and is told whether they are pictures of the same people or not. This technology can also be used to search CCTV footage for criminals, but dimmed lighting and the subject not directly facing the camera affects accuracy.

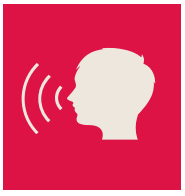
Diagnosing breast cancer: In the past, three specific features of breast cancer were evaluated by a human looking at images through a microscope. Machine learning techniques identified and measured 6,642 features in the cancer and tissue around – performing better than human analysts, but also discovering new features. This technique could reduce cost, improve accuracy and spread expertise.

Monitoring bank fraud: Machine learning acts as a 'data detective' to spot suspicious patterns and halt payments. Fraud costs the financial industry approximately $80 billion annually. Algorithms are trained to spot fraudulent transactions by learning about previous cases and unusual behaviour. The machine can process a large number of transactions as they happen, allowing banks to warn account holders in time.



Maps: Mapping applications use historical data and real-time traffic information to plan the best route for road-users to take. If a main road is busy, these applications might re-route people to a smaller road. They also use GPS data from drivers' mobile phones to provide real-time information on speed and traffic jams. Some people might be concerned about companies using their mobile phone data in this way.

Social care: In the future, machines might be able to help with tasks such as cooking and cleaning – learning an individual's tastes over time. Robots are being developed in Japan that can lift patients from their beds and into wheelchairs, or to help them stand up. In the future, these robots could be used more widely to provide care to the elderly, or even to provide childcare – in and outside the home.



Health – speech recognition and data: Machine learning might be able to help diagnose mental health issues or Parkinson's by analysing speech patterns and identifying features that characterise these diseases. A private company has been granted access to 1.6 million patient records from the past five years to develop an app to help doctors and nurses rapidly identify and treat acute kidney injuries.

Driverless cars: In the future, cars might be able to learn from traffic and weather patterns to predict conditions, or to override controls to brake at certain times. The UK government will be trialling driverless lorries and have committed £1.5 million to test driverless cars in a city centre. We could also see driverless public transport, like tubes and buses. Driverless taxis are currently being developed in America.

Predictive policing: Complex algorithms might be able to detect crime patterns in the future, which would allow the police to resource accordingly. A form of predictive policing technology called 'PredPol' has been tested in Kent; this algorithm re-estimates crime hotspots on a daily basis, using the previous 365 days of crime data, including: type of crime, place of crime, and the time of the crime.

In assessing the case studies, participants focused on:

- How useful they felt the technology was in modern life;
- How comfortable or uncomfortable they were with the technology being used on or by themselves, or to be used in a specific context; and
- How risky they felt the technology was (determined by giving an answer between 1 and 10).

For the final activity, the results of these exercises were presented back to participants and used to explore how they came to arrive at their answers: how they defined 'risk' and why they held these views.

A.2.2 Opportunities for machine learning: Perceived usefulness and comfort towards the applications

Participants saw all of the case studies as being useful to some extent, but the two health-based case studies were seen as having the most potential. Table A.1 below shows the order in which participants ranked the case studies:

Table A.1: Ranking of community case studies by usefulness		
Answers on a five point scale (very useful, fairly useful, neither useful nor useless, fairly usless, very useless)		
1. Diagnosing breast cancer	2. Health – speech recognition and data	3. Social care
4. Monitoring bank fraud	5. Google Maps	6. Facebook image recognition
7. Siri – voice recognition	=8. Driverless cars	=8. Predictive policing

Participants gave several specific reasons for their interest in the health examples, including:

- Producing more accurate diagnoses;
- Reducing human error;
- Reducing costs; and
- Improving survival rates (through earlier diagnosis).

> "It's fantastic, because the earlier cancer is diagnosed, the better the survival rate. So much progress has been made in the diagnosis and treatment of cancer, so this is an added bonus."

Support was particularly strong for the breast cancer example, where participants mostly referred to the greater accuracy of machine learning in terms of the diagnosis and prognosis, when explaining reasons for their support. A couple of participants explained that it was their trust in technology that was behind their comfort.

> "I wouldn't mind this being used on me, as I embrace new technology in science and know it is continually developing."

Despite being overwhelmingly comfortable with this technique being used on them, however, a substantial number of participants qualified their answers, implying that their support was conditional. They explained that they *would* be comfortable *'if'* something else could be guaranteed.

> "If I knew that an early diagnosis meant a better chance of survival, I'd be happy to try this technique."

Using machine learning in a social care setting was seen as the next most useful technology. Participants who reported that this would be useful, said that this was in the context of stretched resources and would also help patients to retain their independence.

> "There aren't enough carers to go round and as our increasing population ages we need to find alternatives, such as this."

Others also saw these benefits, but stressed that the technology should be used in conjunction with human carers – in line with the findings from the qualitative research. Reasons given for suggesting this partnership included: providing oversight of the machine, providing human companionship, and allowing efficient division of labour whereby the machine would undertake certain 'background' tasks, or 'heavy-lifting' tasks, to free up time for, or reduce the strain on human carers.

> "I think there are too many out-of-the-ordinary things that could happen in this setting for a machine to be able to cope. However, the rubbish, mechanical, repetitive jobs that humans have to do (I'm talking about cleaning up body fluids!) could be done by something without emotions or repulsions."

Using machine learning to monitor people's bank accounts to detect fraudulent activity was seen as the next most useful technology. Positive reactions to this technology centred around the machine's ability to analyse a larger number of records than a human could; greater processing speed than a human (perhaps even in real-time); and therefore the ability to stop fraudulent transactions sooner. Many participants saw this as a useful tool for tackling crime and helping victims. However, some participants were less keen, and drew attention to the potential for 'false positives' and questioned how a machine could determine what was and what wasn't a fraudulent transaction.

> "Identifying possible fraudulent transactions could reduce costs and help banks to reduce charges. As long as another method of checking is used to detect possible 'false positives'."

It was common for participants to explain their assessments of the applications as being less useful by drawing attention to the risks involved with each. This conflation of 'uselessness' with risk demonstrates how participants generally assessed the applications in terms of their use to society – by weighing up the risks and benefits, as opposed to simply weighing up the strength of the benefit. For example, participants shared their concerns over invasion of privacy and the potential inaccuracy of algorithms – two key concerns that will be explored in more depth in later sections (8.2.4 and 8.2.5, respectively).

Mapping applications, image recognition on social media, virtual personal assistants and predictive policing were seen as the least useful examples (but were still positively received overall). Participants tended to view these examples as less useful for several reasons (aside from the risks attached) focusing on their limitations, necessity, their personal interest in the technology, and the priority they attached to each.

Participants felt that some of the applications were not yet fully developed and had limitations to their usefulness. For instance, participants commented that virtual personal assistants struggled to understand certain accents and that image recognition systems on social media sometimes tagged the wrong person in images.

They also discussed how necessary technological advances in these areas were. Where it was felt that the applications were not adding anything new, participants assessed them as being less useful, accordingly. For example, participants explained that they could search for items on their phone manually, rather than asking an application to do it for them. They used similar arguments in relation to driverless vehicles; as humans can already drive, they felt that it was not necessary for machines to do this instead.

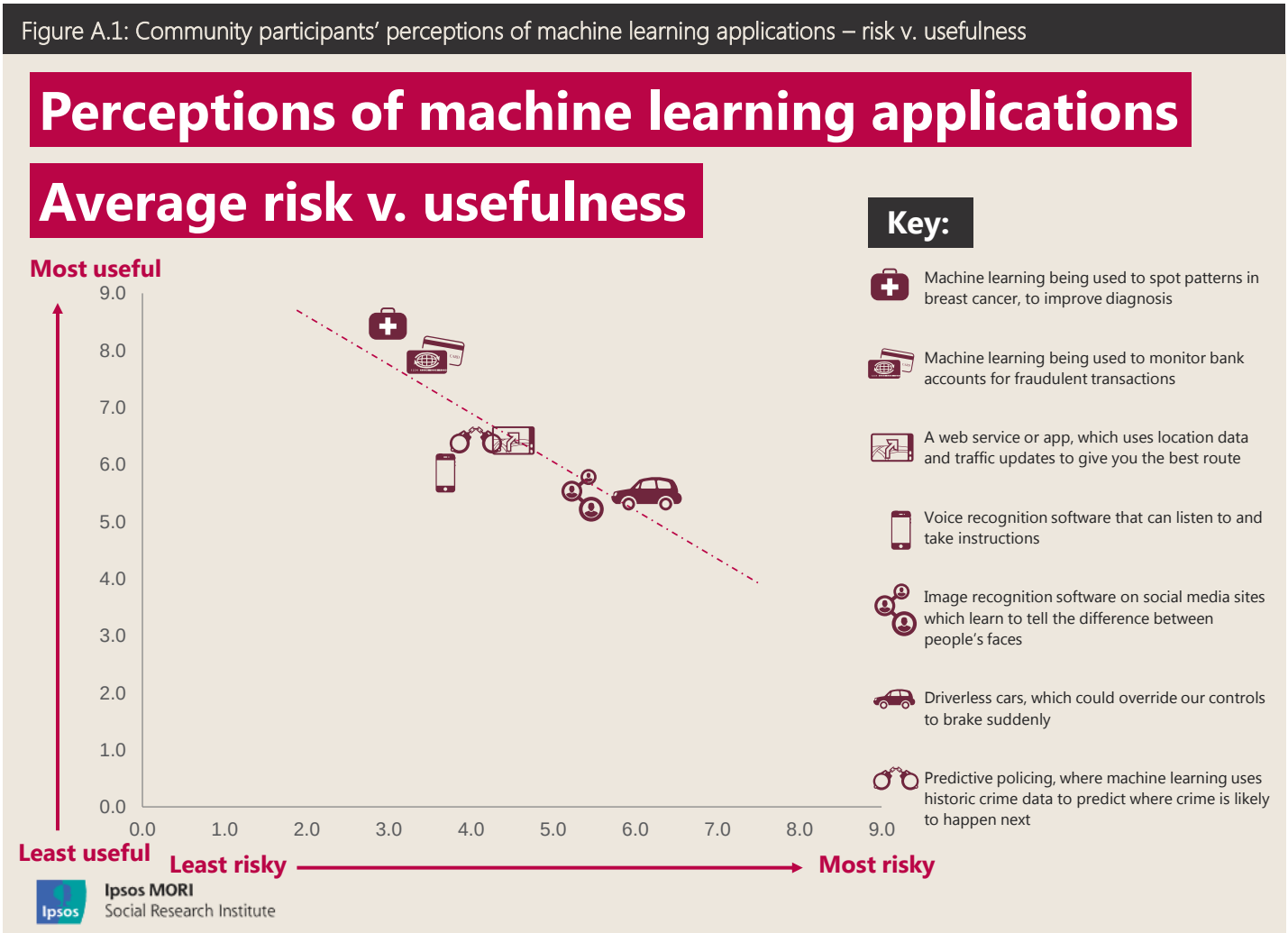
Participants' reflections on the usefulness of these applications also drew from on whether they thought each constituted a worthwhile investment. They also based this assessment on whether they were likely to use the application in question. For example, participants without smartphones (or who used different models) used this to explain why they felt virtual personal assistants were less useful than other applications. In terms of priority, some participants objected to the driverless car example as they felt that the money required to develop the technology could be better spent investing in other areas, such as healthcare or education.

A.2.3 Participants' interpretations of discomfort and risk

After exploring participants' levels of comfort and how useful they thought the case studies were individually, participants assessed both the risks and usefulness of each for society on a scale of 1 to 10 for each.

The average scores for perceptions of risk and usefulness were plotted on the below chart (Figure A.1). Due to the relatively small number of respondents in the survey, these **results are not statistically significant** and should be interpreted **for illustrative purposes only**, showing the significance of context.

Figure A.1: Community participants' perceptions of machine learning applications – risk v. usefulness



This demonstrates that there was a perceived link between usefulness and risk in relation to the case studies presented. These average scores on perceived risk and usefulness were presented back to participants during the final wave of the community, to explore how they defined and interpreted risk – both broadly speaking and in relation to each of the case studies.

Participants discussed whether feeling uncomfortable about an application was the same as viewing it as risky. A consensus emerged that the two feelings often overlapped, but generally speaking 'discomfort' was interpreted as having reservations, or a sense of unease, whereas 'risk' was seen as feeling that there was the possibility of actual harm. This meant that participants generally saw feelings of discomfort as more of an initial, emotional reaction, and views of risk as a more considered, rational response. Participants also noted that discomfort was sometimes a moral or ethical issue, particularly in the case studies involving machines taking on 'human' care-giving roles, such as providing childcare or helping older or disabled people.

Many participants understood risk in a broad sense and also applied it to machine learning specifically. Broadly speaking, risk was defined as:

- A potential for harm (be it: financial, mental, medical, physical, emotional or *perceived* harm) on those either directly or indirectly involved in the original action or decision;
- A negative consequence of an action – either intended or unintended, known or unknown;
- A trade-off between the likelihood of something going wrong versus the benefit of it being successful; and/or
- An emotion: a sense of uncertainty or worry that could undermine confidence in a decision or product.

When considering the risks associated with machine learning, participants drew on these considerations and highlighted specific aspects of the technology that they saw as being 'risky'. However, it seemed that participants' assessment of the risk involved in machine learning was different to how they assessed risk either as an abstract concept, or generally speaking in their everyday lives. The crucial difference seemed to be the risk associated with relinquishing control or decision-making to another entity; in this case, a machine. As was the case with the qualitative phase, participants' initial perceptions of machine learning (as something they were open to or concerned about) did not tend to change as they assessed the case studies in more depth.

> "We, as people, weigh up risks every day (albeit subconsciously in most cases) by our activities, where we go, what we eat, etc. Whereas with machine learning, it's the machine doing it – we don't know how well the machine does it, what exactly it has done and what data it has used to arrive at its outcome."

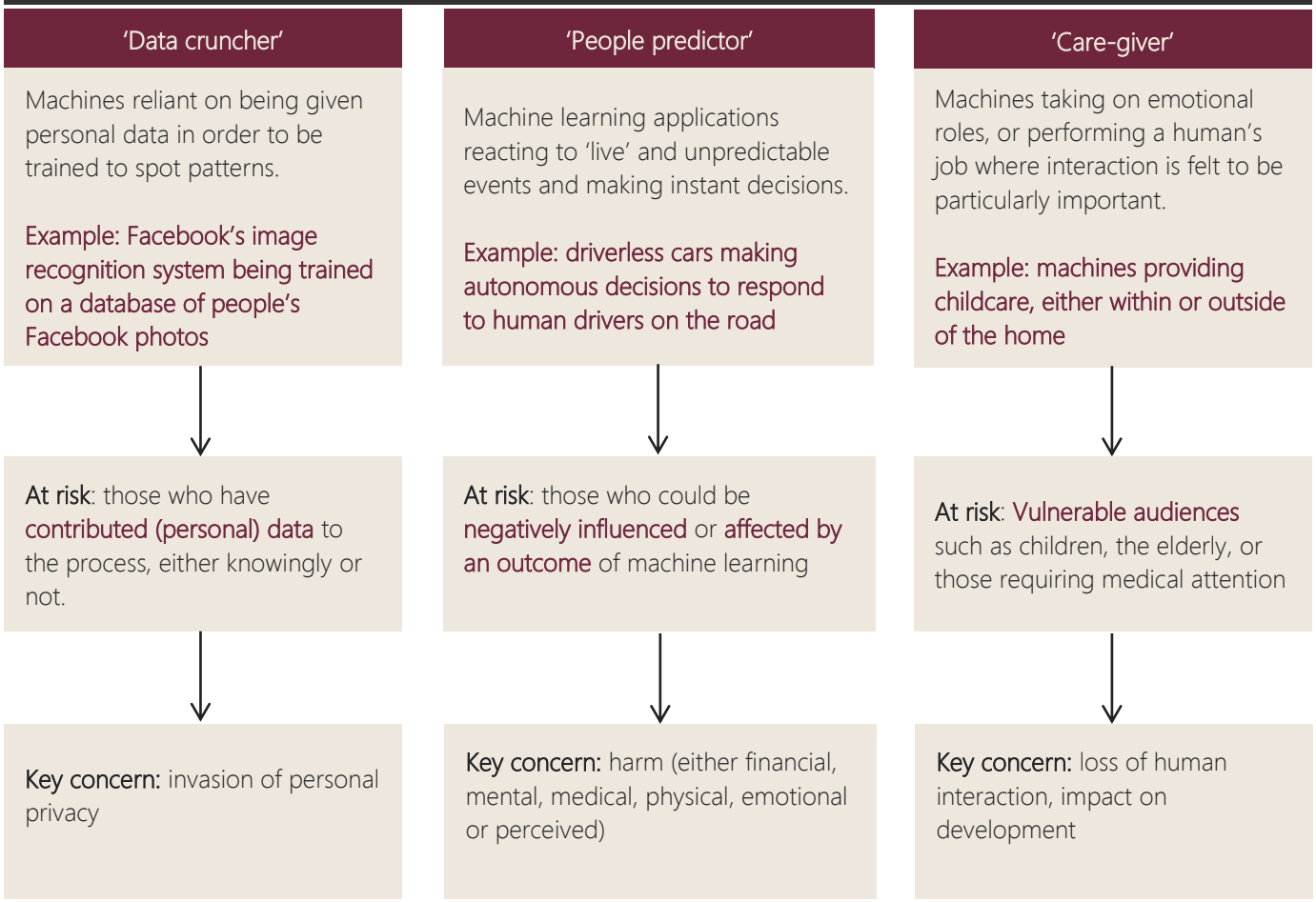
> "We would only see the end results and not how the machine got to that point. People are at risk as you are relying on a machine using data and having to wonder if the data was accurate and all of it was used correctly."

The main specific risks participants discussed focused on personal privacy and harm. There was also some discomfort at the idea of machines taking over from 'human' care-giving roles, which was felt to be inappropriate (from a moral perspective) and risky – in that a malfunction or a poor decision could lead to harm of a vulnerable person.

Participants' definitions of risk in relation to machine learning appeared to be based on their classification of different machine learning applications, as shown below in Figure A.2. The three types of machine learning applications are not mutually exclusive – for instance, some participants were concerned that machines which relied on personal data could

also cause harm. Instead, the diagram highlights the reasoning underpinning participants' key concerns about different machine learning applications.

Figure A.2: Interpretation of risk/discomfort, based on characteristics of machine learning applications



A.2.4 Discomfort and risk – personal data and privacy

Some participants saw risks around using personal data, including issues around: consent, misuse and fearing that individuals' data might be identifiable. Participants understood that machines needed 'training data' in order to learn to spot patterns, but some felt that this use of their personal data was an invasion of privacy.

> "Whenever data is collected, in whichever form that is done, there is always a risk of it getting into the wrong hands, or being used in a way we would not wish or find acceptable. There are many cases each year of companies breaking data protection rules in this country, many do so without knowing they are doing it or realising the seriousness of such breaches."

These concerns were most apparent in relation to the healthcare case study, which discussed two ideas: the use of machine learning in diagnosing mental health issues and Parkinson's, and the use of medical records to develop diagnostic applications. There were very different views of these two examples.

As with the qualitative research, those who were uncomfortable with mental health issues or Parkinson's being diagnosed in this manner tended to be sceptical about the technology's ability to work. They did not trust it and wanted human oversight. Participants also questioned who should have access to medical records which they felt were 'theirs'.

> "People don't know and haven't given permission and I think that's questionable. I would be concerned about whether I could be identified individually… About the safety and privacy of my details/information, about my information being sold or shared with others, i.e. insurance companies, pharmacy research, etc…."

Participants wanted to know whether patients would have been informed that their records would be shared and whether the NHS was able to share records without patients' consent.

Some participants objected to this example either because they felt it to be an invasion of privacy for this data to be shared in the first place, or because they were concerned that they could be identified from the data, and that this might be leaked, or subject to some other security breach. Similarly, some participants were concerned with anyone having access to their location data for use in mapping applications, but were also concerned that they did not know who would be able to access this.

> "This is too much an invasion of privacy, having your every move mapped out by a computer."

Concerns around privacy were also evident in case studies that recorded high levels of comfort amongst participants, including the bank fraud and social media image-tagging examples. In these examples, participants made a trade-off between their concerns over the use of personal data and the service that they would be provided as a result of that data being used to train machine learning applications.

> "I agree that there is a big risk that personal information can be stolen or used by the wrong people, but if I or someone from my family need a diagnosis for serous issues, I want there to be as much information which can help as possible."

A.2.5 Discomfort and risk – lack of trust in the accuracy of algorithms

Discussions of harm took place where it was felt that a machine would not be able to predict human behaviour or react to sudden events. This was particularly the case where the applications would be used in what they considered unpredictable environments, and this concern was stronger where the consequences of an algorithm going wrong were seen as more serious. Some participants expressed concern that the machine would make the wrong decision in reacting to sudden events, and that an incorrect outcome could cause harm. For example, they were concerned about machine learning being used in driverless cars and to provide childcare for this reason.

Their discomfort around driverless car technology centred on a lack of control. They did not trust the ability of the driverless vehicles, and preferred to rely on their own experience and judgement, and that of other human drivers. For instance, participants were almost twice as uncomfortable with the idea of travelling in a driverless car or taxi as they were with the idea of driving a car that could override their control to break suddenly. They were particularly concerned about the ability of machines to adapt to new conditions and their interaction with human drivers.

> "The idea of being transported by a machine that can't think like a human being, when confronted by the actions of a human driver using the same road just scares me to death!"

Similar reasons were given for discomfort about machines providing childcare. Close to four times as many participants were uncomfortable with the idea of a machine providing childcare than a machine caring for older people. This objection was, in part, built on a sense that children were more unpredictable than older people.

> "Kids can manage to get themselves in all sorts of scraps and situations that the elderly don't seem to. I mean, when was the last time you heard about an elderly person with their head stuck in the railings outside a

> school/old people's home? Or an older person having to go to A&E because they have a 5p coin stuck up their left nostril?"

In the social care example, some participants were concerned about the impact that a poor decision would have on a vulnerable audience – either a child or an older person. They identified imbalances in strength and cognition between these vulnerable audiences and a machine that would help them and felt that this heightened the risks of something going wrong.

A.2.6 Discomfort and risk – the loss of human interaction

Participants were the least comfortable with the idea of machines providing childcare. In addition to the reasons outlined above, they were also concerned about machine-provided childcare due to a perceived lack of empathy and reduced human contact. This was seen as being crucial for children's socialisation and development. There were similar concerns around machines being used to care for older people.

> "Where do children learn social skills from? Other human beings with all their foibles, errors and fun."

Participants were concerned that a machine taking on a care-giving role could result in isolation, affecting a child's development or an older person's mental well-being. The 'red line' for participants was not clear, but drawn where they felt that machines would be taking over natural care-giving roles from humans, in regard to a vulnerable audience.

> "Again I feel that if machine learning can HELP in circumstances like these then I would feel comfortable, but if they TAKE OVER and away from human contact for the patient then I would not be content or happy. People of all ages need to be respected, helped and need human interaction. Taking this away from them would undoubtedly have detrimental effects on them."

These concerns were stronger where it was more obvious that a human was missing from an interaction, relationship or process. For example, there were worries over the loss of human interaction for the driverless car case study – a new finding, that did not emerge from the dialogue workshops.

> "The strangeness of it. The facelessness of it. The inability to ask questions about routes, service updated, advice, etc."

Participants would be acutely aware of lacking a human driver on a bus, or someone to interact with on their taxi journey. They explained how a driver represents a visible sign that someone is present – someone who is there to react if something goes wrong, or whose job it is to 'pull the bus over' to deal with any issues on board.

A.3 Mitigating risks and overcoming concerns

When discussing the perceived risks and usefulness of these case studies, and their comfort or discomfort around them, the community participants also explored what might help to assuage their concerns. Figure A.3 below separates their ideas into three broad groups.

Figure A.3: Techniques for mitigating community participants' concerns over machine learning

Information and evidence	Participants needed **more information** to be convinced that machine learning could **work as well as a human** (being as **safe**, if not safer), but also evidence to show that machine learning could be **effective**, without **taking away the role of humans**. **Evidence** was needed to assuage these concerns; participants mentioned **trials** and **triangulation** to verify results. They also wanted to see **endorsements** from **reliable, unbiased experts**.
Human oversight and scrutiny	Participants wanted to see a **governance system** that could ensure ethical use of machine learning. They also wanted the technology to be used **in conjunction with human experts**, and for humans to be able to **override the machine and take control** at any time.
Only time will tell...	Some participants felt that they would **never be reassured** about machine learning. Some could **not trust** the technology, due to knowledge of malfunctions and others were **uncertain about accountability** if the technology were to go wrong. For some, it was not a **priority area for investment**. Other participants felt that **they would probably get used to the technology** over time, as it became more commonplace.

A.4 Online community sample breakdown

The following table shows the sample breakdown of the online community, by characteristics collected during registration. In total, 244 members took part in the community.

		Frequency	Percentage
Gender	Male	100	41%
	Female	144	59%
Age	16-24	10	4%
	25-34	16	7%
	35-44	46	19%
	45-54	69	28%
	55-64	63	26%
	65+	39	16%
	Prefer not to say	1	*
Region	East Anglia	16	7%
	East Midlands	15	6%
	Greater London	30	12%
	North	18	7%
	North West	17	7%
	Northern Ireland	6	2%
	Scotland	22	9%
	South East	36	15%
	South West	24	10%
	Wales	12	5%
	West Midlands	23	9%
	Yorkshire and Humberside	25	10%
Recruitment method	Through our specialist online recruitment company	205	84%
	Through attending an Ipsos MORI workshop or focus group	20	5%
	Through a Royal Society event or newsletter	7	3%
	Other	12	5%

A.5 Quantitative survey – technical note and topline findings

Ipsos MORI interviewed a representative sample of 978 adults, aged 15 and over across Great Britain. Interviews were conducted face-to-face, in home, using Computer Assisted Personal Interviewing (CAPI), between 22nd January and 8th February 2016. Data are weighted to match the profile of the population.

Results are based on all participants, unless otherwise stated.

Where results do not sum to 100%, this may be due to computer rounding, multiple responses or the exclusion of don't know/not stated/refused responses.

An asterix (*) indicates a percentage of less than 0.5%, but greater than zero.

JP01 I'd like to start by asking you if you have ever heard of "Machine Learning", or not?

	%
Yes	9
No	91

JP02 How much, if anything, would you say you know about "Machine Learning"?

Base: All who have heard of "Machine Learning" (94)

	%
A great deal	5
A fair amount	25
Just a little	46
Heard of, know nothing about	25

Machine Learning is when machines or computers have the ability to adapt, learn and make recommendations and decisions on their own without a human giving them ongoing instructions.

JP03 I'm going to read a list of several technologies that use Machine Learning. For each, could you please tell me if you've seen or heard anything about this technology?

	I have seen or heard about this	I have not seen or heard about this
	%	
Computers that can recognise speech and answer questions	76	23
Driverless vehicles which can adapt to road and traffic conditions	75	24
Facial recognition computers which can learn identities through CCTV video to catch criminals	73	26
Computer programmes which show you websites or advertisements based on your web browsing habits	66	33
Computers which analyse medical records to help diagnose patients	47	52
Robots which can make their own decisions and can be used by the armed forces	44	55
Robots that can adapt to the home environment, for example helping to care for older people	41	58
Computers which can make investments in the stock market by adapting to the financial market	30	70

JP04 When thinking about the Machine Learning technologies I've mentioned that you've heard about, which of the following media types, if any, would you say you've heard about them from?

Base: All who have heard of at least one technology (866)

	%
Mainstream media, for example TV, newspapers (including websites), magazines	75
Online media, for example blogs, podcasts and social media websites such as Facebook or Twitter	34
Entertainment, for example books, films, video games (including science fiction)	21
Family or friends	19
People you know who work in the science or technology industry	11
Science exhibitions or events from scientific groups or organisations	6
Through work	1
Other (specify)	1
None of these	1
Don't know	1

JP05 Some suggest that Machine Learning can benefit society by allowing computers to add to what people can already do, such as diagnosing diseases or making public transport more efficient. Others say there are risks, because the learning process of a computer is not always perfect which can present possible dangers if a computer makes a decision rather than a human. Which or the following is closest to your view about the balance of risks and benefits?

	%
The benefits are much bigger than the risks	10
The benefits are slightly bigger than the risks	19
The benefits and risks are both equal	36
The risks are slightly bigger than the benefits	16
The risks are much bigger than the benefits	13
Don't know	7

JP06 For each of the Machine Learning technologies mentioned earlier, can you please tell me which of the following is closest to your view about the balance of risks and benefits with machines doing such tasks?

	The benefits are much bigger than the risks	The benefits are slightly bigger than the risks	The benefits and risks are both equal	The risks are slightly bigger than the benefits	The risks are much bigger than the benefits	DK
	%					
Computers that can recognise speech and answer questions	26	29	27	9	4	6
Driverless vehicles which can adapt to road and traffic conditions	8	19	22	23	22	6
Robots that can adapt to the home environment, for example helping to care for older people	14	24	28	16	12	7
Computer programmes which show you websites or advertisements based on your web browsing habits	7	17	40	17	12	8
Computers which can make investments in the stock market by adapting to the financial market	4	13	31	22	19	11
Facial recognition computers which can learn identities through CCTV video to catch criminals	32	28	18	10	6	6
Robots which can make their own decisions and can be used by the armed forces	7	15	23	20	27	7
Computers which analyse medical records to help diagnose patients	14	26	23	17	13	6

JP07 And which of the following statements comes closest to your view of the role of government regarding Machine Learning?

	%
The government should not provide funding for the development of Machine Learning, but instead let the private sector develop, but with rules and regulations in place	37
The government should encourage the development of Machine Learning by providing funding for further research	34
The government should not provide funding for the development of Machine Learning, but instead let the private sector develop with no rules and regulations in place	12
The government should act against the development of Machine Learning, for example by passing laws to discourage research in it	8
Don't know	8
None of these	1

I'd now like you to consider the following scenario:

JP08a Imagine that you have decided to take your retirement savings and invest them. You take your money to an investment company that uses a computer which, by using mathematical models, can adapt to the financial market to invest your money where it thinks it will provide the best returns. After it chooses where it should invest your money, it automatically invests it without a human being authorising it to do so. To what extent, if at all, would you feel comfortable or uncomfortable with this process?

Base: Half of sample (489)

	%
Very comfortable	4
Fairly comfortable	20
Not very comfortable	29
Not at all comfortable	44
Don't know	2

JP08b Imagine that you have decided to take your retirement savings and invest them. You take your money to an investment company that uses a computer which, by using mathematical models, can adapt to the financial market to invest your money where it thinks it will provide the best returns. After it chooses where it should invest your money, it notifies an investment advisor who will then decide whether or not to authorise the investment. To what extent, if at all, would you feel comfortable or uncomfortable with this process?

Base: Half of sample (489)

	%
Very comfortable	6
Fairly comfortable	28
Not very comfortable	37
Not at all comfortable	25
Don't know	3

JP09 Lastly, I'd like you to think about what happens when Machine Learning technologies go wrong. If that happens, who do you think should bear the most responsibility?

	%
The organisation the operator and machine work for	32
The manufacturer	20
The government	12
The individual operator or user	10
The inventor	6
The machine	3
It depends	9
None of the above	4
Don't know	3
Other	*

A.5 Quantitative sample breakdown

The following table shows the sample breakdown by all characteristics collected in the survey. Please note that these figures are based on an unweighted base size of 978 and a weighted base size of 992.

		Unweighted figure	Unweighted percentage	Weighted figure	Weighted percentage
Gender	Male	521	53%	484	49%
	Female	457	47%	509	51%
Age	15-24	137	14%	153	15%
	25-34	138	14%	165	17%
	35-44	134	16%	156	16%
	45-54	153	15%	169	17%
	55-64	151	27%	136	14%
	65+	265	27%	213	21%
Educational level	GCSE/O Level/CSE/NVQ1+2	264	19%	272	27%
	A Level or equivalent (NVQ3)	186	28%	205	21%
	Degree/Masters/PhD or equivalent	269	18%	286	29%
	No formal qualifications	175	23%	154	16%
Social grade	AB	227	31%	266	27%
	C1	307	20%	266	27%
	C2	198	25%	216	22%
	DE	246	85%	245	25%
Government office region	East Midlands	75	7%	74	7%
	Eastern	67	16%	96	10%
	London	160	5%	129	13%
	North East	46	14%	42	4%
	North West	134	11%	114	11%
	Scotland	106	8%	86	9%
	South East	77	9%	140	14%
	South West	91	5%	87	9%
	Wales	47	9%	50	5%
	West Midlands	92	8%	90	9%
	Yorkshire and Humberside	83	14%	85	9%

A.6 Qualitative sample breakdown

The following tables show the sample breakdown of the dialogue events and the focus groups. Please note that qualitative research does not aim to be representative; a qualitative sample should broadly reflect the population.

A.6.1 Dialogue event 1: London

		Frequency
Gender	Male	11
	Female	17
Age	15-24	6
	25-34	5
	35-44	4
	45-54	4
	55-64	7
	65+	2
Social grade	AB	9
	C1	9
	C2	6
	DE	4
Ethnicity	Asian – Bangladeshi	1
	Asian – Indian	1
	Asian – Pakistani	1
	Black – African	2
	Black – Caribbean	2
	Mixed – White and Black Caribbean	1
	White – British	12
	White – Irish	2
	White – Any other background	6
Children	No children	13
	Older children no longer living at home[31]	4
	Children living at home[32]	11
Working status	Retired	3
	Self-employed	3
	Student	3

[31] This includes people whose partner's children no longer live at home

[32] This includes people whose partner's children currently live at home

	Working full time (30+ hours a week)	11
	Working part time (8-29 hours a week)	8

A.6.2 Dialogue event 2: Birmingham

		Frequency
Gender	Male	16
	Female	11
Age	15-24	6
	25-34	5
	35-44	2
	45-54	6
	55-64	5
	65+	3
Social grade	AB	4
	C1	13
	C2	2
	DE	8
Ethnicity	Asian – Indian	2
	Black – Caribbean	1
	Mixed – White and Black Caribbean	2
	White – British	19
	White – Any other background	3
Children	No children	9
	Older children no longer living at home[33]	6
	Children living at home[34]	12
Working status	Registered unemployed	4
	Retired	2
	Self-employed	3
	Student	3
	Working full time (30+ hours a week)	12
	Working part time (8-29 hours a week)	3

[33] This includes people whose partner's children no longer live at home

[34] This includes people whose partner's children currently live at home

A.6.3 Focus group 1: Oxford

		Frequency
Gender	Male	5
	Female	5
Age	25-34	3
	35-44	3
	55-64	3
	65+	1
Social grade	C1	6
	C2	3
	DE	1
Ethnicity	White – British	10
Children	No children	3
	Older children no longer living at home[35]	3
	Children living at home[36]	4
Working status	Working full time (30+ hours a week)	8
	Working part time (8-29 hours a week)	2

[35] This includes people whose partner's children no longer live at home

[36] This includes people whose partner's children currently live at home

A.6.4 Focus group 2: Huddersfield

		Frequency
Gender	Male	6
	Female	2
Age	15-24	1
	25-34	1
	35-44	2
	45-54	2
	55-64	2
Social grade	AB	1
	C1	3
	DE	4
Ethnicity	Asian – Indian	2
	Black – Caribbean	1
	Mixed – White and Black Caribbean	1
	White – British	4
Children	No children	4
	Older children no longer living at home[37]	1
	Children living at home[38]	3
Working status	Permanently sick/disabled	3
	Self-employed	1
	Student	1
	Working full time (30+ hours a week)	3

[37] This includes people whose partner's children no longer live at home

[38] This includes people whose partner's children currently live at home

Sarah Castell
Research Director
sarah.castell@ipsos.com

Daniel Cameron
Research Director
daniel.cameron@ipsos.com

Steven Ginnis
Associate Director
steven.ginnis@ipsos.com

Glenn Gottfried
Research Manager
glenn.gottfried@ipsos.com

Kelly Maguire
Research Executive
kelly.maguire@ipsos.com

For more information

3 Thomas More Square
London
E1W 1YW

t: +44 (0)20 3059 5000

www.ipsos-mori.com
http://twitter.com/IpsosMORI

About Ipsos MORI's Social Research Institute

The Social Research Institute works closely with national governments, local public services and the not-for-profit sector. Its c.200 research staff focus on public service and policy issues. Each has expertise in a particular part of the public sector, ensuring we have a detailed understanding of specific sectors and policy challenges. This, combined with our methodological and communications expertise, helps ensure that our research makes a difference for decision makers and communities.