

Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance

Big Data & Society
January–June: 1–12
© The Author(s) 2020
DOI: 10.1177/2053951720915939
journals.sagepub.com/home/bds


Aphra Kerr¹ , Marguerite Barry²  and John D Kelleher³ 

Abstract

This article draws on the sociology of expectations to examine the construction of expectations of ‘ethical AI’ and considers the implications of these expectations for communication governance. We first analyse a range of public documents to identify the key actors, mechanisms and issues which structure societal expectations around artificial intelligence (AI) and an emerging discourse on ethics. We then explore expectations of AI and ethics through a survey of members of the public. Finally, we discuss the implications of our findings for the role of AI in communication governance. We find that, despite societal expectations that we can design ethical AI, and public expectations that developers and governments should share responsibility for the outcomes of AI use, there is a significant divergence between these expectations and the ways in which AI technologies are currently used and governed in large scale communication systems. We conclude that discourses of ‘ethical AI’ are generically performative, but to become more effective we need to acknowledge the limitations of contemporary AI and the requirement for extensive human labour to meet the challenges of communication governance. An effective ethics of AI requires domain appropriate AI tools, updated professional practices, dignified places of work and robust regulatory and accountability frameworks.

Keywords

Ethics, expectations, science and technology studies, artificial intelligence, performativity, communication governance

This article is a part of special theme on The Turn to AI. To see a full list of all articles in this special theme, please click here: <https://journals.sagepub.com/page/bds/collections/theturntoai>

Introduction

Today artificial intelligence (AI) technologies, such as machine learning (ML), are embedded in everyday communication services. AI is also at the centre of an immense positive, and future orientated discourse disseminated by national research programmes, consultancy reports and corporate statements. AI and data science are European research and innovation priorities with significant resources being directed towards solutions which may deliver economic and social impacts. However, a number of critical voices and corporate scandals have highlighted the potential negative impacts of contemporary AI on employment, democracy and equality. This article identifies the key actors shaping contemporary societal expectations of AI in

the UK and Ireland over the past decade and the mechanisms they use to do this. We also map the range of expectations being created, and the emergence of a discourse focussed on ethics and AI. Finally, we explore

¹ADAPT Research Centre & Department of Sociology, Maynooth University, Maynooth, Ireland

²ADAPT Research Centre & School of Information and Communication Studies, University College Dublin, Dublin, Ireland

³ADAPT Research Centre & Information, Communication and Entertainment Research Institute, Technological University Dublin, Dublin, Ireland

Corresponding author:

Aphra Kerr, ADAPT Research Centre & Department of Sociology, Maynooth University, Maynooth, Co. Kildare, Ireland.
Email: Aphra.kerr@mu.ie



the implications of our findings for AI in communication governance.

We situate this study within the sociology of expectations (Pollock and Williams, 2016; van Lente, 2012) and apply this approach to identify the actors who are issuing both formal and informal articulations that shape public expectations of AI, and to assess what these expectations are attempting to do. Our empirical work proceeds at two levels: first, an analysis of public documents to identify the key actors, mechanisms and issues currently shaping societal expectations of AI; and, second, an analysis of a survey of members of the public exploring awareness and expectations to AI. Our analysis reveals shared expectations that we can design ethical AI and that developers (especially academic ones) will behave ethically; however, challenges remain around understanding the most important ethical issues, how these vary from domain to domain, and deciding who should be responsible and accountable for negative impacts.

The sociology of expectations distinguishes between generic, or weak, and effective, or strong, performances of expectations. Strong societal expectations can influence the dynamic, direction and focus of technological innovation. In our discussion, we explore the implications of societal expectations of ethical AI for current professional practice in the design and application of AI in communication governance. AI is just one dimension of contemporary communication governance assemblages for transnational digital communication services (Kerr et al., 2014), yet AI is often presented as the sole solution to the scale, speed and complexity of that task. This solutionism is driven by powerful stakeholders and backed by significant investment. It obscures the socio-technical limitations of Big Data based forms of AI when faced with highly contextual cultural communication and the requirement for extensive human labour to train, update and support the technologies. We ask if societal expectations that we can create ‘ethical AI’ is simply a generic discourse for reassuring investors, governments and users? Furthermore, is the focus on ethical AI diverting attention away from other important ethical issues, including how data are gathered, the conditions under which humans develop and deploy AI in communication governance, and the impact on users?

AI and the sociology of expectations

The history of AI is an undulating one, with positive and negative expectations, and periods of growing hype followed by disappointments and decline. After the Second World War a number of new research areas developed in the US and the UK including cybernetics, communication theory and AI. Unlike early

communication theory, which was dominated by transmission models of communication, AI had competing definitions, goals and agendas. Fleck (1982) traces the first use of the term ‘artificial intelligence’ to the US in 1956 where the focus was on developing computer systems with high-level reasoning in planning and mathematics. This meeting gave rise to a ‘proto-paradigmatic’ (Fleck, 1982: 177) approach which legitimised and coordinated subsequent AI research and investment in the 1960s and 1970s and gave rise to the first significant period of positive expectations in AI. However, by the mid-60s it became apparent that the early approaches to AI could not scale to real-world problems. This created the first ‘AI winter’.

A second period of growth in positive expectations in AI occurred between the late 1960s and the early 1980s. This time the technological driver was a focus on encoding expert domain knowledge into ‘expert systems’. For example, the MYCIN system used approximately 450 hand-coded rules to diagnose blood infections (Buchanan and Shortliffe, 1984), the encoding of which was a highly labour intensive and costly approach. Consequently, although expert systems proved successful in a number of specialised applications, by the late 1980s many investors became disillusioned. Further, the negative findings of the Lighthill review of AI research in the UK undertaken in 1973 significantly depressed expectations in AI. This resulted in much negative media coverage and led ultimately to a reduction in investment in AI research in the UK and the US (Fleck, 1982: 191–192).

Current interest in AI can be viewed as a third period of growth in expectations and, as with each of the earlier cycles, it has a distinct technological trigger (Department of Business, Enterprise and Innovation (DBE), 2018; European Commission (EC), 2018; House of Lords Select Committee on Artificial Intelligence (HLSCAI), 2018). The promise is that ML algorithms can extract non-obvious patterns from large datasets, and that the revealed patterns will provide (novel) insights into problems that will ultimately result in better decisions (Kelleher and Tierney, 2018). The connection between modern AI and Big Data positions it within the broader developments of the growing proliferation of sensors throughout society, the shift to online communication, and the growth of business models and national policies which rely on the tracking of individuals. One consequence is that modern AI (unlike earlier generations) is essential for deriving maximum value from data. Another is the development of consumer products/services, such as speech interfaces for smart phones, smart speakers or facial recognition software that AI facilitates. This could mark a change in its fortunes, as technologies

that become commercialised and domesticated are likely to endure (Barry et al., 2017).

This brief history reveals that the meaning of AI has evolved over time, and today the term can mean different things to different people. For some, the ultimate goal of AI is to develop autonomous agents with general intelligence, similar to human intelligence. For others, the goal is to develop systems capable of specialist expert behaviour in a narrowly defined domain. Further variations arise in the methods used to create AI systems, with early research characterised as rule-based symbol processing where the rules were hand-crafted by a knowledge engineer. Recent AI systems, however, are better characterised as data driven ML systems (Kelleher, 2019). This diversity in the meanings of AI and the methods used is important in any discussions of expectations and impact. All of these AI systems raise important ethical questions, but the questions differ. Further, while data driven AI systems are already present in communication, the ‘singularity’,¹ may never be possible.

In order to understand the current ‘turn to AI’ this article draws upon the sociology of expectations from science and technology studies (STS). STS has long explored the role that expectations perform in the management of uncertainty in innovation processes. Expectations can be defined as statements that say something about the future (Borup et al., 2006, van Lente, 2012). A range of actors, and both formal and informal mechanisms, shape expectations (Pollock and Williams, 2016; van Lente, 2012). Formal mechanisms include foresight and research prioritisation exercises, which are deployed by governments, consultancy firms and companies to rationalise future innovation agendas and investment. Informal expectations are ‘images, statements and prophecies’ (van Lente, 2012: 772) which circulate through social networks, the media, conferences and meetings. Some informal images and statements are provided by experts. Non-experts can also play a role in shaping expectations via the media and other activities. In the construction of expectations, a range of actors draw from, and add to, the repertoire of visions that create the innovation dynamic around a technology.

Collectively shared expectations can have a significant influence. Previous studies have identified three ‘forces’ of expectations: they raise attention and legitimate investment, they help to coordinate networks of companies and research institutions, and they provide heuristic guidance and direction to research and innovation activities (Van Lente, 2012: 773–774). Expectations can be said to be ‘performative’ in the sense that they ‘do’ something, and may prompt certain social actions. However, in his analysis of the use of econometric models in financial markets, Mackenzie

(2008: 17) distinguishes between ‘generic’ and ‘effective’ performativity. Generic performativity is when theoretical models, language or approaches are taken up in the real world but do not make a difference to how things are actually done, while effective performativity is when the language or models make a difference in practice.

Expectations can operate at different levels, from the European research area to the national or the company level, and the content can vary from a focus on technical, commercial or social aspects to more fictionalised narratives. In applying this approach to AI, we conducted two studies to provide a multi-level analysis of expectations of contemporary AI. First, we analysed key formal and informal documents and statements on AI released since 2011 in Europe, the UK and Ireland. Second, we conducted a survey of awareness and expectations of AI amongst members of the public in Ireland. Together, these studies are used to identify key actors, formal and informal mechanisms and shared or diverging expectations of AI. In the discussion we assess the performativity of these expectations and their implications for AI in communication governance.

Societal actors, mechanisms and expectations of AI

The first study followed the principles of theoretical sampling (Gentles and Vilches, 2017). We identified a range of secondary documents based on their theoretical relevance, i.e. they outlined the strategic visions of key actors in AI research and policy in the EU, UK and Ireland and thus are ‘performative’ in relation to the construction of expectations (Van Lente et al., 2013). Our sample consists of 41 documents published since 2011 by seven types of actors: international consultancies; the EC; the UK and Irish research administrations; expert consultations in the UK and Ireland; public surveys; professional associations and, statements by workers and whistle-blowers (see supplementary file for details). We view these documents as ‘conduits’ of communication (after Prior, 2008), which are impactful beyond their immediate publication. In what follows we focus on the key actors, the mechanisms used, and the ‘order of discourse’ linking, for example, consultants reports to public policy documents, and policy documents to statements in the media. We also detail those actors promoting a positive view, those promoting a negative view, and an emerging focus on ethics in/for AI.

Positive expectations – Governments, research institutes, IT consultants

The agendas of the national research councils in the UK and Ireland are developed in consultation with international IT consultancies and experts from academia and corporations, and influenced by strategic developments by the EC. Following a decade of austerity in Europe, governments, public research agencies and companies are working hard to legitimise public and private expenditures on research and innovation. Further, AI research is expensive in resource and computational terms. Many of the documents in this analysis justify investments in terms of economic growth, jobs and increasing efficiency. Some refer to competition from the US and China. The UK and Irish governments rely heavily on commissioned reports from business analysts and foresight exercises with experts to inform their national research programmes and priorities. In the UK this is supplemented by regular surveys of public opinion.

In both the UK and Ireland, we can identify coalitions of government departments, public research agencies, public research institutes and technology companies who are active in shaping expectations of AI. This aligns with the ‘co-ordinated’ approach to research policy, with a strong role for public bodies (Cath et al., 2018: 503). In our sample, we identify a largely positive set of expectations about the potential returns from investing in AI and can map significant investment in AI research, infrastructure and institutions (often of at least five-year duration) over the past decade. For example, the Turing Institute was established in 2015 as the UK’s National Institute for Data Science following a government and public consultation process (Department for Digital, Culture, Media and Sport, 2018).² It received almost £42 million in government funding. In 2017 it was renamed the National Institute for Data Science and AI, and received another £48 million in 2018 (Press Association, 2013). Data science and AI are listed as national priorities in recent UK research policy documents and the Turing Institute involves collaboration between government, corporations and universities around the UK. In Ireland, the national research prioritisation exercise (2018–2023) added AI to its priorities and noted a ‘potentially disruptive effect of Artificial Intelligence and Machine Learning’ (DBEI, 2018: 6). This was informed by a technology futures exercise involving corporate, governmental and STEM academic experts. These reports drive the agenda of multi-institutional academic research centres and other innovation programmes. In both the UK and Ireland, a shared feature of strategic and competitive

public research funding is that it must involve collaboration with industry.

The second actors shaping expectations around AI are IT analysts and consultancy firms. Pollock and Williams (2016) argue that industry analysts provide a significant new form of expertise for technology companies, investors and governments, and are a crucial ‘knowledge intermediary’ in relation to complex and uncertain futures. They also play an important role in naming emerging technical fields. McKinsey, Accenture and Gartner are particularly important as their reports and expertise are evident in UK and Irish government policy and programme statements. The Gartner Group describes itself as a ‘decision-support’ company, and not a forecasting or market research company. It is best known for its IT Hype Cycle, which identifies five phases of a technology’s lifecycle: the innovation trigger, the peak of inflated expectations, the trough of disillusionment, the slope of enlightenment and the plateau of productivity. The last decade of Gartner’s Hype-Cycle reports includes a number of AI related technologies: Big Data was at peak hype in 2013, ML in 2016, and deep learning in 2018. While Gartner presents its hype cycle as an educational tool, it can also be understood as a branding and promotion tool (Williams and Pollock, 2015: 1387). Crucially from a sociology of expectations perspective, while Gartner reports are a formal mechanism shaping expectations, images of Gartner’s Hype-Cycles are an informal mechanism which circulate widely in the media and in professional talks. They define key concepts for technological trends and thus can be seen as ‘generally performative’ in establishing expectations around emerging IT fields, including at the moment, AI.

Our analysis of publicly available consultancy reports on AI found a focus on ML and deep learning. They identified significant ‘cost reductions’ and ‘efficiency’ gains to be made by corporate investors including around the automation of decision making and customer engagement. Using reassuring language and a positive tone, these documents provide legitimisation through authoritative statements. For example, McKinsey (2018: 7) provides hundreds of AI use cases to show where ‘value is to be captured’, while Accenture Labs (2018) model how AI will double the growth of economies by 2035. Gartner (2019) helps clients to ‘separate AI hype from reality’ according to a five phase ‘AI Maturity Model’: ‘as enterprises evolve their AI expectations and projects, the technology will mature to have more transformative and strategic impacts’. These positive visions help to mediate uncertainty (Borup et al., 2006), and their terminology is referenced in many policy documents. For example, in 2014, the Irish policy agency Forfás published a

report on future Big Data skill needs. Their definition of ‘deep analytical talent’ came from a McKinsey Global Institute study (Manyika et al., 2011). The EC (2018) Communication on AI also cites McKinsey and Accenture reports.

Negative expectations – Workers, whistle-blowers and contrarians

Since 2011 whistle-blowers and workers have made a number of revelations surrounding the use of data science and AI by companies and governments. The first involved Edward Snowden in 2013, a contractor for the National Security Agency in the US. In a series of coordinated media stories, he revealed the extent of the transnational state security surveillance apparatus and dataveillance of online communications (Lyon, 2014). Then the Cambridge Analytica scandal in 2017/2018 saw contractor, Christopher Wylie, revealing problematic data practices at Facebook and other companies (Cadwalladr and Graham-Harrison, 2018). Both these revelations triggered major controversies, significant media commentary and public inquiries. Further, the media has covered how groups of employees at Google and Amazon have publicly refused to cooperate with AI development for military and law enforcement purposes,³ while other stories have revealed problematic work practices for moderators of AI-driven social media content (O’Connell, 2019).

Between 2014 and 2018 another set of well-known individuals made high profile negative comments about future AI, including the academic physicist Stephen Hawking, and entrepreneurs Bill Gates and Elon Musk. These commentaries largely focus on the potential that AI might become super-intelligent. Galanos (2019) traces how these ‘contrarians’ have been widely cited in EC, UK and US policy documents and notes that while these entrepreneurs and academics are experts in their own domains, none are experts in AI. While statements by whistle-blowers and workers focus on the use and misuse of AI in everyday practices, contrarians focus on a future AI, which may never exist.

Ethical issues and a European approach to AI

These negative interventions have been, in Mackenzie’s terms, ‘generically’ performative in structuring expectations around AI. Since 2016 a focus on ethics in relation to the development, deployment and use of AI has emerged in AI policy and strategy documents in our sample. The Royal Society in the UK has conducted a programme of activities on AI and ML, identifying public concerns about potential harms, worker replacement and impacts on human experience and choices

(Royal Society, 2017). In 2016, the Turing Institute added a data ethics research group and an ethics advisory group. In the same year, the Ada Lovelace Institute was established (£5 million over 5 years) to specifically ‘inform public understanding of AI and data science’, to create a ‘shared understanding of the ethical issues arising from data and AI’ and to ‘define and inform good practice in the design and development of AI’. Finally, a House of Lords’ Select Committee on Artificial Intelligence stated that: ‘the Government must understand the need to build public trust and confidence in how to use artificial intelligence, as well as explain the risks’ (HLSCAI, 2018: 25). In Ireland, a number of recent EU funded research projects have focused on ethics training in computer science and some research centres are adding a focus on ethics; however, to date there has been no significant public programme or research funding on the societal and ethical aspects of AI.

A similar turn to ethics is also evident in IT consultancy reports since 2016. In the McKinsey 2011 report, ethics is positioned as a ‘soft skill’ but there is no discussion of ethical or public interest issues. However, another McKinsey report (Henke et al. 2016) notes that ‘thorny ethical and societal questions ... will have to be addressed as machines gain greater intellectual capabilities’ (p. 81) while observing that these questions have ‘entered the public sphere most recently in the context of autonomous vehicles, and they seem likely to be relevant in many other contexts’ (p. 93). PWC (2017: 21) advise that organisations ask themselves if they have ‘considered the societal and ethical implications’. Accenture Labs (2018) advocated a code of ethics for AI, suggesting that ‘ethical debates should be supplemented by tangible standards and best practices in the development and use of intelligent machines’. However, none of the consultant documents in our sample offer any detail on specific ethical challenges, or how to deal with them. Meanwhile, the codes of ethics for professional organisations including the Association for Computing Machinery (ACM), and the Institute of Electrical and Electronics Engineers (IEEE) are being updated to reflect the impact of AI on practice, e.g. the ACM Code of Ethics and Professional Conduct (2018) has expanded the responsibility of computer engineers to include ‘awareness of the social context in which the work will be deployed, and competence in recognizing and navigating ethical challenges’. The IEEE (2019) has released a set of AI-related standards that advise developers to follow ‘values-based design methods’ but acknowledge a preference for identifiable and quantifiable ‘norms’ over values.

A growing focus on ethics is particularly evident in European policy documents. In March 2018, the

European Group on Ethics in Science and New Technologies called for the EC to develop a common ethical and legal framework for AI (European Group on Ethics in Science and New Technologies (EGE), 2018), and shortly after an EC Communication proposed a ‘European approach’ to AI (EC, 2018). This communication called for applying AI for ‘good and for all’, and the centrality of the ‘Union’s values and fundamental rights as well as ethical principles such as accountability and transparency.’ The EC proposes tripling research investment in AI at the European level, supporting increased training and education in AI, and development of appropriate legal and ethical frameworks. Ethics is specifically mentioned as important in relation to the development and the use of AI (EC, 2018: 13). In late 2018, the EC released its draft guidelines for ‘trustworthy AI’ based on the work of 52 experts around Europe. These generalised ethics guidelines tend to conflate all AI applications,⁴ implying that the negative potential of AI is not influenced by the context and purpose of its use. While formal principles-based initiatives can shape societal expectations, and offer some reassurance, it is unclear as yet how these rhetorical shifts will impact on practice in specific domains.

Public expectations and the ethics of AI

Public surveys on attitudes to science and technology are frequently conducted in Europe and the UK. By contrast, and despite the significant rise in investment in AI in Ireland, there has been no research investigating national or local attitudes to AI since 2011.⁵ Therefore, an exhibition featuring international artistic and research installations on AI at the Science Gallery in Dublin 2017 provided an opportunity to engage directly with members of the public. The survey was administered face-to-face over one week, and 164 individuals completed questions on their awareness of AI, outlined their positive and negative expectations around AI, and gave their views on the most important ethical issues associated with AI.

The survey is not a representative public survey, but rather provided an opportunity for researchers to explore the expectations of a scientifically interested public. Respondents were 55% male, aged 18–34 years (75%) and well educated to degree level (70%). They were predominantly white (84%), heterosexual (87%) and many were non-religious (62%). Respondents came from 25 different countries including 27% from Ireland, 13% from America and 9% from Britain, followed by Germany and Brazil. Many had moved to Ireland for work. Half were working full time, and another 10% were self-employed. The balance were working part time, were unemployed, not

available for work or were visiting as tourists. They worked across a range of service industries, including IT, education, the arts, finance and hospitality. Only two worked in manufacturing, and one each in agriculture and construction. Just over 60% of respondents earned less than €35,000 a year (see supplementary files for full demographics).

Respondent familiarity with AI ranged from expert (1.8%), and very familiar (18.3%), to those with good (25%), some (51.2%) or no familiarity with AI (3.7%). Respondents associated AI with gaming, scientific research, manufacturing and communications. The technologies most associated with AI were self-driving cars, robots, and communication assistants (e.g. Siri). Other popular answers included a range of communication technologies including games, internet ‘bots’, recommender systems and search. The fact that respondents ranked technologies they had not experienced (robots and self-driving cars) most frequently indicates the influence of positive and negative utterances on shaping societal expectations about AI.⁶ However, negative statements in the media and everyday experiences of applications, such as games and communication applications, were also important in relation to their understanding of AI.

Positive and negative expectations of AI

Given that our respondents had relatively good awareness of AI, in this section we focus on answers to two open ended questions which asked about the primary positive and negative issues related to AI. We conducted a thematic analysis (Braun and Clarke, 2006) on the 198 positive and 144 negative responses, to identify patterns of meaning that respond to our focus on actors, mechanisms and expectations of AI. This theoretically independent method is appropriate to our sociology of expectations approach. Both positive and negative responses were coded, with particular attention paid to future orientated phrasing such as ‘may’, ‘might’, ‘could’ or ‘would’. From the codes, we identified a set of themes describing positive and negative expectations.

Overall respondents felt that AI presents good opportunities for economic growth and social progress. Participants most frequently identified automation and efficiency as positive aspects of AI, stating that AI will automate dull, menial or repetitive tasks and make certain types of services more efficient and accessible. Respondents thought AI would impact positively on medicine, science, knowledge and the environment. Some felt that AI might be more reliable and eliminate human bias and errors. Some described AI as ‘servant’, or ‘helper’ and other broader claims are provided below (see Table 1).

Table 1. A sample of positive expectations of AI.

Theme	Participant responses
AI as servant	'Menial tasks that restrict our time <i>could</i> be done using AI and we could have more time for better pursuits in many areas including creativity and scientific advancement.' 'Robots <i>could</i> do hard repeating work.'
AI as helper	'AI <i>may</i> help in our daily lives. It solves technical problems and makes technology more efficient.' 'Advancements in technology that <i>could</i> help people around the world. AI being implemented into medicine <i>will</i> be especially beneficial.'
Other	' <i>Potential</i> for improvements in communication, healthcare etc.' 'Grow the <i>potential</i> for human knowledge and achievement.'

Table 2. A sample of negative expectations of AI.

Theme	Participant responses
AI as a thinking machine	'Human laziness, we <i>may</i> use our brains less if a machine can think for us, it <i>could</i> affect our free will by making decisions based on the expertise of AI.' 'Humans <i>will</i> lose sight of what it means to work hard and <i>could</i> become too dependent on machines to the point where we're unable to live without them and . . . live independently.'
AI and inequality	'The effect that it <i>could</i> have on certain working classes and unemployment that <i>could</i> arise from this improved efficiency.' 'The same problems that humans have <i>may</i> imprint on the system.'
AI and humanity	'It <i>could</i> take away some of the basics aspects of human social life' ' <i>Could</i> dehumanise. <i>Could</i> be abused.'
Fears, including of mis-use	'Humans <i>potentially</i> playing God and not being conscientious of the repercussions.' 'With the amount of corruption and unethical behaviour in this world, AI in the wrong hands <i>could</i> be extremely dangerous.'

Respondents also listed automation most frequently as a negative aspect of AI, stating that automation might result in job losses, too much standardisation and out of control machines. Automation was closely followed by concerns related to security, privacy and surveillance. A third set of concerns related to the impact of AI on 'humanness, knowing and knowledge'. Respondents worried about the impact of AI on personal and social relationships, on emotions and subjectivity. They were concerned about the level of public expenditure on making machines smarter, rather than humans, and wondered if humans should trust machines to predict, rationalise, and rank information. While the loss of jobs and computational creativity were themes in the exhibition, and considerations about unemployment and surveillance are widely discussed in public articulations, there was a lot of concern in these responses for the trustworthiness of automatically created information, a loss of human creativity, and the lack of empathy in automated decisions (see Table 2).

Overall, the positively expressed expectations reflect the benefits of AI to *individuals* – by taking over jobs that humans may not wish to or cannot do. Thus,

positive expectations view AI as a beneficial tool in relation to social progress. The negatively expressed expectations tend to focus more on *group* impacts such as unemployment, or how AI might affect human creativity and subjectivity. Further, they noted the potential for increasing economic and social inequality or malicious uses, suggesting a view of AI as a set of interconnected systems rather than individual applications. We note that some of the positive aspects of AI (as helper or servant) were directly related to the negative (unemployment). These responses echo concerns found in public surveys on AI in other countries including: inappropriate or invisible use of AI (Fast and Horvitz, 2017), removal of human oversight, judgment and decision making (Ipsos/MORI, 2016), making generalised predictions about individuals (Royal Society, 2017) and wealth inequality (Eurobarometer, 2015). However, our survey goes beyond previous work as open-ended questions allowed respondents to outline how automation and efficiency could be both positive and negative. Further, concerns emerged about the centralisation of control, the declining quality of jobs and the errors created by these systems. They were also concerned

about the vulnerability of these systems, although there were no direct references to military applications.⁷

Expectations around ethics and responsibility

Our respondents mostly experience AI in everyday communication and gaming applications, but the ethical concerns of our respondents tended to focus on ‘robots’ and future technologies. Half of our respondents had ‘some familiarity’ with ethics in technology design. When offered a list of ten ethical principles to rank in importance, the four most highly ranked were, respectively, safety, privacy, transparency and security. These were significantly beyond other principles including integrity, non-discrimination, autonomy, dignity, efficiency and equity respectively.⁸ Privacy and transparency are dominant in formal and informal statements on ethics of AI in Europe and are the focus of attention in technology developments. However, the significance of safety to our respondents is notable and further supported by their responses in relation to positive and negative expectations of AI.

Over two-thirds of our respondents agree that it ‘is “possible” to design AI in an ethical manner’. However, participants were evenly split between those who agree and disagree that we can ‘design AI to respect human dignity and values’. A clear majority felt that AI demanded a ‘higher ethical standard’ than other areas of technological design. When probed, a majority (63%) agreed with the statement that ‘AI reflects the biases of its engineers and designers’. When asked whether AI designers are ‘responsible for the use and misuse of the technical systems they build’, over half of participants agreed. When asked ‘who is in control of AI design?’ they answered: industry (28%), academic research (23%), and research funders (21%) followed by governments, regulators, other (UN, social media corps, no-one). When asked ‘who do you feel *should be responsible* for AI impact’ respondents gave different answers: governments (23%), state regulators (21%), industry (18%), academia (18%), followed by funders, others (independent authorities). In other words, industry and academic researchers are viewed as *in control* of the trajectory of innovation, whereas public authorities *should* be responsible for impact.

Our interpretation of these findings is that the issues which relate most directly to a positive *individual* experience with AI, and are widely discussed in society by a range of actors, are dominant for our respondents. Safety is a key issue in relation to future technologies, including autonomous cars, but it is also relevant to location enabled communication technologies. Safety also features in fictionalised narratives and respondents mentioned the Terminator movies. Privacy,

transparency and responsibility are important for our respondents and highly evident in our sample of formal and informal statements on AI.⁹ Other ethical issues may be less clearly understood, although there appears to be agreement on a role for governments or public agencies in managing AI impact while greater trust is placed in AI researchers working in universities. These findings are again mostly in line with large-scale surveys in the UK, which found an expectation that academics ‘would be the driving force behind it (AI) in its embryonic and early stages,’ (Ipsos/MORI, 2017: 50) with concerns that private industry is ‘not subject to the same scrutiny and accountability as the public sector’ (Ipsos/MORI, 2017: 47). The public expectation of state responsibility for controlling AI impact in the UK and Ireland is important, but complicated by the collaborative nature of AI research and innovation programmes in these countries, as identified in our document analysis.

Overall, it is clear that our respondents have some difficulty conceptualising abstract ethical principles and values, but their attitudes are influenced by societal expectations, high profile news and fictional stories and to a more limited degree their own experiences. They are more concerned with things that *might* occur rather than more mundane questions such as what the most appropriate levels of automation in information, entertainment and public services should be. The latter is as much a social and regulatory question as an ethical one, and is linked to our expectations and traditions of political accountability and transparency (Eubanks, 2018). If scientifically literate members of the public have difficulty grasping abstract ethical issues, we need to carefully consider how employees who are tasked with designing and deploying these systems will fare.

The performativity of expectations of AI and the practice of communication governance

Our multi-level analysis suggests that a range of actors construct the expectation that AI technologies will provide significant opportunities for economic growth but also raise a range of ethical concerns. These expectations are useful for justifying significant investments, and supporting a particular innovation dynamic, but they say little about the challenges of applying these technologies in particular social contexts. Contemporary AI is dominated by ML approaches including deep learning. It is methods focussed and domain agnostic. If we are to understand the effective performance of societal expectations of AI and emerging ethical frameworks they must be tailored to specific

domains. One such domain is communication governance.

As communication platforms and services like Facebook and Twitter extend across borders, the scale and speed of data being dealt with has grown exponentially. The current application of AI in commercial communication governance highlights a number of limitations. For example, AI technologies base their predictions on pre-existing big datasets and thus they make their decisions by looking backwards, and without understanding culture, context and meaning. Moreover, human communication constantly evolves and is highly contextual. Some clearly defined words, images and practices can be automatically detected and removed, but identifying new forms of appropriate and acceptable content is challenging (Gillespie, 2018). Moderating harassment, racism, sexism, and misinformation online is highly complex, and cannot be achieved by AI alone (Ging and Siapera, 2018). Large scale gaming communities like Activision Blizzard's *World of Warcraft* use AI but are still plagued by privacy violations, user harassment and unauthorised cheating (Kerr, 2017).

Contemporary communication governance is a multi-level phenomenon where solutions like AI are only one part of a complex assemblage involving human and non-human actors and including legal documents, in-house community management policies and national or regional regulatory frameworks (Kerr et al., 2014). Despite the application of AI, most commercial games and communication platforms rely on users to flag unfair, abusive or disturbing content. They also rely on databases being constantly updated by significant human labour. This is not the highly skilled, well paid work that most documents on AI refer to. Despite societal expectations that AI will automate poorly paid occupations, AI requires extensive human decision making – from training tools, to moderating and removing unwanted content. A large number of poorly paid community and content roles are required in communication governance (Roberts, 2018). They make decisions on removing culturally and contextually difficult flagged content and whistleblowers have revealed the psychological damage of this work (Kerr and Kelleher, 2015; O'Connell, 2019). While AI may reduce the financial and psychological costs of some human moderation at scale, AI still requires human workers to make complex communication decisions, and in some instances to explain unfair automated decisions. Thus, AI can be a helpful servant, but in many instances the humans are serving and helping AI overcome its limitations.

Our study reveals a disconnect between societal expectations and practice around the design and deployment of AI. For current computer and data

scientists, ethics and responsibility remain at an abstract level and have not been part of their professional training. Nor are they part of current agile and iterative approaches to data science projects such as 'The Cross-Industry Standard Process for Data Mining' (CRISP-DM). The 76-page CRISP-DM manual contains just one reference to 'ethics' where it is framed in terms of 'constraints' and limited to the use of the data (Chapman et al., 2000: 33). The most recent professional guidelines from the ACM Code of Ethics and Professional Conduct (2018) and the IEEE (2019) that we reviewed recognise that those building intelligent systems in different domains may face ethical challenges, but as yet there are few practical solutions as to how these challenges can be addressed, and it is unclear what sanctions are imposed on those who violate them. Abstract values, while useful for guiding high level ethical discourse, are difficult to apply in practice, as observed with the variety of competing metrics to measure 'fairness' in AI development (Bird et al, 2019). And while 'transparency' is a prominent public concern in our survey, studies suggest we are unlikely to develop a kind of AI transparency the public can understand (Burrell, 2016), and it may be an unrealistic ideal distracting us from persistent civic rights concerns and deeper ideological and political questions of accountability when AI solutions are deployed in different domains (Ananny and Crawford, 2018; Dencik et al., 2017; Eubanks, 2018).

Both high level guidelines and professional codes of ethics are an example of the formal articulation of societal expectations of AI. They tend to reflect ethical frameworks most aligned to deterministic views of prediction and are focused on outcomes, rights and duties of what are assumed to be autonomous individuals. These frameworks mostly operate as corporate and professional responses to societal calls for more regulation of AI, but may be inadequate for dealing with highly contextual and dynamic human communications. Alternative ethical frameworks might more effectively describe the relational interactions involved in developing and deploying AI – especially for public services and communications. For example, virtue ethics can address cultural and political aspects of design practice (Barry et al., 2017), while an ethic of 'care' could focus on data relationships in communication governance (Barry and Kerr, 2018). Further, existing communication policy and practice has a history of ideals, norms, and laws that seek to balance the individual and collective good in communication. There is now a strong need for these broader national and transnational communication policies to evolve to take account of the role and impact of AI on mediated forms of human communication.

Conclusion

The sociology of expectations offers a useful approach for identifying the actors, mechanisms and expectations emerging around AI, and assessing the divergence between expectations and current practices in the application of AI. In our analysis, we identified a range of actors articulating positive expectations of AI over the past eight years through formal reports, public consultations and research programmes. From 2016 we identified significant activity at European, national and professional levels to agree high level ethical principles and offset both negative statements and the revelations of whistle-blowers and workers on current applications of AI. While ethical guidelines and principles are proliferating, and far from uniform, it is unclear how they can be operationalised in different domains. In practice, corporations and coalitions of industry and public research bodies *design* AI, and both corporations and public organisations apply AI, but there is ambiguity around who is responsible for monitoring its impact, or what sanctions can be imposed. Meanwhile, members of the public believe it is possible to design ‘ethical AI’ and that biases present in AI reflect its designers, who are also in part responsible for its use and misuse. This quite accurately describes the bind in which many of those working in AI development find themselves, both under increasing pressure from society to engage in ‘ethical’ AI practice, and expected to adhere to a range of organisational and professional incentives which may be in conflict with these ethical positions.

Recent attempts to develop ethical guidelines and to design ethical technological solutions, including on explainability, are clearly responses to wider societal concerns about the limitations of contemporary forms of domain agnostic AI. However, these approaches are still limited and problematic when applied in the communication governance contexts. At best, ethics guidelines are generically performative, operating at a linguistic level to assuage and deflect critique and regulation. Indeed, we argue that ethics discourses and solutions are currently operating as *assurance* for investors and the general public, rather than as an effective tool for governance – either of developers or of communication service users. The gap between societal expectations of AI, and AI in practice, will remain until we understand and accept the limitations of AI in complex social contexts and recognise that non-technological policies and human workers are required to make AI work ethically. Organisational and professional policies must be supplemented by robust transnational policies and regulations that make corporations, research institutes and organisations responsible for the human labour involved in deploying AI, and accountable for its social impacts.

Acknowledgements

We would like to thank the reviewers for their helpful comments. We would like to acknowledge Profs Linda Hogan, Declan O’Sullivan and Dave Lewis of the ADAPT research centre for their support with the survey. We would like to thank Ms Clóna Rooney and Mr Joshua Savage (MU) for their research assistance, and Ms Dearbháil Ní Chúirc, Ms Maighread Tobin (MU), Dr Wessel Reijers (DCU), Ms Ramisa Gadpez Hamed, Mr Ensar Hadziselimovic, Mr Pandit Harshvardhan, Dr Kevin Doherty, and Dr Kevin Koidl (TCD) for administering the survey. We would like to thank the Science Gallery, Dublin for facilitating the research. Aphra would like to thank Prof. Robin Williams, Prof. Steven Yearley, Dr James Stewart and Mr Vassilis Galanos, at the University of Edinburgh.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work was supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. Aphra would like to acknowledge the support of the Institute for Advanced Studies at the University of Edinburgh from February–May 2019.

ORCID iDs

Aphra Kerr  <https://orcid.org/0000-0001-5445-7805>

Marguerite Barry  <https://orcid.org/0000-0001-5271-2673>

John D Kelleher  <https://orcid.org/0000-0001-6462-3248>

Supplemental material

Supplemental material for this article is available online.

Notes

1. The singularity refers to the prospect that ‘ordinary humans will be overtaken by artificially intelligent machines or cognitively enhanced biological intelligence or both.’ See Shanahan (2015).
2. See <https://www.turing.ac.uk/about-us>
3. See supplementary file for sources.
4. For example, ‘citizen scoring’ appears alongside ‘lethal autonomous weapons systems’ in High-Level Expert Group on Artificial Intelligence (2019: 34).
5. In October 2019 the Irish government launched a national public consultation on AI.
6. Eurobarometer (2017) found very low levels of use of robots at home or in work in Europe.
7. Public research programmes in Ireland do not invest in military research.

8. The list of ethics principles was based on our literature review and inputs from the ADAPT ethics and privacy working group.
9. The EU General Data Protection Regulation was introduced in 2016, elevating awareness of privacy.

References

- Accenture Labs (2018) Understandable AI: Explaining Machines. Available at: https://www.accenture.com/_acnmedia/pdf-85/accenture-understanding-machines-explainable-ai.pdf
- ACM Code of Ethics and Professional Conduct (2018) Association for Computing Machinery (ACM). Available at: <https://www.acm.org/code-of-ethics>
- Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20: 973–989.
- Barry M, Doherty K, Marcano Bellisario J, et al. (2017) mHealth for Maternal Mental Health: Everyday wisdom in ethical design. In: *Proceedings of CHI conference on human factors in computing systems 2017*, May 06 – 11, 2017, Denver, CO, USA. DOI: 10.1145/3025453.3025918
- Barry M and Kerr A (2018) Everyday justice in data science: What can ethics do? In: *Data justice conference*, Cardiff, UK, 21–22 May 2018.
- Bird S, Kiciman E, Kenthapadi K, et al. (2019) Fairness-aware machine learning: Practical challenges and lessons learned. In: *WSDM proceedings of the twelfth ACM international conference on web search and data mining*, Melbourne VIC Australia February, 2019, pp. 834–835.
- Borup M, Brown N, Konrad K, et al. (2006) The sociology of expectations in science and technology. *Technology Analysis & Strategic Management* 18: 285–298.
- Braun V and Clarke V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2): 77–101. DOI: 10.1191/1478088706qp063oa
- Buchanan BG and Shortliffe EH (1984) *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison Wesley.
- Burrell J (2016) How the machine thinks: Understanding ‘opacity’ in machine learning algorithms. *Big Data & Society*. Epub ahead of print 2016. <https://journals.sagepub.com/doi/full/10.1177/2053951715622512>
- Cadwalladr C and Graham-Harrison E (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, 17 March.
- Cath C, Wachter S, Mittelstadt B, et al. (2018) Artificial intelligence and the ‘good society’: The US, EU, and UK approach. *Science and Engineering Ethics* 24: 505–528.
- Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C and Wirth R (2000) CRISP-DM 1.0 Manual: Step by Step Data Mining Guide. NCR Systems, DaimlerChrysler AG, SPSS Inc., and OHRA Verzekeringen en Bank Groep B.V.
- Dencik L, Hintz A and Carey Z (2017) Prediction, pre-emption and limits to dissent: Social media and big data uses for policing protests in the United Kingdom. *New Media & Society*. <https://journals.sagepub.com/doi/full/10.1177/1461444817697722>
- Department for Digital, Culture, Media and Sport (2018) *Centre for Data Ethics and Innovation. Government Response to Consultation*. London: UK Government.
- Department of Business, Enterprise and Innovation (DBEI) (2018) *Research Priority Areas 2018–2023. Department of Business, Enterprise and Innovation*. Dublin: Government of Ireland.
- Eubanks V (2018) *Automating Inequality. How High-Tech Tools Profile, Police and Punish the Poor*. New York, NY: St Martin’s Press.
- Eurobarometer (2015) *Autonomous Systems. Special Eurobarometer 427*. Brussels: European Commission.
- Eurobarometer (2017) *Attitudes towards the Impact of Digitisation and Automation on Daily Life, No. 460*. Brussels: European Commission.
- European Commission (EC) (2018) Artificial intelligence for Europe. *European Commission COM (2018) 237*. Brussels: Communication, 25 April.
- European Group on Ethics in Science and New Technologies (EGE) (2018) *Statement on Artificial Intelligence, Robotics and ‘Autonomous Systems’*. EGE. Brussels: European Commission.
- Fast E and Horvitz E (2017) Long-term trends in the public perception of artificial intelligence. In: *Thirty-first AAAI conference on artificial intelligence*, San Francisco, California, USA, February 4-9, 2017. <https://aaai.org/ocs/index.php/AAAI/AAAI17/index>
- Fleck J (1982) Development and establishment of artificial intelligence. In: Elias N, Martins H and Whitley R (eds) *Scientific Establishments and Hierarchies. Sociology of the Sciences*. London: D. Reidel Publishing Company 6: 169–217.
- Galanos V (2019) Exploring expanding expertise: Artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018. *Technology Analysis & Strategic Management* 31: 421–432.
- Gartner (2019) CIO’s guide to artificial intelligence. Available at: <https://www.gartner.com/smarterwithgartner/the-cios-guide-to-artificial-intelligence/>
- Gentles SJ and Vilches SL (2017) Calling for a shared understanding of sampling terminology in qualitative research: Proposed clarifications derived from critical analysis of a methods overview by McCrae and Pursell. *International Journal of Qualitative Methods*. Epub ahead of print 2017. <https://journals.sagepub.com/doi/full/10.1177/1609406917725678>
- Gillespie T (2018) *Custodians of the Internet. Platforms, Content Moderation and the Hidden Decisions that Shape Social Media*. New Haven, CT / London: Yale University Press.
- Ging D and Siapera E (2018) Special issue on online misogyny. *Feminist Media Studies* 18: 515–524.
- Henke N, Bughin J, Chui M, et al. (2016) The Age of Analytics: Competing in a Data-Driven World. *McKinsey Global Institute*. Available at <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world>
- House of Lords Select Committee on Artificial Intelligence (HLSCAI) (2018) *AI in the UK: Ready, Willing, and Able? Report of Session 2017–19*. London: HLSCAI.

- Institute of Electrical and Electronics Engineers (IEEE) (2019) *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. 1st ed. Available at: <https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html>
- Ipsos/MORI (2016) *Public Dialogue on Ethics of Data Science in Government*. London: UK Government Data Science Partnership.
- Ipsos/MORI (2017) *Public Views of Machine Learning*. London: Royal Society.
- Kelleher JD (2019) *Deep Learning*. Cambridge, MA: MIT Press.
- Kelleher JD and Tierney B (2018) *Data Science*. Cambridge, MA: MIT Press.
- Kerr A (2017) *Global Games. Production, Circulation and Policy in the Networked Age*. New York, NY: Routledge.
- Kerr A, De Paoli S and Keatinge M (2014) Surveillant assemblages of governance in massively multiplayer online games: A comparative analysis. *Surveillance & Society* 12: 320–336.
- Kerr A and Kelleher JD (2015) The recruitment of passion and community in the service of Capital. Community managers in the digital games industry. *Critical Studies in Media Communication* 32: 177–192.
- Lyon D (2014) Surveillance, snowden, and big data: Capacities, consequences, critique. *Big Data & Society* 1. DOI: 10.1177/2053951714541861
- Mackenzie D (2008) *An Engine, not a Camera. How Financial Models Shape Markets*. Cambridge, MA: MIT Press.
- McKinsey (2018) Notes from the Ai Frontier. Insights from Hundreds of Case Studies. Available at: www.mckinsey.com/mgi
- Manyika J, Chui M, Brown B, et al. (2011) *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. New York, NY: McKinsey Global Institute. Available at: <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation>
- O’Connell J (2019) Facebook’s dirty work in Ireland: ‘I had to watch footage of a person being beaten to death’. *The Irish Times*, 30 March.
- Pollock N and Williams R (2016) *How Industry Analysts Shape the Digital Future*. Oxford: Oxford University Press.
- Press Association (2013) Eight great technologies’ benefit from £600m in government funding. *The Guardian*, 25 January.
- PWC (2017) Sizing the Prize: What’s the real value of AI for your business and how can you capitalise? Available from <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>
- Prior L (2008) Repositioning documents in social research. *Sociology* 42(5): 821–836.
- Roberts ST (2018) Digital detritus: ‘Error’ and the logic of opacity in social media content moderation. *First Monday*, 23.
- Royal Society (2017) *Machine Learning: The Power and Promise of Computers That Learn by Example*. London: The Royal Society.
- Shanahan M (2015) *The Technological Singularity*. Cambridge, MA: MIT Press.
- van Lente H (2012) Navigating foresight in a sea of expectations: Lessons from the sociology of expectations. *Technology Analysis & Strategic Management* 24: 769–782.
- van Lente H, Spitters C and Peine A (2013) Comparing technological hype cycles: Towards a theory. *Technological Forecasting and Social Change* 80: 1615–1628.
- Williams R and Pollock N (2015) Industry analysts – How to conceptualise the distinctive new forms of IT market expertise? *Accounting, Auditing & Accountability Journal* 28: 1373–1399.