Ipsos MORI
Social Research Institute

October 2017

Public views of machine learning: Digital Natives

Supplementary research conducted on behalf of the Royal Society

Daniel Cameron and Kelly Maguire

THE ROYAL SOCIETY

Contents

1 Executive summary

1.1 Background and objectives

Machine learning is a branch of artificial intelligence that allows computer systems to learn directly from examples, data, and experience. 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 commissioned Ipsos MORI to carry out research into public knowledge of, and attitudes towards, machine learning in 2015. This was part of a wider project on machine learning, which aimed to increase awareness of the technology, demonstrate its potential, and highlight the opportunities and challenges machine learning presents[1]. Building on this study, the Royal Society wanted to engage with a specific audience, 'digital natives', to understand whether they had different views on machine learning.

There is no broadly agreed definition of digital natives, but for this study, they were defined as follows:

- **Age:** 18-29
- **Experience of technology:** Using the internet was an important part of their life when growing up
- **Comfort using technology:** Very comfortable using new technology and accessing services online (such as sharing photos and posting on social media, using smartphone apps that track users' location, reading the news, or online shopping)

This supplementary research was commissioned to explore the views of digital natives, and how they differed from those of the general public as a whole[2]. As with the previous machine learning study, the digital natives study focused on: initial reactions to machine learning; the perceived benefits and risks associated with six case studies; the potential value and risks of machine learning for society; and how machine learning should be developed.

1.2 Comparing the digital natives 2017 and the general public 2016 dialogues

This report will provide a snapshot of how digital natives compare to broader society overall, in terms of their attitudes to machine learning. However, it is important to note that being confident about the extent to which any

[1] Research for this project was carried out in 2016. Both the Royal Society's and Ipsos MORI's reports were published in April 2017:

- Royal Society (2017) *Machine learning: The power and promise of computers that learn by example*, available at: https://royalsociety.org/topics-policy/projects/machine-learning/
- Ipsos MORI (2017) *Public views of Machine Learning: Findings from public research and engagement conducted on behalf of the Royal Society*, available at: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

[2] The challenges surrounding differences in views between the two groups are noted in Section 2.4.1

differences in views are a result of life stage or cohort effects is challenging. Caution needs to be taken when comparing the findings from the digital natives workshops with the research carried out last year. This is due to:

- **The length of time between the two pieces of research**: Technology, and machine learning specifically, has developed during this time, and awareness of applications (and their perceived risks and benefits), may be higher with increased media attention (for example, in the field of driverless technology); and
- **The research design:** The digital natives study was commissioned to supplement the 2016 research with the general public. The digital natives research occurred at a specific point in time, without anything to compare to at a similar point in time with previous generational cohorts. We are therefore unable to assess whether differences are down to age or cohort. Furthermore, digital natives were included in last year's dialogues.

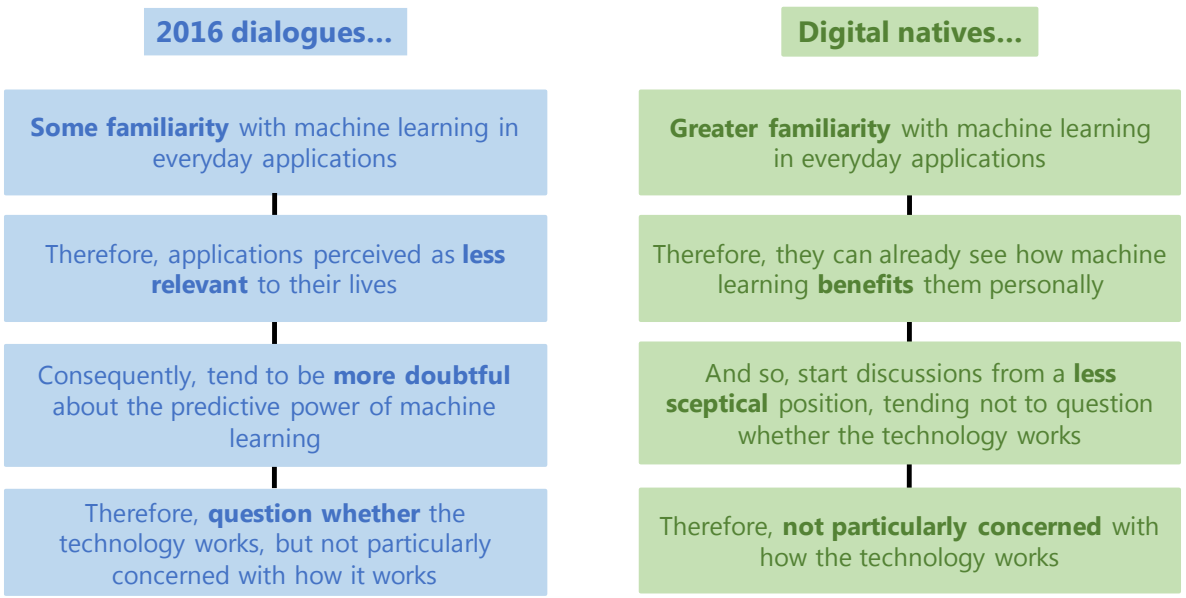
What follows is a summary of the key findings from the digital natives 2017 and the general public 2016 workshops, but it is important to bear these caveats in mind when interpreting the results.

1.3 Initial reactions to machine learning

Digital natives were familiar with some, predominantly consumer-focused, applications that use machine learning – such as recommendation-based services like Netflix and Spotify. They used these applications regularly and being familiar with these examples helped them to understand how machine learning worked in practice. They were also able to speculate about possible real-life uses of machine learning, having quickly got up to speed with the basics of how the technology works.

Digital natives could more readily accept that the technology was already a part of their lives, and felt that its development was inevitable. They were less sceptical about the idea that the technology worked – that computer systems could learn from data and generate new insights – than participants in last year's groups.

The digital natives' greater familiarity with machine learning-based applications meant that their approach to the discussions was quite open, relative to the more sceptical approach of last year's participants, who tended to question in more detail whether machine learning 'worked'.



1.4 Views on specific machine learning case studies

The table below presents a summary of participants' thoughts on the case studies discussed during the 2016 dialogues with the wider public, and the 2017 workshops with the digital natives. Caution should be taken when interpreting these findings: a greater number of case studies were discussed during the 2016 dialogues, and they were discussed in greater detail as these dialogues were longer than the 2017 workshops. Therefore, these findings are not directly comparable.

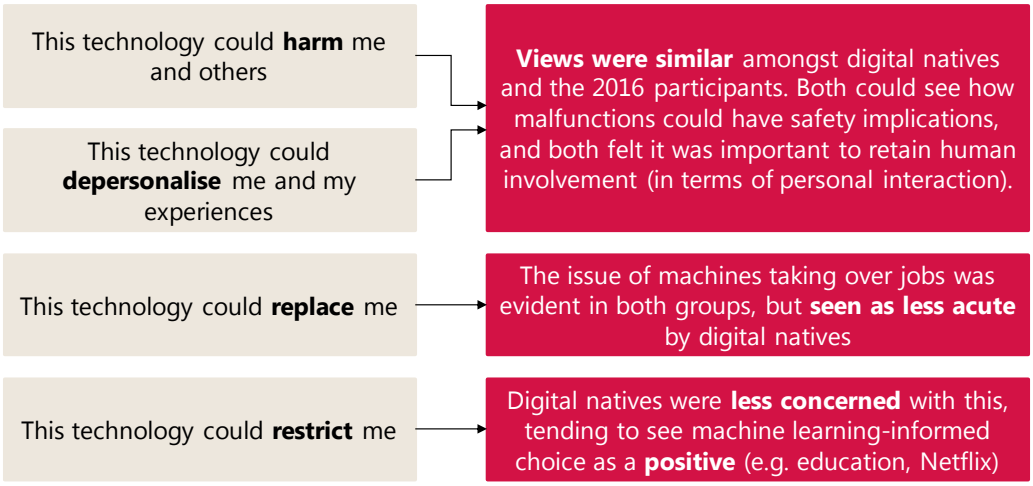
	2016 dialogues	Digital natives
Crime 	Participants tended to think that using machine learning to spot patterns in crime was a good idea in principle, but struggled to see how it might work accurately in practice. They saw it as a useful tool to aid with limited police resources, but were also concerned about the consequences of stereotyping individuals or groups.	Digital natives were generally supportive of this example, and even discussed other potential benefits and uses: to reduce human confirmation bias; linking CCTV to social media; and analysing intelligence related to terror attacks. They wanted to ensure that the technology would not be used in a way that would compromise the 'innocent until proven guilty' principle.
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.	Digital natives were very supportive of machine learning's use in the health sector, and whilst they felt that a human doctor would be needed to deliver serious or life-changing news, they identified several areas where they felt that human involvement would be less essential: symptom checking to help with triage; ordering minor tests for patients; and delivering diagnoses for minor ailments.
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.	Digital natives had nuanced discussions about a range of issues relating to driverless technology, including the need for all cars to be self-driving and the potential vulnerability to hacking. While safety concerns persisted, there was a level of acceptance that driverless cars would eventually become the norm – reflecting their initially more open approach to machine learning.
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.	Digital natives responded very positively to this idea, and there was greater appetite for machine learning to take more of a lead in this example than in the others; the need for humans to be involved in checking the algorithm was less acute than it was for other case studies. However, they still felt a human teacher was important: to inspire and motivate pupils.

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.	Digital natives felt that machine learning was best used to carry out background or administrative tasks, freeing up human carers' time to spend with those they looked after. They balanced resourcing issues with loss of human interaction, when assessing how acceptable this would be. Some participants felt that this would feel normal by the time they were older, and possibly in need of care.
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.	Digital natives focused on machine learning-produced film, music, or books (as most struggled to identify personally with poetry). The debate focused on whether the purpose of art was about reflecting the experiences of the artist, or audiences enjoying the art – with views on this fairly equally split. This was then reflected in whether or not they could see a role for machine learning.

1.5 The risks and value of machine learning

Digital natives and the 2016 participants identified similar benefits of machine learning. Participants felt that machine learning: had a lot of potential to benefit individuals and society; could save a lot of time; and could give people better choices.

Digital natives tended to approach discussions in a more open and less sceptical way, and as such, were generally more positive and accepting of machine learning's current and future potential uses than last year's groups. They identified many of the same concerns, but tended to feel these to varying extents:



Participants weighed up the risks and values that each of the case studies posed to society, and had broadly similar views for most of the examples.

However, their opinions differed on social care and education, with digital natives typically feeling that the potential risks that these examples posed to society were much less. Digital natives were generally more positive about the

prospect of using the technology in education – possibly linked to their familiarity with technology in supporting their own education.

1.6 Considering the development and application of machine learning

Digital natives found it hard to articulate clear, consistent views on how machine learning should be developed. In part, this was because of the breadth of different applications they could envisage, and their expectation that this technology would change everyday life in fundamental ways. Questions about how the technology should be developed fell into five groups, summarised below.

Similarly, the 2016 participants tended to focus on the risks and benefits of individual applications rather than more general conversations about governance. While not directly comparable, in conversations around *risks and benefits*, they developed a number of overlapping criteria to evaluate machine learning applications, which then determined how readily they could engage with them.

2016 dialogues

1. **What is the intention behind using the technology in a particular context?**
   Participants felt that the motives of those involved in developing an application might shape its success and direction as it progressed.

2. **Who would the beneficiaries be?**
   Views 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.

3. **How necessary is it to use machine learning?**
   Participants sometimes struggled to see why machine learning was necessary in some contexts, particularly where humans were seen as being as good as or better than machines at completing the task.

4. **How appropriate is it for machine learning to be used?**
   Participants were more concerned with appropriateness, particularly where machine learning would reduce valuable human-to-human contact.

5. **Will a machine make an autonomous decision?**
   If the example required a machine to make a decision, the importance of getting that decision right was a key factor in the assessment of all groups.

Digital natives

1. **What is the application?**
   The digital natives felt that the applications they had discussed were on a spectrum and that governance and oversight were more important the more 'serious' the context.

2. **Who is responsible for decisions and outputs?**
   The digital natives identified several ambiguities where machines and humans interacted and wanted clear guidelines to be in place and continually reviewed.

3. **Is there someone who can understand how the system works?**
   The digital natives felt they did not need to fully understand how an algorithm worked to trust the application, but wanted experts to ensure that the technology was working properly, particularly in applications with significant personal or social consequences.

4. **Who will guide machine learning's development?**
   The digital natives wanted a prominent role for independent experts, without an agenda, to shape the technology in a way that had broad benefits.

5. **How can we be confident in machine learning applications?**
   The digital natives generally felt that extensive testing would be essential for building trust amongst the public.

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 branch of artificial intelligence that allows computer systems to learn directly from examples, data, and experience. 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 aimed to increase awareness of the technology, demonstrate its potential, and highlight the opportunities and challenges machine learning presents. The project's focus was on the current and near-term (5-10 years) applications of machine learning, and the Royal Society published its report in April 2017[3].

[3] Royal Society (2017) *Machine learning: The power and promise of computers that learn by example*, available at: https://royalsociety.org/topics-policy/projects/machine-learning/

The UK public was a key audience for the Royal Society's project, and public engagement continues to be an integral part of its programme of work. At the end of 2015, the Royal Society commissioned Ipsos MORI to carry out research into public knowledge of, and attitudes towards, machine learning. This research focused on the wider public and the report was published in April 2017[4].

Building on this study, the Royal Society wished to engage with a specific audience, 'digital natives', to understand whether they had different views on machine learning.

2.2.1 Who are 'digital natives'?

The precise definition of 'digital natives' is not broadly agreed. The term was originally coined by Marc Prensky who used it to describe those who had grown up with digital technology (computers, videogames, mobile phones, the internet etc.) [5]. While it does not refer to a specific age group, people born in the 1980s and who grew up during the 1990s are generally considered the first potential digital natives. The defining characteristic of digital natives is their regular use of digital technology from an early age, resulting in both a **longer experience** of using technology, and typically **greater breadth of using technology** – which is ingrained into their lives to a greater extent than for non-digital natives[6].

Whilst not focusing on digital natives explicitly, previous research undertaken by Ipsos MORI has found that younger cohorts have different attitudes towards science and technology than older generations[7,8]. The Royal Society therefore wished to conduct supplementary research with this group to see how they currently differ from the general public overall, in terms of their attitudes to machine learning.

2.3 Objectives

This supplementary research was commissioned to explore the views of digital natives, and how they differed from those of the general public as a whole[9]. As with the previous machine learning project, the digital natives study focused on:

- Initial reactions to machine learning, including previous awareness of the technology and its applications;
- The perceived benefits and risks attached to the technology, explored through case studies of machine learning in practice;
- The potential value and risks of machine learning for society as a whole; and

[4] Ipsos MORI (2017) *Public views of Machine Learning: Findings from public research and engagement conducted on behalf of the Royal Society*, available at: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

[5] Prensky, M. (2001) 'Digital Natives, Digital Immigrants', *On the Horizon, MCB University Press*, Vol. 9 No. 5, October 2001, available at: http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf

[6] Helsper, E. and Enyon, R. (2009) 'Digital natives: Where is the evidence?' *British Educational Research Journal*, pp.1-18, available at: http://eprints.lse.ac.uk/27739/1/Digital_natives_%28LSERO%29.pdf

[7] Ipsos MORI (2014) *Public Attitudes to Science 2014*, available at: https://www.ipsos.com/sites/default/files/migrations/en-uk/files/Assets/Docs/Polls/pas-2014-main-report-accessible.pdf

[8] Ipsos Connect (2015) *Attention Generation Next! Beating the attention deficit for young audiences*, available at: http://m.ipsos.fr/sites/default/files/doc_associe/ipsos_connect_tp_gen_next_nov2015.pdf

[9] The challenges surrounding differences in views between the two groups are noted in Section 2.4.1

- How machine learning should be developed.

2.4 Methodology

The project consisted of two workshops to explore digital natives' views about machine learning, held in London and Sheffield in June 2017. The digital natives workshops were shorter than those held last year (a day, rather than a day and a half), due to the target audience's anticipated greater familiarity with machine learning-based applications.

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, according to our primary quotas:

- **Age**: 18-29 years old;
- **Experience of technology**: 'Strongly' or 'Tend to agree' with the statement, *'Using the internet was an important part of my life when I was growing up'*; and
- **Comfort using technology**: 'Very comfortable' in response to the question, *'People use new technology and media in many ways: sharing photos and posting on social media, using smartphone apps that track your location, and to access public and commercial services online (such as reading the news or doing your shopping). To what extent are you comfortable using new types of technology and accessing these services online?'*

Recruitment quotas were also set to ensure that, overall, people of a range of ages and from a variety of ethnic and socio-economic backgrounds took part[10].

2.4.1 Caveats around comparison

This report will provide a snapshot of how digital natives currently differ from the general public, overall, in terms of their attitudes to machine learning. However, it is important to note that being confident about the extent to which any differences in views are a result of life stage or cohort effects is challenging. Caution needs to be taken when comparing the findings from the digital natives workshops with the research carried out last year. This is due to:

- **The length of time between the two pieces of research**: Technology, and machine learning specifically, has developed during this time, and awareness of applications (and their perceived risks and benefits), may be higher with increased media attention (for example, in the field of driverless technology); and
- **The research design:** The digital natives study was commissioned to supplement the 2016 research with the general public. The digital natives research was done at a specific point in time, without anything to compare

[10] Please see appendix for a detailed sample breakdown

it to at a similar point in time with previous cohorts. We are unable to assess whether differences are down to age or cohort. It is also important to note that digital natives were also present in last year's groups.

2.4.2 A note on interpreting qualitative research findings

Qualitative approaches (including workshops) are used to explore the nuances and diversity of views, the factors that shape or underlie them, and the ideas and situations in which views can change. The results are intended to be illustrative, not statistically representative.

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 an important emergent view or trend. The value of qualitative work is to identify the issues that 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, either reflecting a 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 Initial reactions to machine learning

This chapter covers digital natives' understanding of machine learning and its applications, as discussed during the workshops. Broadly, digital natives were more familiar with machine learning-based applications than participants in last year's groups, and as a result required less time to get up to speed with how the technology worked. As they were familiar with some machine learning applications, they were also less likely to question whether the technology worked.

3.1 'Familiarity breeds favourability': Digital natives' initial openness

Digital natives were familiar with some applications that use machine learning, and spontaneously raised examples of where the technology was already being used in their everyday lives. These examples were predominantly consumer-focused, such as recommendation-based services like Netflix and Spotify. Most were examples that digital natives used regularly and that were embedded into their lives. Being familiar with how machine learning worked in practice helped the digital natives to grasp the concept of machine learning. They were able to quickly develop a broad understanding of the principles underpinning the technology and its applications.

"Spotify do a* [Discover Weekly] *playlist, where they make a playlist made up of music that's similar to what you already listen to, and I really like that."
Sheffield

As a result of their experience with machine learning-based applications, digital natives were able to suggest a wider array of potential uses for machine learning, before further examples of how the technology works in practice were discussed. For example, one participant in London suggested that an algorithm might be able to analyse students' grades, personality traits, and other data to suggest possible career options. This participant thought that an algorithm might be able to produce a more tailored suggestion than a human careers advisor whom they felt might not always understand individual pupils' preferences. A second participant made the following health-based suggestion:

"You only get 15 minutes with your GP, so sometimes people do use online sites to see if they can understand what they have before they go. If they were to set up something that had all of these thousands of people putting in their symptoms and what they had and things… They'd have a lot more data to look at."
Sheffield

Both the 2016 participants and the digital natives could see how wide-ranging the potential applications of machine learning were, and the ways in which it was already being used. Consequently, the idea that the development of machine learning was inevitable was evident in both groups. However, the digital natives could quickly see and *more readily accept* that machine learning technology was already part of their lives, compared to the 2016 participants. As such, they were less concerned by what they saw as the inevitable development of machine learning, compared to the 2016 participants who were more resistant to the idea that the technology worked in the first place.

Digital natives' familiarity with machine learning-based applications was evident in how they described their use of technology: in habitual, or routine terms. For example, they discussed the trade-off between providing data and

receiving a service, such as setting up a social media account, or doing online shopping. While they were not *positive* about giving up their personal data in this exchange, they were not *negative* about it, either. Instead, they accepted such exchanges as 'normal' much more readily than participants in last year's groups – where the most sceptical participants would avoid certain situations where they were required to input their personal data in order to access a service.

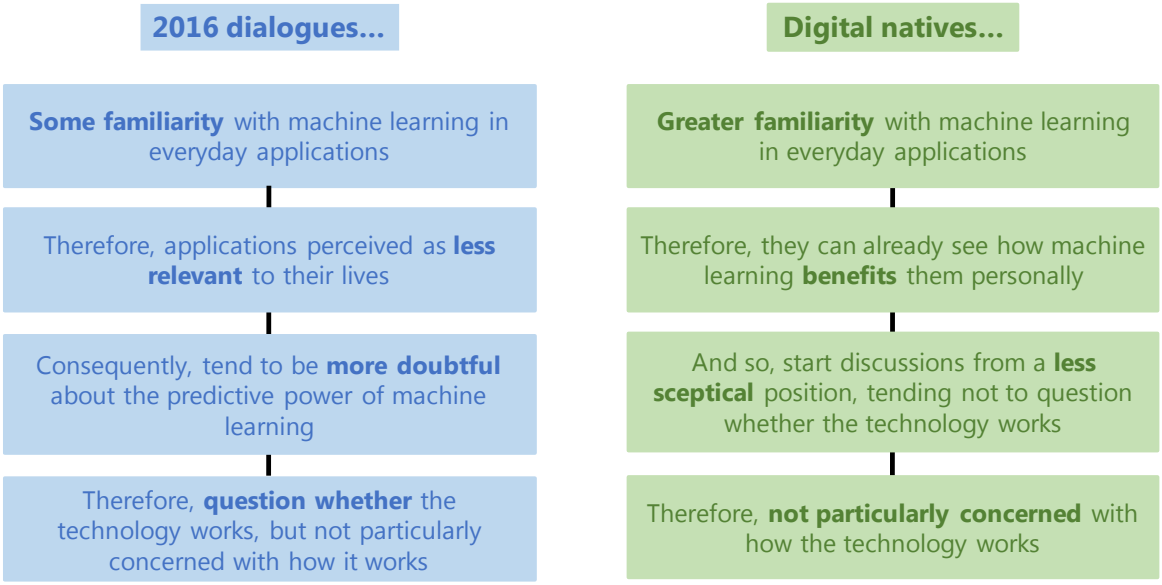
"If you put something on social media, it's not really yours anymore. You don't own the picture, it's out there on the platform – anyone can copy it, download it."
London

Last year, participants displayed four spontaneous reactions to machine learning:

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 risk 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'

These spontaneous reactions were also seen in the digital natives, but the key differences were that they tended to have a more positive approach to the technology, and there was little evidence that they were suspicious about the technology's purpose.

Figure 3.1: Key differences in approach to discussions on machine learning



Digital natives, on the whole, more readily accepted that machine learning worked and therefore took a more open approach to discussions than those in the 2016 dialogues. This is because the digital natives were more familiar with everyday machine learning applications – they tended to have both *more* experience and *positive* experience of using the technology.

Neither the digital natives nor the 2016 participants were particularly concerned about the intricacies of *how* machine learning worked in practice. For both groups, this was in part due to the complexity of the technology being something they assumed they would not be able to understand. However, as discussed above, familiarity with machine learning applications was also important in shaping reactions to the perceived inevitability of the technology, and helps explain the differences in attitudes between digital natives and the 2016 participants.

4 Case studies

Digital natives were asked to consider six different case studies[11] and discuss the potential risks and benefits of each. 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 takes each case study in turn, describing digital natives' reactions, and then summarises the findings from the 2016 groups at the end of each section.

4.1 Crime and policing

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.

Digital natives could see the potential benefits of using machine learning in a crime context: to help the police gain an advantage over criminals, and help to manage resources more effectively – with some participants considering cuts to police budgets when framing their discussions.

"They could use it to know what's coming, though, and prevent it. So many police cars sit on the road monitoring people; if we have more machines monitoring then it frees up manpower to deal with things."
Sheffield

They also pointed to the potential for machine learning to remove human confirmation bias and felt that an algorithm, subject to being given the right data to analyse, had the potential to reduce bias through analysing large datasets from many different sources. Those who raised this idea, or supported it, felt that an algorithm could reduce the risk of human police officers (subconsciously) gathering evidence to support their own hypotheses, or preconceptions about who might have committed a crime.

Despite seeing the potential benefits of this technology, digital natives clearly identified a 'red line' for the use of this technology, stressing that it would be unacceptable to use it to inform pre-emptive measures against individuals. Participants were clear that moving to arrest potential perpetrators, without sufficient evidence beyond that provided by the algorithm, would contradict the idea of 'innocent until proven guilty'. This would therefore be unacceptable from a legal and ethical perspective. This view was shared by all, including those who were open to the idea that machine learning algorithms would always be improving their predictions, and may eventually be capable of predicting at least some future crime with good accuracy.

[11] Two additional examples were covered in the 2016 groups that were not used with digital natives (due to time restrictions). These were: **marketing** (tailored marketing based on previous behaviour, and drawing on data from other people who have behaved in similar ways) and **finance** (banks monitoring spending patterns to detect fraudulent activity, and an automated service warning against purchases when people had insufficient funds, or where the algorithm had identified times where individuals were historically prone to overspend).

"If you're going to start to predict who is going to commit what crime, I don't think that's fair. The algorithm is effectively saying that person doesn't have free will [...] you could have that prediction and it becomes a self-fulfilling prophecy."
London

Whilst initial reactions to this example were quite positive, as discussions progressed, some participants were more sceptical about how effectively the technology would work in practice. Certain participants felt that algorithms would not be able to predict more opportunistic crimes, such as muggings or break-ins, due to their 'one off' or spontaneous nature. Other participants felt that the algorithm may become a victim of its own success. For example, one consequence might be that crime would simply move to another area, or that criminals would catch on to the algorithm and adapt their behaviours in an attempt to outwit the algorithm's prediction.

Furthermore, a few digital natives raised the idea that police presumably already know where crime hotspots are. Some participants therefore questioned what more machine learning could add in this context, and argued that limited police resources were a stronger factor in determining police success in combatting crime. Reflecting their open-minded approach to the case studies, digital natives built on this idea to speculate about possible contexts where machine learning could add more value:

- CCTV facial recognition to be linked to social media sites (such as Facebook) to enable algorithms to draw on more data to identify perpetrators
- Algorithms should focus on more planned, coordinated incidents, such as analysing intelligence on terrorist attacks and helping to predict likelihood

2016 dialogues

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.

Participants identified the same primary benefit and ultimate 'red line': the potential for machine learning to help manage police resources effectively, and the need to protect suspects' rights with regard to being 'innocent until proven guilty'. However, last year's participants tended to be more sceptical about the use of machine learning in a crime setting, in particular being mistrustful of the integrity of the predictions. These participants doubted whether the technology could work in practice and were concerned about individuals being labelled as a result of an algorithm's prediction. Furthermore, they also felt that using historic data could reinforce stereotypes, and justify the targeting of certain individuals or groups within society.

"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, 2016 workshop

In addition, 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 increased police monitoring, which they saw as an 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, 2016 workshop

4.2 Health

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

Digital natives were very supportive of machine learning's use in the health sector and could intuitively see the potential benefits to individuals and society. They felt that machine learning could improve accuracy and permit more variables to be considered when assessing physical health conditions than was currently possible with human doctors. Digital natives were given the following example of how machine learning could be used to improve breast cancer prognoses:

Machine learning in action: Breast cancer prognosis

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[12].

Participants discussed what role human doctors might have in these scenarios, agreeing that there would still be a need for them to act as a second pair of eyes in terms of the data analysis, and to retain personal interaction with the patient. They felt that the latter would be particularly important when delivering more serious news. However, digital natives discussed situations where they felt that human involvement could be reduced, again demonstrating their more open-minded approach to the case studies. These examples included:

- Symptom checkers that could help with triage and waiting times;
- Ordering minor tests for patients; and

[12] Myers, A. (2011) '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

- Delivering treatment news or diagnosis for minor ailments.

"It all depends on the topic. If you've got a fractured ankle, whatever, you wouldn't mind it saying, 'you have got a fractured ankle and will need this treatment'. I don't mind that – I'm not going to die."
London

By contrast, using machine learning to predict whether someone may be suffering from a *mental health* condition was one of two areas across all the case studies where digital natives were more sceptical that the technology could or would work (the other being childcare). Some also raised concerns about the possible negative effects on vulnerable or isolated people from only having contact with a machine, as opposed to a human doctor.

There was debate around this issue, though, with counterarguments focusing on the fact that the algorithm could be analysing cases where people had already been referred by a human doctor, and that this would be a tool to aid diagnosis. Those who were more supportive of this idea explained that, in a sense, it wouldn't be the machine diagnosing – at least not taking sole responsibility. This counterargument was used across the case studies by the digital natives more frequently than was the case in the 2016 dialogues – demonstrating the digital natives' more intuitive understanding of machine learning.

One group spontaneously discussed whether an algorithm could analyse people's Facebook statuses to detect changes in their mood, and possibly identify if people were suffering from depression. Whilst they were sceptical about the idea of a machine diagnosing people, they did have an understanding of how this could work in practice.

"With machine learning, there are millions of statuses over time, so it won't just look at someone who's just typed 'sad day', it will look at patterns, aggregate data. It would use habits its picked up, parameters within the algorithm to decide if someone might or may be at risk."
London

2016 dialogues

The use of machine learning in a health context was where participants could intuitively see the greatest potential for benefits to individuals and society. The specific example of machine learning being used to improve breast cancer prognosis was crucial for many to accept that machine learning could actually work in practice, as they could see empirical proof of algorithms analysing more variables than a human (and recognising previously unseen patterns).

Participants felt that machine learning could improve accuracy, as it would be able to consider more data when making diagnoses than humans.

Despite their support for machine learning to be used in healthcare, participants wanted human involvement in diagnosis and treatment to remain, to ensure that the 'personal touch' and reassurance of human oversight were not lost.

"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, 2016 workshop

Participants were, overall, less supportive of machine learning's use in diagnosing mental health conditions. They typically struggled to accept that there would be physical manifestations present in a consistent enough way for a machine to analyse, and they felt that a machine would not be able to take context into account like a human doctor might. They also pointed to other weaknesses such as a machine's inability to understand an accent, or rely on other senses.

Participants were concerned that misdiagnoses would lead to people being labelled (possibly reflecting participants' attitudes towards mental health, rather than machine learning per se), and that machines would replace humans, leading to loss of personal interaction.

4.3 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.

Digital natives clearly identified both risks and benefits associated with driverless cars. They identified safety as the primary risk, and wanted extensive testing to be done before driverless cars became more commonplace on the roads. However, they saw this technology benefitting those who could not drive, by affording them greater independence, and felt that it could be more efficient than human drivers.

"You wouldn't have traffic lights at a junction, cars could just slip in with each other. The time to travel from A to B would be less. For everyone it would be more efficient, even if it'll be less for some individuals."
London

Digital natives' discussions of self-driving cars became advanced quite quickly. However, it is worth noting that there has been much on this topic in the media in the time between last year's and this year's workshops – as such, the 2017 participants' awareness of driverless technology and the associated issues is likely to be higher. Examples of how these groups progressed the discussions included:

- **The need for all cars to be self-driving:** Some digital natives identified the potential risks of having a mixture of human and non-human drivers on the roads, early on in conversations. They felt that this mix would be unsafe, as human drivers would be unpredictable whereas machine learning-based cars would be programmed in the same way
- **More advanced discussions around responsibility:** Some digital natives wanted to know what safeguards would be put in place to ensure that the technology was safe. For example, they discussed controls needed to prevent children from being able to use them unattended, or adults under the influence of alcohol being able to take control of the car (where the car was not fully autonomous)
- **The potential vulnerability of the technology to hacking:** Some digital natives raised the idea that driverless cars, reliant on sensors and shared networks to communicate with other vehicles, could be targeted

by hackers. Digital natives discussed the possible consequences of this, feeling that it posed a significant risk to people's safety[13]

"If you have everyone in these pods, it becomes infrastructure. If it's controlled by a central computer, that can get shut down. All the cars shut and lock all over the country. There has to be some security."
London

There was a level of acceptance amongst digital natives that driverless cars would eventually become the norm. They shared concerns over safety with last year's participants, but seemed to trust that the technology would get to a point where it would be safe to use and were more assured of this than participants last year. As such, driverless cars were less of a conceptual leap for the digital natives.

2016 dialogues

Participants' first thoughts were about how driverless cars might affect them personally, with those who enjoyed driving being concerned that it could reduce their freedom to carry out an activity they took pleasure in, and those who were unable to drive (due to ill health, financial difficulties, or who had never learned) feeling it could be liberating.

They also identified efficiency as a potential benefit. The more technologically engaged participants recognised that cars could be programmed to drive in the same way, and that this would ensure traffic could move in a more uniform and controlled manner, increasing efficiency on the roads.

Safety was identified as the main risk, and participants wanted driverless cars to be tested extensively in a range of scenarios, before they were released to the public. There was an expectation that driverless vehicles would have to be much safer than human drivers, with many even feeling that there would have to be an assurance that driverless vehicles would not cause accidents before they could be used.

4.4 Education

Participants discussed the potential for 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, including the role for machine learning in **marking students' work and tests**[14].

Digital natives reacted very positively to this case study. They were receptive to the idea of identifying students' learning styles early on, and tailoring their learning according to their strengths. They did not raise concerns about pigeon-holing pupils into skillsets and careers, or losing the ability to develop certain skills. Rather, they saw this as

[13] It is worth noting that a global cyber-attack, which also affected the NHS in the UK, took place roughly a month before the workshops. This may have contributed to participants' greater awareness of the issue of cybersecurity and hacking.

- BBC (2017) NHS cyber-attack: GPs and hospitals hit by ransomware, 13 May 2017, available at: http://www.bbc.co.uk/news/health-39899646, accessed 20.7.17

[14] This example was not discussed in the 2016 workshops

an opportunity to help young people, by identifying their strengths – something they felt that human teachers may not always be able to spot.

"I think that's really good because teachers don't always have the time to analyse all students. Some people might be at a disadvantage, as people aren't teaching the way they need teaching. This would be on an individual basis."
London

Digital natives had grown up with technology, and typically already had experience using it to support their learning, either at school, college or university. The idea of using machine learning in education was therefore much more tangible. There was greater appetite for machine learning to take more of a lead in this context than the other case studies. Participants were comfortable with the idea of an algorithm identifying learning pathways for students, and for teachers to use this information to tailor their style (as much as possible) for individual students.

"When you're still under the health visitor, when they check your child's hearing and give the injections, you could be tested then to see what type of a learner you are. Then tested later again, too, to see if you are still a visual learner."
London

The need for humans to be involved in checking the results of the algorithm was less acute than it was for other case studies. However, digital natives still stressed the importance of maintaining a human teacher, who had an important role to play in inspiring pupils, and motivating them to continue working.

2016 dialogues

Spontaneous reactions to this case study were positive, with participants warming to the idea of being taught as an individual – something they felt was not possible in a large, classroom setting. However, there were some concerns that tailoring in education might be taken too far, and could result in young people losing core skills, or restricting their horizons.

"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, 2016 workshop

Participants felt that machine learning should be a tool used by teachers, rather than an alternative way of educating people, as they recognised the importance of teachers as role models and communicators. Other participants recognised the strain on resources in teaching and were supportive of machine learning playing a supporting role, to free up teachers' time to spend with pupils. Ultimately, participants felt that this technology was more appropriate in the context of adult education.

4.5 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**.

Digital natives debated the pros and cons, and appropriateness, of machine learning's use in a social care setting. While they did have concerns, their positive view of the potential of machine learning meant they were open to making the most of technology in this context. For digital natives, machine learning's use in social care was, again, not too much of a conceptual leap.

"It's just an extension of a chair lift, really, isn't it?"
London

Concerns – as with the 2016 dialogues – centred on the risks of reduced human involvement. Participants were uncomfortable with the idea that social care recipients would have a depersonalised service, as they felt that human contact was very important. Indeed, they stressed the importance of 'care' in social care, and were not convinced that a machine could provide this.

In addition, where participants felt that machines could play a role (performing 'background' tasks with less scope for error or harm, such as cooking and cleaning), they still wanted human social care providers to be present, to check that the machine was working as expected. Participants suggested that humans could check machines were delivering the correct medicine, for example. Digital natives saw this as more of a 'spot check' rather than involvement at every point in the care.

"It's like autopilot on planes. It's fantastic, but you still want the pilot there."
Sheffield

Participants speculated about what technology might be in use when they themselves were more likely to be in need of social care services. Digital natives seemed to feel that machine learning would become 'the new normal', as it was used in a greater number of settings, and as people adapted to it over time.

"Not to generalise, but most old people don't like technology … By the time this is available, though, we'll be the old people. I'd love that! With the ageing population, it's also inevitable."
Sheffield

2016 dialogues

Participants felt that social care should be about an emotional relationship and human interaction, and consequently many participants were against the idea of machine learning being used to provide social care. Others also argued that it was undignified to consider a robot helping an older or disabled person with things such as bathing or going to the toilet.

"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, 2016 workshop

However, some participants pointed to the perceived deficiencies in the quality of care received by many older people, and 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 pursue it. Overall, there was consensus that machine learning could play a supporting role in social care, as 'an aid, not a replacement', taking on background tasks to free up social care providers' time to provide genuine support and meaningful interaction with patients.

As with the 2016 participants, digital natives were more sceptical about machines being used for childcare, and this was one of the areas where they were less convinced that it could work (the other being diagnosing mental health conditions). As well as doubting whether it *could* work, as with the mental health example, digital natives were unsure as to whether it *should* be permitted in this setting. They raised several concerns over the use of machine learning in childcare, all of which were also evident in last year's dialogues:

- **The unpredictability of children**: Digital natives doubted that an algorithm could predict the random behaviour of children, or that a machine would be able to react quickly enough to new situations. They were therefore concerned that children's safety would be compromised if they were cared for by a machine
- **Abuse of the technology by parents**: Some felt that there was scope for parents overusing the technology, where participants felt they should be primarily responsible for childcare
- **Development and socialisation**: Related to the point above, digital natives were also concerned that over-exposure to a machine from a young age could harm young children's development, particularly if this came at the price of reduced exposure to other humans
- **'An aid, not a replacement'**: Again, digital natives were more supportive of machine learning being used to support parents or guardians. For example, they suggested that the machine could take care of routine household tasks, freeing up parents to spend time with their children

4.6 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.

Many digital natives struggled to personally identify with the poetry example, so conversations tended to move to discuss machine learning being used to write books, films, or songs – artistic forms they were more familiar with. A

widely-identified objection was that the machine did not know the meaning of the words it was using, which participants felt reduced the value of this technology (including amongst those who were more supportive of the idea).

"It takes the soul out of it. If it doesn't know the meaning of the words it's using, it's not really art, is it? If you're reading something like a poem, you can feel the passion they've put into it."
Sheffield

Discussions centred on a 'philosophical' debate around whether the crux of art is the artist's emotions and experiences, or the reader's interpretation and personal enjoyment. Those who were uncomfortable with this example felt that artistic expression was a specifically *human* activity. They felt that the essential purpose of art was self-expression and story-telling – an artist sharing their experiences and emotions, and the audience connecting with this story.

"It would take something away from it. For something visual, would it really be art if it were just made from an unknown source, created for nothing? Isn't half of it the artist who's creating it?"
Sheffield

Those who were artists themselves, or who had creative experience, were particularly uncomfortable with this example, disliking the idea of a machine trying to recreate what, for them, was a very personal and emotional practice.

Those who were more positive focused on the individual interpreting the art, and how personal enjoyment was the most important factor in determining value. For these participants, it mattered less that the art had been produced by a machine, and that the experiences or emotions presented had no underlying human meaning (in terms of creative intent). In fact, some of those who were more engaged with this example pointed to the fact that the experiences and emotions that the machine would be drawing on *were* real – as the algorithm would produce poetry based on analysing examples of human experiences.

"If you got a feeling from* [the machine learning poem]*, is it a crime that a computer generated it? What if a computer generated a novel from an author you really liked, but who had died and you couldn't get any more books from? With a book or music, if I just enjoyed it, maybe I wouldn't think about its provenance."
London

The idea of a machine being able to create new, original work from an author, poet or singer who had died was considered by participants. Those in favour liked the idea that their favourite artists could, in a sense, be preserved. Meanwhile, those who were less positive again felt that this would take something away from the original work, which in their view would seem 'inhuman'.

Examples of these two contrasting views are shown in Figure 4.1.

Figure 4.1: Reflection versus reaction

Reflection

*Say Whitney Houston, who's dead but influential, what if all her songs were put into a generator – **would people like that?** I think it would be interesting, but **not the same**. It's that person's words recycled. **It's like a remix.**"*

London

Reaction

*I like one author, Stuart MacBride, and have read all of his books. **Could a computer follow his same style?** If he died and I couldn't get any more of his books, but a computer could generate a novel **inspired** by him, then **I'd definitely read it.**"*

London

2016 dialogues

Most of the participants believed that the machine learning poem had been written by a human, because of the language used, and the case study produced two competing views:

- **'Reflection':** These participants 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 therefore never produce true 'art'
- **'Reaction':** These participants cared more about the effect the poem had on them, not how it had been written; the machine-written poem gave them more as a reader than the human poem. Their preference was for a poem they could relate to, regardless of whether it was written by human or machine

Participants of both mindsets thought that machine learning poetry was not particularly risky to society, and that it had very low social value. Much of this was due to the fact that humans can write poetry already – in contrast to some of the other case studies which emphasised machine learning's superiority over humans in terms of data analysis.

"The examples you initially gave were about things that…* [aren't] *feasible for us as humans. We're now talking about something that we can do and we've been doing for* [years] *… This is just taking away the last few things we've got. I don't see why it's important."
Oxford, 2016 workshop

5 The risks and value of machine learning

This chapter explores digital natives' take on the risks and benefits associated with machine learning, in different contexts. Participants were asked to consider the risks and benefits of individual case studies, and also to discuss the social value and social risk of each. This chapter sets out the overall risks and benefits identified by digital natives, and then presents how they defined 'risk' and 'value' when discussing the case studies, relative to each other.

5.1 Key themes: benefits and concerns relating to machine learning

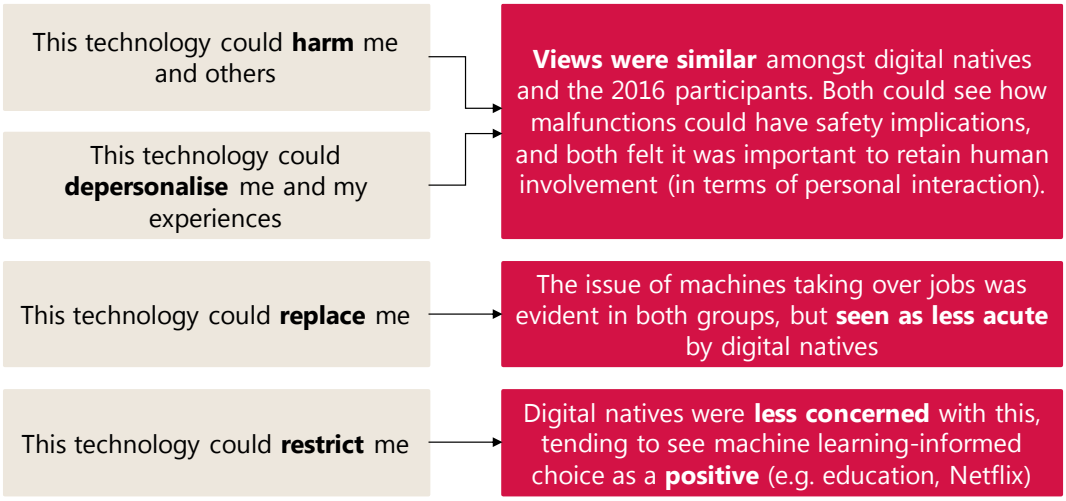
The perceived benefits and concerns that participants had about machine learning were similar across the two studies. Participants felt that:

- Machine learning had a lot of potential to benefit individuals and society;
- Machine learning could save a lot of time; and
- Machine learning could give people better choices.

The main difference was that digital natives tended to approach discussions in a more open and less sceptical way and, as such, were generally more positive and accepting of machine learning's current and future potential uses. As discussed in Chapter 3, digital natives were also able to come up with more examples of what machine learning's potential future uses may look like in practice.

The digital native groups expressed concerns that were similar in scope and nature to the 2016 dialogue groups. However, the strength with which the groups expressed these concerns varied, as shown below in Figure 5.1.

Figure 5.1: Concerns identified by the 2016 participants, and views from digital natives

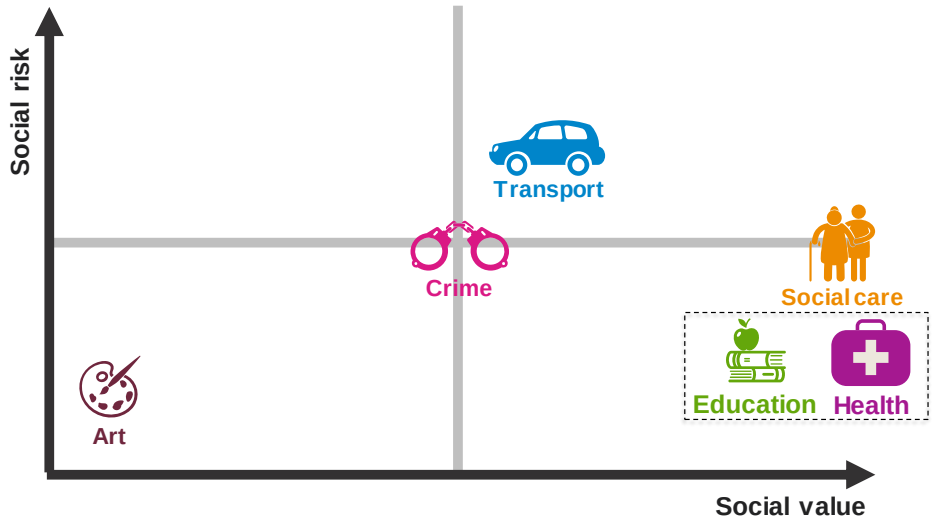


5.2 Considering social value and social risk

Digital natives were asked to consider the value and risk that each of the applications posed to society. Across the discussions, there were felt to be risks associated with machine learning generally (as discussed above), as well as in relation to specific applications. As with last year, the actual process of machine learning – the computation and 'data crunching' – was not seen as being particularly problematic, although there were concerns about how the technology would be applied in practice, and the consequences of any problems or errors.

Digital natives were 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 5.2 below shows where the different case studies were typically placed by participants.

Figure 5.2: Digital natives' social risk versus social value assessment



15

Digital natives weighed up the risks and values of the case studies relative to each other, and revised their earlier opinions in relation to discussions on new case studies. In determining an example's risk or value, the main question that digital natives seemed to consider was, 'would this be better than what we currently have?' – which tended to determine the overall acceptability of individual applications. Participants in the 2016 workshops adopted a similar approach.

The education and health case studies occupied the same space on the quadrant – both were felt to have the highest societal value, and relatively low societal risk. All groups[16] had different perspectives on the crime example, hence an 'average' has been taken, and it is shown in the middle of the chart. Discussions on value varied substantially between the groups:

- Those who focused more on the idea that the technology did not add anything new (as police already know where crime hotspots are likely to be) felt that it had low social value

[15] Discussions around societal risk and value, and the individual quadrants from across the groups were reviewed in order to produce this summary chart.

[16] Each workshop was split into two smaller groups of 10-12 participants

- Those who could see machine learning being used to analyse large volumes of intelligence data felt it had high social value, as it could perform this task more quickly and accurately than human intelligence officers

There was more consensus on the risks, with digital natives identifying this as the technology being used to target individuals, and arrest them without sufficient evidence. However, if the technology was purely being used to put preventative measures in place, then the digital natives felt that there was a low risk to society. Importantly, the digital natives judged these two aspects as comparable to the current level of risk associated with human police officers performing the same task.

5.2.2 Determining risk to society

The approach to assessing an application's risk to society was similar across the 2016 wider public and 2017 digital natives' workshops. Both began by thinking about how the scenarios might affect them (or people they knew) as individuals, and then scaled these conversations up to determine the impact on society.

When considering risk, participants were thinking about the possible ways that the technology could go wrong, and the impact that this could have – in terms of harm, safety, and what society might lose as a result of giving machines a greater role. For example, participants could easily identify the impact of driverless technology going wrong, which would result in physical harm. They also felt that transport would have one of the biggest potential impacts, as it would be one of the few case studies to affect everyone – the digital natives worked on the assumption that everyone would have to have a driverless car, to reduce the number of crashes. The possible risk to society was therefore ranked highly. Conversely, art was felt to pose a low risk to society, as if it 'went wrong' it would not result in physical harm (though there was some debate around the risk to society of losing a shared culture, or ability to express ourselves).

The digital natives also considered the role of humans when judging the examples and the level of risk they posed to society. They thought about humans overseeing the algorithm, and also whether personal interaction would suffer as a result of machines taking a more prominent role. For example, education was felt to be low risk as human teachers would still have a role in delivery, and interaction with students. Likewise, social care was seen as low risk – on the assumption that human carers would be present to spend time with patients, address their concerns in person, and also check on tasks performed by the machine.

Digital natives made several references to fictional portrayals of machine learning-based technologies, when they discussed risk[17]. They used examples from books, film, and television where robots or artificial intelligence had malfunctioned, as their framework for the possible risks that machine learning posed to society.

"You watch so many films and stuff, it's hard not to think of the risks. In all the films, the risks are these things going wrong."
London

"How long before technology does everything? It does so much already, how long is it before it takes over? It's like that film, Wall-E, where we're all in chairs..."
Sheffield

[17] In the quantitative survey from the 2016 project, 21% of people had heard about machine learning from entertainment (for example, books, films, and video games – including science fiction)

These examples indicate that the depiction of machine learning and similar technologies in fiction is influential in shaping participants' perceptions of risk.

5.2.3 Determining value to society

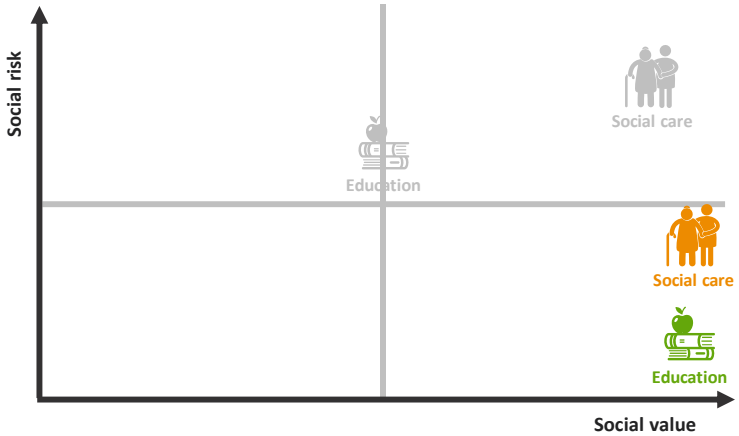
Digital natives found social value easier to conceptualise. They began by thinking about how different types of individuals might be affected. As well as considering what more machine learning could add to society, what pressures it could relieve, and what improvements it could enable, participants from both sets of research thought about the importance of the technology – whether there was a *need* for it in the first place.

Digital natives judged value according to several key concepts:

- **Whether humans could perform the same task:** If the digital natives could see that machine learning was better suited to a task, because it could analyse a greater volume of data, or produce more accurate results, then they judged it to have a higher social value. For example, they could see that machine learning could outperform humans in the fields of education and social care, but did not believe this to be the case for art (which was subsequently deemed to have the lowest social value)
- **The relative importance of the benefit(s) brought by the application:** Digital natives revised their assessments of social value and risk, based on case studies they had already discussed and, indirectly, produced a hierarchy of benefits to society. For example, transport was not felt to have as much social value as the health example, because the benefits of driverless cars were primarily seen as efficiency and convenience – not thought to be as important as saving lives or improving prognosis
- **Whether they felt it could work in practice:** Digital natives tended to be more positive about the examples that they believed could work, and the fact that there was empirical proof of machine learning contributing to improvements in breast cancer prognosis was a key factor in their perception of its high social value

5.2.4 Comparing risk and value: Digital natives versus the 2016 participants

Figure 5.3: Risk and value: social care and education[18]



The digital natives and the 2016 participants had broadly similar views on the risks and values of most of the case studies. However, their opinions differed on social care and education. The digital natives tended to feel that the potential risks to society of machine learning in these areas were much less, and were notably more positive about the prospect of using the technology in education. This could be because of their familiarity using technology in their own education, and because of their relative distance from needing social care services.

[18] In Figure 5.3, the grey symbols represent the 2016 groups and the coloured symbols represent the 2017 digital natives

6 Development of machine learning

This chapter explores digital natives' perceptions of how machine learning should work in practice, particularly focusing on the criteria they felt should guide how the technology would be developed.

6.1 Considering the development and application of machine learning: Context is key

Digital natives found it hard to articulate clear, consistent views on how machine learning should be developed. In part, this was because of the breadth of different applications they could envisage, and their expectation that this technology would change everyday life in fundamental ways.

Digital natives felt that the machine learning applications they had discussed throughout the day were on a spectrum of seriousness, in terms of social impact and consequences of malfunction.

Digital natives felt that governance, transparency and understanding were more important the more 'serious' the context, or the less familiar the application. For example, where there was potential for physical harm, as with driverless technology, or psychological harm, as with machine learning robots being used to diagnose and treat vulnerable individuals, they felt that it was more important to have some sort of oversight in place.

"It depends on what it is. Recommendations, Facebook … It's not the end of the world if you get the wrong thing. A mortgage, a job – that's your life."
London

Similarly, there were some contexts where digital natives considered governance not to be important, because the consequences of something going wrong were perceived as less severe.

"It's when it's used for something more sinister like government spying and listening to your phone calls. The light-hearted stuff is irrelevant; it doesn't matter."
Sheffield

The idea of 'context is key' guided digital natives' ideas about oversight, and their key concerns about how the development of specific machine learning applications would be taken forwards. Key questions that participants felt were important to address in 'high risk' areas fell into the following categories:

- Who is responsible for decisions and outputs?
- Is there someone who can understand how the system works?
- Who will guide machine learning's development?
- How can we be confident in machine learning applications?

6.1.1 Who is responsible for decisions and outputs?

As digital natives discussed the machine learning examples and considered how they might work practically in society, they identified several ambiguities where machines and humans interacted. For example:

- Who owns a song, film, or a poem written by an algorithm, based on someone else's work? What if it received commercial success – who would reap the benefit?
- Who is ultimately responsible for a crash involving a driverless car?
- Who is responsible in situations where an algorithm's prediction is found to be wrong, and who should be blamed? In a medical scenario, who would be to blame if an algorithm predicted a high likelihood of a benign tumour that was later discovered to be malignant? What if this was not discovered in time and the tumour inoperable?

The digital natives recognised that such scenarios were challenging, but thought they were important to resolve. They did not have any clear answers to these questions themselves, but felt that independent experts and industry had a role in answering them.

6.1.2 Is there someone who can understand how the system works?

As discussed in Chapter 3.1, digital natives did not feel that they needed to be able to fully understand how the algorithms would work themselves. However, they felt strongly that *someone*, or a body of people, would need to have this knowledge in contexts where the consequences of decisions or actions were perceived as more severe. Digital natives wanted experts (independent from developers) to verify the technology, and scrutinise its use to ensure that the algorithms were working correctly and safely.

6.1.3 Who will guide machine learning's development?

Related to the point above, digital natives wanted independent experts to have a prominent role in the development of machine learning. Overall, their preference was for the technology to be developed by those perceived not to have an agenda – be it political or profit-driven. Digital natives felt that this was important to ensure machine learning would not be used to the detriment of society, particularly to the detriment of vulnerable groups.

6.1.4 How can we be confident in machine learning applications?

Digital natives felt that extensive testing of machine learning and associated applications would be essential to build trust and support amongst the public. They felt that this was important prior to a product or service's release to the public, as well as being ongoing – continual review by independent experts to ensure that the technology functions as it should.

2016 dialogues

Participants recognised the importance of regulation, but found it challenging to come to a general view having discussed the risks and benefits as anchored around particular case studies. They could see how governance may be framed differently in different areas.

They found these *separate* conversations around the risks and benefits of individual case studies more straightforward, and used a number of overlapping criteria to evaluate machine learning applications, and to determine how readily they could engage with them:

1. **What is the intention behind using the technology in a particular context?** Participants felt that the motives of those involved with an application's development might shape the success, and direction, of the technology as it progressed. Consequently, participants generally wanted to know *who* would be involved with the development and delivery of the technology.

2. **Who would the beneficiaries be?** Views were more positive when they thought there would be worthwhile benefits for individuals, groups of people, or society as a whole, than with examples that may, initially, only be available to further private interest. Applications were considered less worthwhile where they were primarily profit-oriented.

3. **How necessary is it to use machine learning?** Where humans were seen as being as good as or better than a machine at completing tasks, some participants struggled to see why machine learning was necessary. The clearest example of this was creating art.

4. **How appropriate is it for machine learning to be used?** Many of these concerns centred around the loss of human-to-human contact, with examples including robots in the home, and an increasing role for machines in education, facilitating personalised learning.

5. **Will a machine make an autonomous decision?** If the example involved a machine making a decision, the importance of getting that decision right was a key factor in the public's assessment.

Appendix

Qualitative sample breakdown

The following table shows the sample breakdown of the focus groups. Please note that qualitative research does not aim to be representative; a qualitative sample should broadly reflect the population.

		London	Sheffield
Gender	Male	11	10
	Female	12	11
Social grade	AB	4	3
	C1	10	9
	C2	6	6
	DE	3	3
Ethnicity	Asian – Indian	2	-
	Asian – any other Asian background	-	1
	Black – British	1	-
	Black – Caribbean	1	2
	Chinese	-	1
	Chinese – any other background	-	2
	Mixed – any other mixed background	2	1
	White – British	17	14
Working status	Not working	1	2
	Student	8	1
	Student and working (part-time)	-	6
	Working (full-time)	2	10
	Working (part-time)	12	2

Daniel Cameron
Research Director
daniel.cameron@ipsos.com

Kelly Maguire
Senior 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.