EXECUTIVE SUMMARY

The **100+ Brilliant Women in AI & Ethics Conference** was created to highlight three key issues:

1. The absolute need for greater diversity in technology and STEM fields
2. The challenges facing the world in terms of AI/machine learning integration
3. The necessity for immediate action on the climate emergency

The day incorporated keynotes, a conversation between Jeanette Winterson and Baroness Beeban Kidron, and workshops for all delegates.

Key outcomes from the workshops included:

- The problem of late-stage capitalism as a systemic and economic underpinning for these technologies. There were many issues identified that stemmed from this, including:
  - Harms that have been identified from technology frequently have at their root a profit-making motivation – for example the tracking of children online that lays the groundwork (as far as some tech companies are concerned) for their smooth transition into consumers to be advertised to.
  - The need to address the question of whether the current model is the way society wants to incorporate life-changing technology that may permanently alter its course. (cf Surveillance Capitalism, Shoshanna Zuboff).
  - Themes such as the difference between growth and flourishing – measuring GDP and growth of AI cannot indicate the quality of life for human beings or the health of the planet and its non-human inhabitants. Altering measurement away from ‘growth’ metrics would be a powerful change in the public discourse, as current economic models drive development down paths that make it difficult to envision different possibilities – we need richer models for social success. Metrics are therefore key – GDP is inadequate in many ways⁠¹ and also unsustainable – the ‘growth’ model cannot continue. Specific interventions may be needed in order to shift the measures of health and success of current institutions.
  - The proprietary ownership of enormously powerful and influential algorithms is directly linked to this economic model and creates a situation where systems are implemented that have vast effects on citizens’ lives without any oversight or governance.

- Questions of societal values, and more robustly articulating the values that society should be diligently trying to embed, preserve and promulgate. Individual and societal wellbeing, reflected in areas as diverse as ‘work’ as a source of human flourishing, and algorithmic biases that unjustly punish portions of society that are already disadvantaged, led to deep thinking around the values that should be more
prominently foregrounded. These might include, for example, values of solidarity, cooperation and community – these currently lack weight when measured against values that prioritise individual protection, and individual justice and fairness. We need to foreground more community-driven values. It is also difficult to see how to shape an ethical future for algorithms in society when there are many different visions of the ‘good’ society. We need to investigate political and technological reforms that can help to fulfil the existing social contract – which is not being fulfilled by the ways that algorithms are currently developed and implemented. There was a desire for algorithms to enable higher and richer human modes of consciousness.

- Finding **new ways to think about and frame machine learning technologies** is key to understanding how public discourse and society can better grapple with the challenges they present. Discussions highlighted the value of social science concepts such as metaphor, analogy and narrative to understand the ways in which technology and humanity can co-exist and co-evolve. This was seen in examples such as the ‘dancing with the machine’ analogy in the World of Work session, and the question of ‘who is in control of the public narrative’ that came out of Algorithms in Society. It also underlies questions about whether social media should be viewed as more of a public utility. There is also a need to demystify AI so that genuine questions and challenges can be distinguished from hype and marketing.

- The **importance of retaining the ‘human’ element in machine-assisted systems** (rather than handing over to the machine altogether) – for example, requiring intelligibility from systems and accountability for algorithmic decisions. There is a need to articulate whether decision-making systems act in support of human decisions, or replace them. This might also be considered as ‘areas of effect’ for AI and algorithms – whether there are areas of decision-making that we decide are ‘no-go’ regions for algorithms. This was also reflected in the desire for tech solutions that are designed to be supportive and unobtrusive, running in the background of society rather than in the foreground.

- **Mapping out the relationships that are the fundamental building-blocks of society.** These can be as diverse as relationships of trust between citizens and their governments, relationships between individuals, relationships between employers and their employees. Not only can misdirected or misconceived technologies damage, corrupt or destroy these relationships, but even well-intentioned technologies may not always sustain or support ‘good’ relationships at scale. Community is an essentially federated enterprise that recognises different requirements and drivers at different granular levels. This includes relationships such as those between employers and employees – how do we enable co-creation and co-constitution in and of the future workplace? Questions of scale also arose in relation to the ways in which ethics can be operationalised; ethical and responsible modes of operation can be easily subsumed by drivers of growth and return on investment associated with the capitalist model. Perhaps ethics do not scale.

- **Drawing attention to and reshaping the power relations around technology.** There are clear imbalances currently, ranging from areas as diverse as the difficulty for researchers of gaining access to ‘proprietary’ data held by private companies, to the challenges of creating a level playing field in stakeholder discussions when citizens hold so much less power than technology firms.

- **Using questions such as ‘which benefits would arise’, and ‘for whom’,** not simply as rallying cries, but also the fundamental way in which we should organise our planning for the future. This also relates to the ‘societal’ and ‘GDP’ questions above, but viewed...
specifically as a means of possibly rebalancing some of the inequities currently visible in the system by moving away from ‘efficiency’ as a key measure. Efficiencies in the world of work generally operate in favour of management and shareholder groups rather than workers.

- **Professionalisation of software and programming-adjacent careers** was considered – especially in relation to notions of, for example, medical and legal ethics. Medical staff and legal professionals have codes of professional conduct and industry bodies to which they are answerable, which back up and reinforce consideration of ethics in their work. There is currently no societal consensus around the ethical status of software and technology workers. There are also questions of framings – for example medical professionals vow to ‘first do no harm’ – but ‘harm’ may not be the right word/concept here. There is therefore a need to define the parameters and consequences of possible action in order to create boundaries for ‘professional’ behaviour and related standards. This also raises questions around use of the term ‘responsibility’. Responsibility has the meaning of ‘liability’ as well as ‘care’ or responsiveness. In aviation there is responsibility without blame, which may be a useful model.

- The **future for scrutineer-type roles, such as journalistic and academic communities**. There are clear challenges for journalism in the current age, but as vital mediators and communicators, there is a responsibility to test hype, identify genuine (rather than spurious) issues and inform publics so that citizen populations are able to assess risks accurately. Academic communities have a responsibility to inform policy and engage with governance bodies in order to create evidence-based courses of action.
WOMEN IN STEM

It is well known that STEM disciplines have a diversity challenge. Women, minorities, people of colour – all are underrepresented in the STEM subjects. Homogeneity of background, education, training and experience leads to homogeneity of thought – this leads to gaps in knowledge that can be significant when developing technology that will affect the whole of society. Given this, the conference chose to focus its attention on the work being done by women in this particular field of AI/ethics. In particular, a women-first ethos was adopted that allowed for a greater range of voices to be heard than is usual at STEM-related conferences.

Partly to this end, the day was made as accessible as possible, welcoming babies and those who needed space for breastfeeding or expressing. All delegates were encouraged to communicate any particular needs they had, and much thought was given to ensuring that everyone felt they were in a safe and welcoming space.

In the interests of the women-first ethos, women were prioritised as speakers, workshop leaders and panel chairs. Male speakers were not excluded, but women were invited as a priority, in recognition that much of the pushback on the ‘move fast and break things’ trope is being led by women.

Delegates told us this created an atmosphere unlike any other tech-focused conference they had ever been to. Many made a point of reaching out after the event to comment on the welcoming, collegiate atmosphere, and the sense of community strength and fellowship. It was satisfying to have created an event so different from many others, and one where women felt appreciated and supported.

AI

The key theme of the event was AI. Artificial intelligence – or machine learning and big data – has recently become one of the most penetrative, ubiquitous and persistent technologies of our age. Its presence is everywhere, from GANs learning to ‘create’ human faces that have never existed, through facial recognition used by police forces to try and catch suspects on the street, to data-gathering programmes installed in children’s classrooms to monitor, record and predict their behaviour.

These technologies are proliferating far faster than regulation or enforcement can keep pace, and so it falls to those working in these fields to anticipate and reflect on the consequences that might be seen in society. The conference sought to provide a context for some of the issues that were chosen for the workshops – our keynote speakers gave their insights into current challenges and the issues raised were tackled collectively in the workshop sessions.
CARBON-NEUTRALITY

Evidence of the climate emergency, animal extinction, and impacts on planetary life is all around. It was therefore one of the earliest considerations of the 100+ Brilliant Women in AI & Ethics Conference that the event should be carbon-neutral.

Advice was sought from an expert group of climate-mitigation professionals, including Greenpeace scientists, whose most common response was advice not to hold a conference – the single most effective method of carbon mitigation is not to create that carbon at all. Offsetting is problematic, and not close to as effective as cutting carbon. Having committed to the conference with a strong rationale for bringing the 100+ women together in the same space, carbon-assessment company Carbon Footprint Ltd was commissioned to calculate the carbon that would be generated. Subsequently every decision was assessed on the basis of its potential for carbon-generation.

Every aspect of the conference and its impact on the environment was explored. For example, the printing company used was selected on the basis of its own carbon-mitigation efforts. The one chosen maintains a tree-planting programme to offset its carbon as well as providing numerous climate-friendly options in terms of materials. Some decisions were simple – the conference dinner was vegetarian with the food sourced as locally as possible. There were no conference give-aways due to the extraordinary waste generated by free items (which are often also plastic). Branding was kept to a minimum to create items that could be re-used for other activities.

After working to minimise the carbon footprint, Carbon Footprint measured the carbon that would be generated. Details were gathered during registration of where each attendee came from, their mode of travel, class of flight and other details. Lady Margaret Hall also supplied figures to calculate its own carbon footprint. The final figure calculated was 80 tonnes of carbon – a relatively low amount for an international conference. The largest single source was the flights of attendees – 26 flights generated 96% of the carbon.

Carbon Footprint was then commissioned to offset this 80 tonnes of carbon through a mix of rainforest-protection in the southern hemisphere and tree-planting in the northern hemisphere. Some of the carbon-offsetting activities chosen will involve schoolchildren, as it is our belief that an awareness of how to take care of the planet should be taught at the earliest possible age. That was also one of the rationales for running a schools panel at the conference.

The final result was having the conference certified as a carbon-neutral event, which makes it the first carbon-neutral conference that Oxford has hosted. Working to create a carbon-neutral event, however, has highlighted a clear dilemma. There is a clear need to stop excessive travelling – flying in particular – to help address the climate emergency, and of course academics are aware of this. However, as it essential that they collaborate to carry out and communicate their work, there is an urgent need for the research community to investigate new ways to address this issue.
METHODOLOGY

The conference aimed to produce some useful outputs that could be taken forward, and so a set of workshops was designed to ensure that all delegates had the opportunity to discuss and contribute their insights on the topics of the day. With so many attendees coming to the conference possessing enormous depths of expertise and breadth of knowledge, it seemed like an opportunity to ensure that as many voices were heard as possible.

All delegates therefore chose and attended a two-hour workshop split into two sessions. Each workshop was chaired by a leader in the particular topic being discussed. All four workshops followed a similar format in order to ensure consistency. Every workshop had a scribe in attendance to record and give their impressions of the session – these scribes’ reports are included later in this report.

Chairs were invited to structure their sessions around the questions that seemed most relevant to them but were also asked to consider that much of the current discourse focuses on the problems being caused in society. The aim was to build out from this and extend the debate – to move outwards from the current challenges and consider: ‘what might the future look like if we get what we want?’

The first session in the morning therefore sought to imagine how a positive future in this space might look. It centred on a horizon-scanning exercise in which participants discuss their responses to the question: ‘in the best of all possible worlds, where would we want to get to with this?’ i.e. what would the ideal future be? The aim was to encourage participants to think positively, creatively and ‘outside the box’. At this stage they did not need to be constrained by the practicalities of what might or and might not be possible.

This was a 60 minute session.

Session two in the afternoon brought together the same participants to extend the discussion. The aim of this session was to take some of the ideas formed in the morning and scope out a roadmap for how to get there. Participants once again discussed this in groups and were invited to consider what steps might be necessary to achieve some of the positive visions agreed in the morning session. Chairs were invited to select some of the ideas from the morning that they particularly wanted participants to focus on, or alternatively invite groups to select those which they felt most strongly about.

Session three was a full conference plenary session. In this, the Chairs of each workshop reported back on what was covered in their sessions. Each Chair had around 10 minutes to summarise the visions of the future identified in the morning and the pathways to the future identified in the afternoon.

After the conference all the sources of data were drawn together. These included:
• The talks delivered by the workshop chairs at the plenary sessions
• The reports written up by the scribes from each session
• Notes and posters created during the sessions by delegates

A data-analysis session was held with a small group of researchers to discuss and draw out themes from the materials and identify key issues to take forwards. All these sources were used to create this report and to provide themes for the call for papers.
THEMES FROM THE WORKSHOP

What is the question?

Work is a key element in society in terms of the connection it provides between humans, as well as the value of the work per se. It is therefore a vital component in making meaningful human activity, and thereby offers mechanisms both for human flourishing and appreciation of human value. It is also important to recognise that there are huge gender disparities in the world of work. The key question decided on for the workshop, therefore, was “How do we ensure that AI benefits all?” while using work and workplaces as the frame.

The starting point was that the concept ‘work’ helps us think about an ethical AI future, but generalities are unhelpful – there is a need to focus on practical questions. These practicalities may not have been considered, leaving serious gaps in the way AI ethics is discussed that omits labour and work.

What we also see are questions and challenges about impact in this area where the thinking is still in broad strokes rather than in terms of specific impacts.

Who benefits?

The portion of the question about AI benefiting all led the group to query the ‘all’. We need to question more strongly for whom – benefits for whom? Are benefits societal or at the social level? Or are they still framed in terms of individuals?

This leads on to questions of ‘values’ – whose values and what values, and what ‘is valued’.

It was noted that questions of efficiency are often uppermost in discussions and decisions about work – however if work is such a significant form of human connection, flourishing and value (as noted above), goals should not simply be efficiency or profit.

This drew out the question of whether late-stage capitalism is the best context for AI to emerge. The profit-motive as a driver for such influential technology may be perceived to be a poor choice when it comes to attempting to ensure beneficial outcomes.

What tools do we have?

1. New measures: if there is to be a serious conversation about human-centred AI, it’s crucial to think about the kinds of wellbeing measures that would be needed.
2. Solidarity between workers: for example, ‘ghost workers’ need to be brought back into the centre of how communities think about AI systems.
3. Collective action: for example, as can be seen where tech employees push back against decisions taken by their corporate bodies. This will take a different form than it did in the industrial age.
4. Co-design: bringing in stakeholders – not simply to listen, or educate, or inform, but actually to work together to build something new in a new way.
5. Education: education isn’t simply the responsibility of individuals to retrain and skill themselves. Education prepares workers and citizens for claiming their own roles, but also for companies to claim their responsibilities to citizens of the world. These questions were framed in terms of benefits, wellbeing, managing job changes and education – academia, policy, advocacy, and industry all have roles to play.

6. Citizens’ assemblies: stakeholder bodies that would be tasked to think about the impact of AI on the future of work, to consider roles that industry could be playing, to potentially be teaching, and developing new kinds of measures that allow companies to push themselves for higher accountability and responsibility. These would need to be matched with serious policy and regulatory teeth from government.

7. Analogies/narratives i) ‘dancing with the machine’ – co-constitution of how people will adapt, and how those adaptations will then be brought into further design. Neutral shaping and the co-constitution of society, this metaphor helps us consider how human and machine roles can be considered within society. ii) move away from thinking about work as a ladder through an organisation. Rather, think about workers today as on a climbing wall, going from hand-hold to hand-hold relatively precariously. So, we can refocus on the harnesses, the safety nets that are holding workers together as they make this pathway. iii) the need to work together on creating better narratives of what possible futures could look like.

ADDITIONAL POINTS FROM DATA SESSIONS

The question of what we measure and how we measure it is key – what does not get measured does not get valued.

Efficiencies – doing things faster at lower cost shouldn’t necessarily be the only goals.

Ensuring all players and actors are recognised, drawing back on those who don’t currently have a voice and creating truly representative solutions.

Key points: the distribution of benefits, worker well-being, the evolution of work and jobs and the challenge of factoring in intersectional groups to that change, and the co-design of systems that are intelligible not just intelligent.
How will work look different in a world that is increasingly automated? How can people find value or meaning in a world where things we traditionally consider work might no longer be done by humans? The first session invited participants to throw out broad, challenging questions to shape our understanding of the problem we wanted to tackle.

Using the workplace as a lens, there was immediately a focus on ‘levelling the playing field’, keeping intersectionality and diversity at the forefront of our discussion. We talked about existing gender and racial disparities in the workplace and how the introduction of AI could worsen them, as well as the different understandings of workplace culture that exist across the globe and what new technologies would mean to each of them. However, when the problem of fair redistribution of the benefits of AI in the workplace was raised, we again hit the roadblock of lack of concrete definitions. Naturally, financial benefits were considered, but also intangible ones, such as worker wellbeing. This engendered a discussion about the perspective from which we should view the problem: is it a societal or individual responsibility to manage the impact of revolutionary technology? Is there even an opt-out scenario for working with intelligent and automated systems? How can we suggest societal level solutions to problems of personal well-being? The group came to the collective conclusion that it was important to keep both corporate and private contexts in mind.

The subject of education was particularly contentious. The group was immediately mindful of placing too much responsibility on individuals to re-skill, and that corporate and government efforts needed to be made to ensure all people had the ability to engage with AI on some level.

At the end of the first session, we had established four main themes around which to build our roadmap: the distribution of benefits, worker well-being, the evolution of work and jobs and the challenge of factoring in intersectional groups to that change, and the co-design of systems that are intelligible not just intelligent. Even with the focus of our discussion narrowed, these themes presented a challenge in their enormity and complexity. The first session had made it clear that stakeholder groups would have different concerns and priorities regarding AI and work. Additionally, they have different roles in establishing norms within society that could be utilised to solve the issues we envisaged as inevitable within this framework. Consequently, during the second workshop session we sectioned off into smaller groups, each taking on the role of a significant stakeholder, to discuss our four key themes: academia, industry, policy makers, and civil society.

Unsurprisingly, education fell under the umbrella of academia, also. Japan and Germany were suggested as case-studies upon which to base a system for universities to organize re-skilling programs for people who aren’t studying. Once again, the emphasis was placed on remaining human-centric, reiterating the theme of distributing the benefits of AI by providing everyone the opportunity to actively engage with the new technology.

Of course, the academic community alone cannot be expected to shoulder the responsibility of educating and upskilling an entire workforce. The group representing industry stakeholders acknowledged that there was some responsibility for those members to establish a pipeline for knowledge sharing with the wider community. In particular, they examined the top-down approach to ‘giving-back’ to society, of which academic-industry partnerships would be a central component. We discussed existing examples of industry heavyweights who spend part of the work week sharing their expertise through university or school programs. It was suggested that certain jobs could have a mandated ‘outreach’ responsibility.

In an ideal future, industries would push themselves to achieve higher accountability. On the intersection of AI and ethics, it was queried whether it would be possible to establish an accreditation scheme, taking holistic view of a company and providing an assessment...
Scribe’s report

of its ethics. While most members of the workshop were enthusiastic about the idea in theory, there was scepticism about the practicalities, given the challenges associated with understanding causality in the context of intelligent systems. Subsequently, one of the most interesting discussions regarding industry centred on the changes to the priorities of the future workforce. It was accepted that a millennial workforce was focused not only on profit, but also finding work that gave purpose to an individual. Perhaps the changes to industry would be demanded by the workforce, rather than solely the technology.

Policy makers also have a responsibility to support the development of both foundational and advanced skills in AI, through further investment in research programs, rethinking what foundational skills are needed and changing curricula accordingly. On this matter, the policy group stressed the importance of teaching non-technical skills, such as critical thinking, that will be vital in our imagined future. In terms of job creation and reskilling initiatives, the group suggested redesigning the way in which governments help job seekers, by facilitating more paths to move between industries.

We discussed existing legislation in terms of worker well-being and the need to ensure that policy remained up-to-date with the altering work environment, so that no one would slip through the gaps. This may mean re-envisioning what the social contract of work means to a future society, and taking lessons from what others have done.

We closed the second session with the role of civil society (non-governmental regulators, non-profit organisations, grassroots movements and unions, and citizens). We saw these groups acting as the facilitators of trans-sectoral partnerships, as participants in research, lobbyists on behalf of workers and job-seekers, and coordinators, providing mechanisms like online community forums, workshops and public events. Unions were highlighted as particularly important to the conversation, as they could act as true representatives of the workforce.

In talking about work, we found that the subject of education arose repeatedly and we wanted to be clear that education should not be considered a panacea that can be handed to individuals, as their responsibility to retrain and get more skills. Rather, education should be thought of as a means to not only prepare citizens for their role in the coming world, but also as preparation for companies and organisations.

The second metaphor homed in on the fears that are prevalent within discussion of automation and the workplace. We endeavoured to do away with the ‘ladder of work’ metaphor in organisations and rather think about a ‘climbing wall’. Even today, a typical career path is unlikely to be a simple bottom-to-top journey. Workers move on the ‘climbing wall’, going from handhold to handhold relatively precariously. The consequence of changing the imagery used in these sorts of discussions is that we focus more on establishing harnesses and safety nets within workplaces and, more broadly, the world of work, to hold workers up as they make their path.
AI & GLOBAL GOVERNANCE

THEMES FROM THE WORKSHOP

How could global governance be structured?

There is a requirement for different levels of granularity of governance dependent upon the 'layer' of social/community that is being addressed.

1. Incentive structures: one action for good global governance is incentive structures – for executives, companies, policy makers, what does it look like to have an incentive structure that holds them accountable and responsible? The UK has some regulation around accountability for executives since the 2008 financial crisis – potentially a model for building an incentive structure for accountable and responsible AI.

2. Standards: In sustainability, for example, there are standards from the Sustainability Accounting Standards Board or the UNPRI. These are environmental, social and governance factors for how to measure companies’ performance against given standards. Potentially this kind of system could be a model for how to think about incentives and disincentives for companies. Measurable performance against agreed standards and metrics would also be relevant for investors looking at these companies.

3. Regulation and law: it is difficult to create regulation without sufficient knowledge of an industry; enough data; or a workable model of a constructive relationship between industry and government.

4. Certification: for example, the ICCC in climate change and the IEEE working on certification for AI. A ‘regulatory sandbox’ has been used in innovation in a couple of countries, including the UK, and could be worth investigating.

5. Community ethos: In a small enough community – for example the open source community – it is possible to instil community standards.

Potential roadblocks

1. The ‘race’ for AI pushes both countries and companies to hype their discoveries, push untested or unready models out, and fail to take the time to consult with stakeholders. Stakeholders may be the same individuals from a country or company perspective (for example, voters from a country perspective are also customers or shareholders from a company perspective) so they do have power when it comes to where they spend and where they vote, but this goes unrecognised.

2. A lack of understanding about what is ‘good’ in good global governance. How can ‘good’ be defined? For whom is it good? There are many questions to be considered and many potential trade-offs such as, between the individual and the community, or between majorities and minorities – for example, should concerns err on the side of marginalised groups? There are also challenges of scale here such as, regional level ‘good’ versus global ‘good’.

Chair
Bulbul Gupta
3. A lack of agreement as to where accountability and responsibility lie, and at what level. Does this responsibility lie with corporate executives, developers, policy makers, or at a societal level?

4. Self-regulation has been the norm to date. However self-regulation by companies is not the answer when societal ‘good’ is at stake.

5. Human rights standards. It was questioned as to whether these are good enough to move forward with good global AI governance or whether an update is needed based on the changes wrought by emerging technology.

**ADDITIONAL POINTS FROM DATA SESSIONS**

It is vital to recognise that stakeholders are approaching these questions from different viewpoints and to recognise that there will be frictions – there is a need to seek ways of working together that take account of this.

The speed of innovation, as well as its direction, is problematic. It is extremely difficult to do complex work like governance or regulation fast and well. Usually governance is either slow and good or fast and bad.

Any investigation of regulatory frameworks or governance would have to take into account the global nature of these systems and the ‘wicked’ problem of any globalised solutions.

Relationships are vital here – relationships between companies and their shareholders, within various communities. Strong relationships generate trust and trustworthiness. This can be more difficult where there are asymmetries of power.

It is easy to talk about ‘society’ and too easy to think of it as one homogenous mass – in reality there are multiple societies and many layers – different solutions will be required at different levels, depending on the granularity adopted.

Anticipatory governance can be regarded as preparedness – plans are not made to be adhered to, but to provide resilience and broad templates for how to address particular challenges.
The workshop began with an outline of the main questions for the day: what does the future of global governance look like if we get what we want as a community; and when we think about the innovation we want to have and promote, what are the bright lines we think are required or red lines we don't want to cross. In the first session, participants discussed questions around what constitutes ‘good’ global governance for AI, how to define and measure its core components, as well as what effective models of global governance for AI might look like. Groups discussed individual questions and reported back to the workshop as a whole, leading to wider discussion of the points raised.

While deciding what is ‘good’ in global governance presented little challenge (e.g. fairness, welfare, equity, and equality), the extent to which these ideals should be traded off when balancing needs of individuals or marginalised groups against the wider population was recognised to be highly contextual. Overall the discussion highlighted the tension between creating a global forum to bring organisations and institutions together, and the need to consider differences in goals and customs at the national, regional, and individual levels. On a more positive note, the commonalities that do exist between countries and industries are something that we can foreground and build on. However, the developing nature of artificial intelligence means that these tipping points need to be revisited as the risks of new technologies became apparent over time. There was also a sentiment that models should be principles-based in order to keep ahead of rapidly developing technologies and prevent the stifling of innovation by being too prescriptive.

A recurring theme within this discussion was the importance of autonomy, both individual and collective, as a core topic of AI ethics. The scale and pace of development is greater than ever before, and this is something we need to respond to by increasing efforts towards education and awareness beyond the core governance stakeholders. When stakeholders are engaged, this needs to be done in a practical way in order to reach a decision making consensus. Education and general understanding for the general public was also seen as key, not least because citizen juries will be increasingly required to make judgements about AI and algorithms.

There were widespread concerns about large technology firms being the loudest voice when it came to governance of AI, and too powerful for self-regulation to be effective (implying that they should be government led). Participants agreed that we need to come together as a community to be big enough, taking inspiration from changes to banking standards post 2008, and how the introduction of the GDPR has prompted many companies to adopt compliance in jurisdictions outside of the EU. Looking for ways to motivate compliance, examples were given of the UK introducing criminal liability at the individual level for directors, and the ability of German companies to sue each other for breaching the law; if companies could sue other companies for gaining an advantage by flouting national laws, then perhaps governments could sue other governments for breaching global AI agreements. This top down approach was seen as necessary given that those on the front lines (such as engineers) are rarely incentivised to optimise for ethics, particularly with respect to the big picture.

In the afternoon session, participants revisited these challenges with a view to determining the practical steps required to move forward with the morning’s discussions, as well as the risks and opportunities that these might bring with them. The session began with a time for individual contributions that were then discussed in groups, and then by the workshop as a whole. After the initial steps needed to identify and bring together stakeholders from the public, legislative, private, citizens, academia spheres, the discussion turned to the possibility of creating a global regulator or similar intergovernmental organisation that would help to keep governments
accountable. While metrics (similar to the UN Sustainable Development Goals) and corresponding enforcement mechanisms were seen as important, they brought with them the challenge of deciding what those metrics should be. Regulatory sandboxes were seen as another potential way to begin this process, and the possibility of leveraging the shared ethos of existing open source communities was seen as an opportunity. The discussion also prompted a question over the extent to which existing international agreements (such as those governing basic human rights) would need to be updated to cover developments in artificial intelligence.

The creation of a set of guiding principles or standards was seen as key, as these can often lead to (and form the basis of) subsequent legislation. A representative from the IEEE standards association's group of P7000 standards described the progress of current efforts on a variety of areas relevant to the discussion, including the addressing of ethical concerns during the design process, transparency of automated systems, algorithmic bias, and the use of facial analysis technology. These were seen as giving nation states a solid foundation on which to lead their own governance efforts. One key point raised with reference to the rollout of the GDPR was that any efforts to regulate AI would need to give companies clear routes to compliance.

In terms of the risks of following these approaches, some participants questioned the practicality of making guidance precise enough, especially given the chicken and egg nature of creating ethical frameworks, where it can be difficult to evaluate use cases until new technology is designed and understood. Another risk identified was that agreements on AI development would not be followed, especially owing to the dual use nature of AI as a tool that could be used for civilian and military applications. The option of moratoriums was broached for types or uses of AI that were widely regarded to be beyond the pale. Another approach, drawing the comparison with conventional weapons, was of the separate regulation of cyber “arms” (potentially including acts like interference in the 2016 US presidential election that skirt current definitions of warfare).

To conclude, the workshop tackled a number of difficult problems with respect to AI and global governance. From what is good in global governance, and for whom, to questions around promoting accountability through incentive structures (carrots) and regulatory frameworks (sticks). Examples from UK and German legislation, regulatory sandboxes, and the UN were identified as promising approaches to emulate, and current work by IEEE in standardising terminology and procedures around algorithmic fairness, accountability, and transparency will be worth watching as they develop.
THEMES FROM THE WORKSHOP

Who and what are within the scope of our concerns?

1. Who is the ‘everybody’ when it comes to society? A great deal is at stake when operating with a majoritarian model, which is currently the dominant discourse. But is there actually some type of collective we and everybody for which there is social good that might emanate from AI?

2. Can algorithms help humans to make better, fairer decisions?

3. What are the moral and ethical frameworks from which those decisions would emanate? – this is an area where there is no clarity. Different groups have different expectations and assumptions.

Questions arising

1. Areas of effect for AI and algorithms – are there areas of decision-making and human existence that should be ‘no-go’ regions for algorithms?

2. Transparency – society should be clear on whether AI is an extension of human decisions – a support mechanism – or a replacement?

3. Efficiencies – efficiency tends to be selected for in creating value for shareholders or management, which may result in trade-offs to quality of life.

4. Narratives are important – in order for non-specialists and publics to grapple with the challenges in these areas, there is a need to frame the problems in such a way as to be properly understood. Narrative and analogy (for example the ongoing discussion about whether data is ‘the new oil’) are important sociological tools in enabling more accurate and fruitful discourse.

5. Workers’ rights – there is a lacuna in the discourse, given that AI is made by workers, and is about workers. Citizens need to understand that they are not solely consumers, but also labour. Labour is not frequently analysed with regard to AI. The question needs to be asked whether workers should unionise, should they be at the core of decisions with respect to AI?

Keywords from the discussions

human-ness and human life, accountability, power, inclusion, agency, freedom, context, human value, emotions, creativity, culpability, transparency, quality of life, sustainability, pleasure, wisdom, environment, reflection, choice, feedback

Nuance

The debate needs to be driven towards more nuance and specificity, rather than generalities, and to demystify AI so that genuine questions and challenges can be discerned. There is not currently enough nuanced conversation that also understands people are coming at many of these questions from different starting points – academia, industry, government.
Data

Vast quantities of data are controlled by companies and governments. This means that very often researchers and institutions encounter significant difficulty in accessing the data. There are clearly asymmetrical power relations in this space – it is extremely hard to make progress when the only people who have the data are private companies. There are issues around trade secrets and proprietary IP etc that make it extremely difficult for researchers to have any kind of power in trying to access these datasets.

Growth vs flourishing

There is a difference between growth and flourishing – the growth of AI is generally discussed in terms of GDP, but not in terms of the quality of life for human beings, or other measures such as planetary health. Altering measures away from ‘growth’ metrics would be a powerful change in the public discourse.

Capitalism

The current global model puts AI in service of extraction rather than in service of, for example, ending oppression. There is little attention paid to the question of ‘for whom’ is this technology in operation? This is also part of the asymmetrical power relationships when considering the role of AI and algorithms in society.

Accessibility

There is a convenient tech myth that users can be in control of their own data, their own levels of privacy, their own safety in the digital world. This is clearly unrealistic, not least because these areas are frequently filled with jargon, with technicality, and when even developers may not know what they’re working on. There’s a handy expectation that somehow the public is going to have a high level of digital literacy. This can make discussions highly inaccessible.

Narratives

Articulations about what technology is and what it could be are frequently in the hands of those with the deepest pockets – realistically this means tech companies. These companies and, frequently, their large professional partners, are able to speak powerfully about the future they want to see and they have the resources to broadcast that message globally. Those whose focus is on studying the harms do not have the resources to tell the story in a way that can compete with this. This is nothing less than a battle of ideas – many of which are incredibly political – but the asymmetry of power is also political.

ADDITIONAL POINTS FROM DATA SESSIONS

Metrics are key – how is ‘good’ defined and how can that be measured. GDP is inadequate in many ways and also unsustainable as the ‘growth’ model cannot continue.

The power relationships between tech giants and everybody else are so wildly skewed that they can probably never be readjusted in a way that could compensate for the current asymmetries.

The proprietary ownership of enormously powerful and influential algorithms is directly linked to the economic model we use in the West – how can this be addressed.

There is huge difficulty not just in bringing all stakeholders to the table but also ensuring that they are all participating in good faith, with similar values and from similar positions of understanding.

Relationships are key – people exist in their own ‘ecologies’ but need to understand alternative points of view in order to make progress collectively and collaboratively.

1 Cf Diane Coyle GDP: A Brief But Affectionate History
As is perhaps inevitable with broad questions, significant parts of the first session were dedicated to deciphering the question itself. Who do we – as participants in this workshop – represent when we ask questions about the future of algorithms and society? Are we totally inclusive? Do we include harmful people and their viewpoints?

This reflects an old debate: to what extent do we tolerate the intolerant? Is an ideal world one in which such conflicts no longer exist? The participants in Group A concluded that an ideal world would likely need a forum for settling disagreements, which included diversity of thought. Participants noted that a lack of diversity among the majority of technology designers is reflected in many science fiction dystopias and utopias, which either reduce the universe to Western cultural values or allow enormous diversity through assuming everyone has the same privilege, which reduces diversity to aesthetic. "Real diversity"—of thought and of experience—would be a crucial value in an ideal world.

Many participants agreed they wanted humans to remain morally culpable in algorithmic decision-making. However, as one group pointed out, individual accountability may be a wonderful thing, but is it a western value rather than a universally shared one? Should companies, researchers or governments be in charge of determining who is responsible for rectifying and compensating for mistakes? Group A argued that the government should ultimately be responsible for adjudicating such conflicts, while Group B emphasized a close relationship between developers and users, so that two-way communication allows for user needs to be met.

Group C focused on the potential benefits algorithms could bring to the areas of disability and care. One participant related a conversation with a man who asked: "Why are there smart cars but not smart wheelchairs?" Artificial intelligence could help with social integration, sense support, communication enabling, and companionship.

Group F pointed out that different groups might have different preferences for the scope of algorithmic decision-making. For example, people of different ages are familiar and fluent with different technologies. In fact, technology is increasingly a defining feature in differences between generations.

Another shared concern related to companionship. For example, doctors provide pastoral and emotional care which algorithms cannot offer. We should be careful not to take that away. However, artificial intelligence (AI) could also offer companionship; for example, some technologies attempt to recreate the voice of deceased partners or loved ones. Participants discussed the importance of "human-ness": we need to keep hold of some values and activities which are essential to human life.

Lastly, a central concern was the question of control: as participants in Group E put it, "algorithms should augment and enhance but not take over." In the medical context, this might mean supporting a doctor’s diagnosis but not dictating it.
THEMES FROM THE WORKSHOP

Challenges

1. *Who is the 'we'?* In discussions of algorithms in society, ‘society’ is being used in a way that presupposes much consensus around what precisely is meant. The world has many societies and many communities that AI is affecting in different ways. Assumptions about whose values apply – even in whose language the narrative of AI is told – will inevitably affect the conclusions drawn. It is difficult to see how to shape an ethical future for algorithms in society when there are irreducibly many plural visions of the ‘good’ society.

2. *How can the future of AI be freed from a customer-driven model of development in which every stakeholder is seen as a customer or potential customer – therefore creating a vision of a future that’s constrained by present day economic forces and incentives?* Current economic models drive development down paths that make it difficult to envision a wider scope of possibilities.

3. *Enriching the model?* There is a need to broaden measures of success to include richer measures of social success – this is necessary to counter the narrow quantitative metrics so frequently utilised.

Destinations

1. Societies that are structured by more ethical power relations. New distributions of power are going to be needed that address some of the current imbalances and asymmetries of power. This might include algorithms that enable and are shaped by goals of liberation, equality, justice, opportunity, solidarity, social mobility and reflexivity, rather than goals, or norms of dominance and exclusion.

2. Societies strengthened by ethical cultures, enriched cultural values and institutions, algorithms that support the development and flourishing of humans and non-humans and that provide social and environmental resilience for a future driven by humane values. Values that aren’t limited to quantitative or consumerist frames – values of solidarity, cooperation and community. These values are under-emphasised in the discourse around algorithms and our institutions in favour of values that prioritise individual protection, and individual justice and fairness. We need to foreground these more community-driven values.

3. Societies supported by trust and responsibility in AI development – more of the burdens and costs and responsibility for social protection being shared by and shifted to AI developers who can be more responsible and trustworthy, who can develop tech for the public sector. Society needs greater integrity, diversity and regulation of data, expanded access and opportunity in tech, and in the fields opened up by tech. Tech that’s designed to be supportive and unobtrusive, running in the background of society rather than in the foreground. And societies that are more resilient to machine error, but also able to appropriately trust and receive algorithmic benefits.
Routes/roadmaps

1. Society structured by ethical power relations – political and technological reforms are needed that will help fulfill the existing social contract. Existing social contracts are not currently being fulfilled by the ways that algorithms are being developed and designed and implemented. Therefore, there is a need to identify how to fulfill the existing social contract and then iterate and improve upon it.
   a. Influence the tech community to serve the broader community and own their responsibility – for their power, for the bias in our systems.
   b. Through mechanisms and consensus realign the share value, the dollar value of algorithmic products and services, with their value to the community and their value to workers.
   c. Think of open data and related instruments to reassign the value and power of data back to its source – the human communities from which that data is generated. And then to enforce the social contract through more effective regulations and liability structures.
   d. Algorithms that serve and are governed by communities that represent the interests and agency of human beings, and not the tech itself. Re-centring citizenship as the nexus of power through reforms in education and diversity practices.

2. More ethical cultures. This would see algorithms enabling and being embedded in higher and richer human modes of consciousness and valuing rather than impoverished consciousness of value and possibility. This was viewed through six institutions, since culture is driven by institutions.
   a. justice and security,
   b. social welfare,
   c. infrastructure,
   d. education,
   e. the economy
   f. the domain of human spirit and meaning, which includes art and all other forms of meaning making.

Specific interventions may be needed in order to shift the measures of health and success of current institutions. For example, in education, this could mean focusing less on optimising for individual performance, and more on goals like shared learning and collaboration, and the joy of learning, which is increasingly being lost in educational institutions as a result of algorithmic assessment. In the economy, this might mean measures like opportunity, and wellbeing and sustainability being used – a model beyond GDP – as measures of economic growth and success. In particular, the most effective interventions were considered to be:
   a. better metrics, which law and policy can generate, that describe and assess those goods that are not easily counted.
   b. more activism and public engagement.
   c. new modes of storytelling and moral narrative, in art, education and other cultural institutions. One of the most powerful tools that society has is narrative and storytelling in order to shape a future with AI and with algorithms that preserves those higher modes of consciousness of which the human animal is capable.
3. A culture of responsibility in AI and algorithmic development. Four core elements were identified here:

a. More diverse inputs to AI and algorithmic development. Diverse data origin and usage, better representation and tech, diverse methods, goals and visions of the future.

b. Better forums for reception of these tools, more support for new institutions for education about and discussion of algorithms, more transparency in design, development, implementation, and better auditing of the track record of specific algorithms.

c. The need for improved fora for critique of algorithms, for the cultivation of appropriate and balanced distrust and room to acknowledge failure in algorithms and improve upon them.

d. Better structures for responsiveness to critique of algorithms and tech, more structures of explainability, value alignment. New structures of distributed responsibility for these algorithms and their impact, because mapping responsibilities on to particular actors is increasingly difficult.

**ADDITIONAL POINTS FROM DATA SESSIONS**

Some of the challenges here are ones of scale. A search engine company might begin with a clear ethos to do things differently from predecessors, and may start out that way. But after several rounds of investment, the need to make returns on that investment, and huge growth, the company becomes unrecognisable from the way it started out as. Perhaps ethics do not scale.

Relationships can also be difficult to scale – a small company that works with its customers who also work with each other, can be a community. Past a certain size-point, this becomes increasingly difficult to sustain. Growing a business can be a significant driver of ‘value drift’. This may particularly apply where initial management ethos can be diluted by investors who potentially have alternative views of how the business can/should develop.

The industrialised model creates efficiencies, in the way that Henry Ford did – breaking tasks into smaller pieces – and this has enabled high-speed technological development. It also creates a many-hands problem. But this hasn’t been matched by a collective responsibility paradigm, such as that found in the aviation industry. Efficiency may also not be the best overall goal for a society, creating a misalignment with economic models that drive development in the direction of ever-greater efficiencies and profit growth.

Engineers who write code are generally at an operational level and have very little say in how their work is used. Is it reasonable to hold them responsible?
We were tasked to envision a positive future society that uses algorithms, and how to get there. We started with many challenging questions: in the best of all possible worlds, what rights, values, principles, laws and norms would guide the development and use of algorithms within society? How would human agency, as well as characters of people, institutions, and societies, be impacted? What harms should we prevent, minimise, and avoid, and what alternatives to the current societal patterns would we find?

We agreed that not everything needs to have a monetary value, and that regulation is key – there should be some sort of accountability to the developers of systems, something that we have not seen enough in today’s society.

As individuals in societies, we are shaped by our technological environments. One participant mentioned that to envision a society of the future, we should allow ourselves to be guided by how we want our children’s environments to be shaped by the surrounding technological systems. We want algorithms to help facilitate learning for all, but especially our children.

The current development of and conversations surrounding technologies have been largely centred on harms and benefits to individuals (e.g. ‘the user’); we need to engender a shift or expansion of concern to considering their effects on groups, institutions, and social structures.

When discussing the issue of algorithms and society, the primary question arises: whose society? Algorithms have global reach, but visions of societies vary – even today, accents and languages aren’t well represented in technological platforms, whilst algorithmic outputs are clouded by gender and racial biases. It is said that algorithms should free people of menial tasks, but for who, and for whose benefit?

The norms and values between different cultures and ideologies vary massively, hence this can lead into a majoritarian approach, with some participants agreeing that this seems the way forward, and others voicing concerns. Regardless, there was a consensus on equality and diversity being the major guiding values in such production of algorithms; further, in this imagined society, forms of human agencies including liberation, opportunity, and autonomy should be enhanced and not diminished.

We posited that there should be a new distribution of power. Compared to today, power distribution should be shaped by equality, justice, liberation, opportunity, social mobility, and reflexivity, as opposed to domination and exclusion. This power should be constrained by regulations, and those who do have power should not be able to disown their responsibilities. The relationship between trust and power should be studied, while questioning the appropriate relationships of trust and responsibility are possible in our new trust network.

Lastly, the most important point for our envisioned society: algorithms should function unobtrusively in the background as a support, not as a central pillar of society. This is not to imply that such technologies are closed off from public inspection, or immune from accountability or social critique – AI should be unobtrusive, but transparent.

On Power

To achieve the distribution of power we envisioned, we needed to imagine power shifts in society. An intuitive start point would be to look at the tech communities today and their influence, and how their aim should be to serve and influence the human communities around them. We want there to be engagement and trust. Onwards, the consensus product should be determined by communal values -- which includes the employees -- aside from the share value. The data obtained by the products should be open, giving availability and access to the communities.

Regulations should be guiding communities’ decisions, alongside social contract – keeping humans in the loop is key. This might mean a human that has the power to stop the system altogether, and that perhaps no machine that is ‘too intelligent’ for human oversight should be deployed. With that, liability and
accountability come naturally. Above all comes diversity, education, and citizenship which should follow.

**On Trust & Responsibility**

A culture of trust and responsibility should start within organisations. When we are focused on the use of algorithms in a society, as our previous discussion on trust, it is intuitive to focus on those organisations which designs, distributes, and develops AI algorithms. To such, we can apply a social epistemology framework.

In such organisations, there needs to be a forum for critiques of ideas, along with an infrastructure to allow a systemic response to such critiques. In such a response, there needs to be the means for explanations – we must also consider the mechanism for explanations - what role do they actually play?

A possible answer is simple: explanations play a role in informed decision making, which increases the likelihood of trust between the users and the systems. With such a platform for critiques and responses, the agency and autonomy of users are respected.

**On Ethical Culture**

An ethical culture is one which the human consciousness is raised, and the human intelligence enriched, enlarged, and deepened. To envision how such culture can be achieved, we looked at different institutions, and which values we want these institutions to uphold. We limited our discussion to six institutions: education, economy, social welfare, justice and security, human spirit and meaning, and infrastructure.

On education, the current use of technology surrounds the student as an individual and is used primarily as a method of assessment. In our imagined society, the use of algorithms will shift from an individual level to a community, by facilitating shared learning and collaboration whilst bringing enjoyment into the process of learning.

On the economy, society needs to go beyond GDP and measure success towards more on well-being, opportunity as different economic agents, and sustainability. The social and political health, as part of social welfare, should be driven by solidarity, autonomy, civic friendship, and trust. Mobility, freedom, accountability, and fairness should be the building blocks of justice & security institutions, and human spirit & meaning should maintain optimism, creativity, compassion and benevolence, as well as tolerance and inclusion of other cultures.

Lastly, to accommodate all this, the infrastructure of systems should prioritise resilience, sustainability, flexibility, and transparency.

To accommodate these institutional values, there should be an emphasis of human values beyond the current favoured quantitative metrics. With such, there is a challenge to create a new mechanism for such shift: firstly, a society needs to reshape its narratives by new designs of storytelling, which may include literature, art, and education. Secondly, new modes of law and policy can help determine what we assess in terms of success and growth – this brings another challenge, in which the words that are used today in law and policy are quantitative-based, such as ‘count’ and ‘measure’. Lastly, activism and public engagement should be encouraged, so that members of the society outside of the public policy sphere is able to contribute.
THEMES FROM THE WORKSHOP

Case studies

Case studies on smart speakers and voice-assistants; deep-fakes; and personalised insurance were used as a basis for discussion. This included discussions on algorithms and business models, services, the internet of things, the advantages and opportunities of these technologies versus the disadvantages, downsides, and threats.

Forms of regulation and governance were also discussed in terms of regulating and/or governing them. This might include law, ethics, norms, code, software, codes of conduct in an effort to maximise the upside and minimise the downside from viewpoints both of societal value, and economic and business value.

Points from case study discussions

1. Despite the difficulty of some technologies having more downsides/threats than upsides – for example, deep-fakes which were agreed to be egregious – there was reluctance to agree that some technologies therefore should be banned. Discussion focused on whether there may be a tipping point where a technology is so inherently toxic, that society decides not to pursue it. Can decisions like this be left to the market?

2. Timing of any kind of regulation or governance is difficult – for example, whether you act ex post or ex ante. In the example of deep-fakes, enforcing any kind of sanction against them after the fact will be very difficult in a globalised universe, so might there be a way to filter them out in advance, to do something proactive? This generated discussion of what the responsibility of platforms should be.

a. Regulation: There was discussion around the UK’s Online Harms White Paper. Using regulation to push more responsibility on to platforms such as Facebook and Google would place them in the position of being moral and ethical guardians or censors. This model would have clear difficulties – for example they might disproportionately favour certain types of content, which could introduce discrimination against already vulnerable groups. They would likely also favour material that they can most easily data-mine to produce the most revenue from the most targeted adverts. Current content regulation systems are already not working,\(^1\) so it is likely that this would simply transfer over to new forms of AI products and services as well.

b. Ethics: The insurance industry was offered as a current example of effective self-regulation driven by ethical concerns. Insurance companies already are willing to discount premiums if customers wear a Fitbit or have a device in the car that measures how safely it is being driven. They have undertaken not to drive up the cost of insurance for people who do not provide that data. However, there is no guarantee that this will continue into the future. So, is there a tipping point in this space where regulation needs to replace ethics – a point at which even effective

---

\(^1\) Sarah T Roberts, *Behind the Screen*
self-regulation stops being appropriate/sufficient? Can this point be identified in advance? It was also noted that the insurance companies seem to be using a concept of solidarity here to spread risk, and might that be derailed by big data, analytics, and hyper-personalisation?

c. The role of code and design. There was discussion about the role of the software employee or designer. When people join a software company, they are exposed to training about their values – if they follow company rules then can they be regarded as acting unethically? It is clear that there could be conflict between their job and being asked to produce products that society may view in certain contexts of deployment as unethical – for example, facial recognition. Thus, there needs to be a debate about how these conflicts – whether or not they result in whistleblowing – are addressed. There is also a lack of societal consensus about the ethical status of the software profession: we have a notion of medical ethics, we have a notion of legal ethics; we know what you can be struck off or disbarred for. It is not at all clear what that is in different sectors of the software industry.

3. Media responsibility. Currently deep-fakes are getting a great deal of media attention, but perhaps this disguises material that might be far more important – for example, the choice of different fairness metrics in the recidivism system. So there needs to be discussion about responsible journalism – Columbia Business Review has suggested that journalists should carefully consider reporting, for example, deep-fake stories which are seen as low-hanging fruit, because such stories can make people scared of things that they shouldn’t be scared of and less scared of things that they should.

ADDITIONAL POINTS FROM DATA SESSIONS

Is ‘harm’ the right word/concept here – in medicine, harm has a very specific meaning, but when applied to problems caused by tech it is unclear.

Professionalisation of software engineering might be one way to address some of the issues we have seen in these spaces. Difficult to do, potentially, but worth investigating.

Is ‘responsibility’ the right term – it has the meaning of ‘liability’ as well as ‘care’ or responsiveness. In aviation there is responsibility with no blame – perhaps a useful model.

Regulation is not keeping pace – for example some large corporates use algorithms to decide who to sack, and thus avoid employment law protections.

There was a reluctance to ban ‘deep-fakes’ but it’s important to distinguish between the technology and its use. The technology may still be useful even if the use-case is egregious.

Academics can be conflicted when it comes to making categorical statements against something – not only might they be risking their career but also an opportunity to work with someone in the future.

Collective action may be key here, whether formally (for example, via unionisation) or informal (for example consumer boycotts) – solidarity2 and recognition that an issue is not irrelevant simply because it doesn’t affect you.

Journalists have a major responsibility here to ignore easy clickbait stories and report the issues that are genuinely of significance.

Divided into two parts, the first session put attendees into groups and provided discussion prompts from which to create a Strengths/Weaknesses/Opportunities/Threats (SWOT) analysis. In the second session, attendees used the earlier discussion to explore and challenge the ideas they had collaborated on as well as their own individual experiences and opinions that emerged.

The session chair welcomed attendees and reiterated that the day centred on working towards a future where artificial intelligence (AI) "works for us". The chair wanted attendees to explore the modalities that could be used to make that happen, such as public education, regulation, international norms and standards, technology standards, design practice and the instillation of ethics into techno-design culture. To prompt these discussions, "Snapshot" briefings from the Centre for Data Ethics and Innovation (CDEI) were provided as cases studies. The case studies provided were "Smart Speakers and Voice Assistants", "Deep-fakes and Audiovisual Disinformation" and "AI and Insurance".

Case studies were used to reduce abstract theorising: the session chair wanted attendees to create a SWOT analysis in groups, focusing on motivating factors as to why these technologies should be engaged with (or not). By the end of the first session, all groups had mapped common ideas and concerns using the SWOT analysis framework (see appendix 2 for an overview of common concerns). The second session built on these ideas, exploring how advantages could be maximised and disadvantages minimised.

**Challenges**

Where are smart speakers getting their information? Empathy? Are we becoming less empathic? Does this threaten even more of an "echo chamber"? Is there a clear process to withdraw consent? Lack of info and data. Public figures v. individuals’ personal censorship. Speed of response and who. Intervention v. in-built systems. Business focus is profitability. Who defines ethics? What is ethical, and for whom? How do we monitor compliance?

**Key points**

The role of platform regulation and governance arose. A recent paper on online harms released by the UK Government (DCMS, 2019) outlines a new regulatory framework for online safety. Attendees called for "transparent practice" and were concerned that regulation may stifle innovation and creativity and while enforcing sanctions in this space may present difficulty, being proactive may help. Regulatory difficulties in this space are pushing us forwards and driving progress, and while a business focus is important, there is a wider view that must account for individuals in a way that is accessible for platform and business owners to implement. Concerns raised included data-mining from "easy targets" and platform owners favouring some situations (or contexts) and groups of people, over others. The market alone cannot solve everything.

While discussing the insurance industry, the question arose as to what "ethical" means. Insurance companies exercise a form of their own regulation – if customers are given data-collecting sensors and use them then premiums are cheaper (but prices are not driven up if individuals do not provide data). Discussion arose around whether a "data free-market" would be possible, where individuals use their data as currency. Instead of relying on the ethics of an insurance company, or any company at all, are we at a tipping point where regulation is needed instead? This problem may need a risk framing, to see how far risk is spread across the sector, across organisations and how much is taken on by those being insured.

The technical design process (referred to as "code design") was identified as a key consideration. Attendees discussed the importance of teaching ethics to those who develop software and encouraging "ethics-centred" approaches and this

---

**Scribe**

Arianna Schuler Scott

---

1 "CDEI Snapshots are briefing papers designed to build understanding on ethical and governance issues related to the development and deployment of AI. They aim to separate fact from fiction, clarify what is known and yet unknown about a particular issue, and outline possible avenues of action by government and industry in the near future" (CDEI, 2019).
is where the room divided. A software engineer follows company rules. In practice, this looks like a development process for creating, maintaining and fixing services and features. It is a software developer’s job to create software and while some people may see a service or feature as unethical when it is released, what is more realistic is that these creations can be put to malicious use after they have been released into the world. It may be unrealistic for an individual to protest an “unethical product” if they wanted to keep their job. In asking what a software engineer’s “ethical locus” is, or where their ethics are specifically situated is not relevant to wider problems around the design of algorithms and other AI tools because individuals have mixed backgrounds that feed into a complex ethical understanding.

Conclusions

This section largely describes the plenary summary given at the event, including extracts from workshop discussion as and when appropriate.

There were three technologies looked at today: speakers and voice assistants, deep-fakes and audio-visual misinformation techniques, and AI in personal insurance. Some of these technologies had more accessible “upsides” than “downsides”. Voice assistants for example, could reduce the feeling of social isolation and make every-day chores like shopping and scheduling more efficient. While these benefits have trade-offs (such as “talking to” a smart speaker anthropomorphising the technology, affecting social interaction and empathy), it was difficult to identify such trade-offs for deep-fakes.

Attendees had difficulty identifying positive uses for deep-fake technology, arguing that it removes the human element. Despite the apparent toxicity of this technology, the group were reluctant to suggest a total ban.

We discussed three modalities; law, ethics and code design. A regulatory approach to AI presents difficulties but these challenges seem to be driving change and innovation law. This is not necessarily a good thing as law has a specific purpose. In a sector where best practice facilitates change and “movement”, there may be room to innovate despite limitations but a proactive approach is needed. Ethics are normative, culture-based and certainly not a rigid framework, so formative discussion is needed that establishes what this baseline looks like. There is no societal consensus of software engineering ethics looks like, whereas legal and medical ethics are immediately obvious to most people. Software development is implemented by individuals but organisations have oversight of this process which is where “ethical” practice needs to be prioritised if it to happen at all. Asking what data is being accessed, where it is coming from and how it will be put to use are all prompts to start thinking about the impact of new algorithms, tools, services and products.

One final point is that in the midst of inaccessible technology and misinformation; responsible journalism cannot die. Informing the general population and holding those with responsibility to account will remain key and journalists need to help shape this conversation by talking to experts and translating accordingly.
ORBIT is the Observatory for Responsible Research and Innovation in ICT.

Responsible Research and Innovation (RRI) aims to ensure the sustainability, acceptability and desirability of research processes and outputs. ORBIT is funded by the UK Engineering and Physical Sciences Research Council. Its purpose is to provide services to promote RRI across the UK ICT research community. It aims to move beyond ICT and the UK and will provide RRI services and knowledge to all interested parties. The project is a joint initiative between the University of Oxford and De Montfort University.