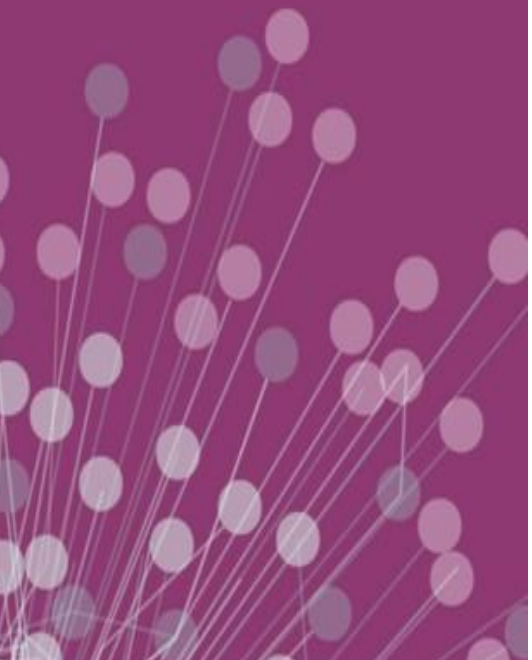
UNIVERSITY OF CAMBRIDGE

University of Southampton

Just Outcomes: How can AI make people's lives better?

A report by the Bennett Institute for Public Policy and the Web Science Institute, funded by the Nuffield Foundation



Bennett Institute for Public Policy Cambridge

WEB SCIENCE INSTITUTE

Nuffield Foundation

Just outcomes: How can AI make people's lives better?

A report by the Bennett Institute for Public Policy and the Web Science Institute, funded by the Nuffield Foundation

This report is the output of a series of four workshops in summer and autumn 2024 delivered by the Bennett Institute for Public Policy and the Web Science Institute and funded by the Nuffield Foundation with input from the Ada Lovelace Institute. The workshops brought people together from across disciplines and practices to discuss how artificial intelligence (AI) can work in society for the common good and, in pursuit of that objective, to find gaps in the research agenda informed by policy needs. The report summarises discussions about applying AI for the public good across four thematic areas – administrative justice, place, public health and market failure - and presents the observations, ideas and research questions generated in the workshops.

Contributors

We are very grateful to all workshop participants for their contributions to this project. We would also like to thank the Nuffield Foundation for their generosity, hospitality and support throughout this project.

Ben Hawes (lead author), Professor Dame Diane Coyle (Co-Director, Bennett Institute for Public Policy, University of Cambridge), Professor Dame Wendy Hall (Director, Web Science Institute, University of Southampton), Professor Pauline Leonard (Associate Dean Research and Enterprise, University of Southampton), Professor Matt Ryan (Associate Director Web Science Institute, Dr Richard Gomer (University of Southampton), Professor Michael Boniface (Associate Director Web Science Institute), Professor Paul Smart (University of Southampton), Professor Age Chapman (University of Southampton) Professor Joe Tomlinson (Professor of Public Law at the University of York).

Acknowledgement

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds research that informs social policy, primarily in Education, Welfare, and Justice. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute and the Nuffield Family Justice Observatory. The Foundation has funded this project, but the views expressed are those of the authors and not necessarily the Foundation or the Ada Lovelace Institute.

Bluesky: @nuffieldfoundation.org

LinkedIn: Nuffield Foundation

Website: www.nuffieldfoundation.org

Contact details

Email: Courtney Lee, Web Science Institute Collaboration Manager (C.K.Lee@Soton.ac.uk) or WSI@soton.ac.uk

Introduction

AI technology continues to develop rapidly and is increasingly being used to automate processes in public services and business and to inform decisions which can have big impacts on people's lives. While there is active research in some areas (including in economics, on possible impacts on jobs, on biases in data, on privacy and security, and within computer science), many open questions remain about the possible impacts of increased automation of information-processing and decision-making.

The aim of this series of workshops, funded by a grant from the Nuffield Foundation, was to scope questions that need answers if the use of AI is to deliver just outcomes for society. What needs to be understood or implemented to steer the development and use of AI for public good? How can such outcomes be brought about? Our underlying hypothesis was that without collective interventions in design and practice, there is no reason to expect that the technology will deliver just outcomes overall. The direction of product innovation in frontier machine learning and AI systems in the UK is largely determined by large US-based companies, many of which dominate digital markets and also have significant political influence. Potential users – particularly in the public sector – may not have the capacity or skills to evaluate products or tailor their use appropriately, although they are under financial pressure to adopt them.[1] One of the hopes for AI is that it will help improve public service productivity, which has been flatlining for nearly two decades in the UK. Many of these services are classic 'burning platforms', contexts in which service levels can no longer be sustained, having experienced substantial budget squeezes and, often, growing and increasingly complex demands.

That AI might contribute to economic growth and public services and create new skilled jobs is widely recognized. However, using the technology to improve productivity is difficult and even successful deployment can create new challenges. Research on the private sector indicates that "among UK firms, higher productivity is linked to the use of digital tools and skills, and the more so for those using more than one digital technology and combining this with in-house skills."[2] However, we know from previous waves of digital innovation that it takes time to understand exactly where new applications can most improve productivity in different roles, contexts and sectors, and to develop skills to use them well. As it is adopted more widely, AI could also have adverse impacts

[1] NAO report https://www.nao.org.uk/reports/use-of-artificial-intelligence-in-government/
[2] The Productivity Agenda Report 2023 https://www.productivity.ac.uk/research/the-productivity-agenda-report/

on the number and quality of jobs and on income inequality. There is a strong societal interest in shaping the evolution of AI from technical research and the direction of innovation all the way through to implementation in everyday activities and products. Developing and debating visions of public interest AI can help to address some of the gaps left by private corporate dominance, so that society and technologists can co-create to become what Tim Berners-Lee has described as ‘social machines’: networks where people and devices interact and produce emergent behaviours and conditions, enabling new kinds of action.

Digital technologies expand in their capabilities much faster than the subjects of policy development have generally changed in the past. As the technologies become more powerful, the range and variety of potential impacts, good, bad and mixed, also expands. Governments as potential users of the new technology could lead by example in adopting AI to make best use of the opportunities it presents in terms of creating more efficient and effective public services whilst mitigating against potential risks. Ideally governments should take far-sighted, nuanced and speedy action to catalyse innovation responsibly and mitigate risks, but that is asking a lot. The public sector has great responsibilities, and essential assets to protect and make use of. It is difficult to act fast and get decisions on novel technologies right. Organisations will be better able to gain public benefits from the technology if they know what questions to ask and what safeguards to put in place though innovation, implementation, and evaluation.

The workshops

The Nuffield Foundation funds research that informs social policy, primarily in education, welfare and justice. The Bennett Institute, the Web Science Institute and the Nuffield Foundation worked together to select the four thematic areas for discussion. These were chosen as priority areas to explore, where there is demand for new solutions and interest in applying AI, but where there are also important public interests to defend and promote. The workshops addressed these themes:

- AI for administrative justice
- Civic AI for place-based solutions
- AI for public health
- AI and markets: what will Silicon Valley not do for us?

The workshops were led by Professor Dame Diane Coyle (Co-Director, Bennett Institute for Public Policy, University of Cambridge) and Professor Dame Wendy Hall (Director, Web Science Institute, University of Southampton).

Attendees included experts from local and central government, thinktanks, research organisations and funders, the NHS and healthcare providers, law, consultancy, and campaign groups. We are very grateful to all contributors to this work.

The aim of the workshops and this report is to contribute to a forward-looking research agenda to help ensure AI brings about positive outcomes. We connected researchers from across disciplines to discuss what 'public benefit' looks like in the context of AI and what sort of AI sector could deliver that, and to identify a research agenda to improve that understanding. Technical choices will have implications for justice and fairness, or for public service outcomes, while the future possibilities for outcomes will be shaped and limited by technical affordances. Research can help us understand the trade-offs between opportunity and risk, and how the balance between them might best be struck.

In the workshops and this report we have use the term "AI" broadly, to include a range of techniques and algorithmic processes. We also recognised that for many organisations, the majority of applications that come under consideration in the near future will be uses of generative AI.

In each section below we present the primer we gave to attendees in advance of each workshop, followed by summaries of discussions and then the research questions and other ideas that emerged in the sessions. The discussions brought out multiple voices, experiences and perspectives, which we represent here, and which we have not sought to harmonise into a single narrative.

Workshop 1: AI for administrative justice

Primer

By Professor Joe Tomlinson and the Bennett Institute for Public Policy

What is administrative justice? There are various definitions of 'administrative justice'. Traditionally conceived, there are two key institutional aspects to the administrative justice system:

- The frontline systems through which people seek to access their legal rights and entitlements - the focus has traditionally been on administrative decision-making, but it can extend to consideration of the wider aspects of service design and management;
- The complaints and dispute resolution processes through which people can seek redress for their grievances about public officials.

These include systems such as internal review, ombuds, tribunal appeals, and judicial review, described together as 'the administrative justice system'. Some would argue that it is imperative to see these systems as an integrated whole, but not everyone does. Some people even argue that it would be mistaken to see them as integrated. In practice, the system is enormous—with millions of important decisions made every day and many more interactions occurring around those processes. The system is also complex as processes, including decision-making processes and what rights of redress are available, often differ between (and sometimes within) different areas.

To (attempt to) simplify things a little, some analyse the system 'vertically' by reference to policy areas or service functions (e.g., tax, social security, immigration, education), and others look at it 'horizontally' by reference to common institutional mechanisms (e.g., first-tier administrative decision-making, tribunal appeals, judicial review). Researchers also often use a combination of the two and look at the functioning of different mechanisms in one policy/service area.

A key factor in the design and operation of the administrative justice system is that responsibility for its design and management is, generally speaking, diffuse. For instance, tribunals might be under the management of HM Courts and Tribunals Service, but the decision-makers being challenged in those tribunals will range from across a whole host of government bodies. Thus, questions of reform often engage multiple government bodies (and other parties) that often have differing interests and viewpoints.

Underlying these institutional components are questions about what requirements the idea of 'administrative justice' imposes on the design and operation of institutions. As with any such use of a concept, there is disagreement on what exactly it requires. However, the arguably dominant strand of administrative justice research focuses on trying to tease out the different ways processes might legitimately be organised, accepting that public service and justice system administration often involves 'doing justice within the limits of the possible.'

What can AI do to enhance administrative justice? To date, the research on AI and administrative justice has generally focused on the use of AI and automation in decision-making systems. The issues generally considered in such research include the sort of errors that AI can make in administrative settings, how they can be challenged (including problems with challenges), and whether the public perceives these sorts of AI applications in this context to be fair and legitimate. These are important issues and must be kept central, but there has been much less consideration of how AI might be used within the administrative justice system, particularly beyond decision-making within cases. However, it appears that AI has the potential to improve management of the administrative justice system, if deployed carefully.

One area is the function of feedback loops in organisational learning. The problem is simple and long-standing: all of the complaints, appeals and challenges to government decision-making are also a source of data from which first-tier decision-makers (and other officials) can learn to improve their decision-making. The goal should be to get decisions right the first time, but decision-making structures often make the same errors repeatedly, which then piles (expensive) work onto tribunals and other complaints systems. The main challenge is getting the processes in place to make sense of the mass of decisions and then communicating this to people with the authority to change decision-making practices. There is a clear opportunity for beneficial impact through AI here, as it could assist in the sort of analysis and communication that could enable better initial decision-making for people and lower costs overall.

Questions

- What research do we need to help advance current policy and practice in this field?
- What characteristics will both AI models and/or regulatory/legal frameworks need to deliver fair processes and outcomes?
- What role can AI play in improving appeals and challenges to unfair administrative decisions?

- How can AI chatbots be fair and effectively deployed to enhance administrative justice?
- How can the justice system be made fit for purpose in challenging unfair uses of AI?
- What uses of AI in the administrative justice system might help or harm public perceptions of the legitimacy of the system?

Discussion

These processes sit within integrated systems which serve both individual needs and rights and public policy objectives. In some areas of administrative justice the status quo is not sustainable, and case backlogs are leading to injustices. Justice is not a service, but access to justice has service elements and there is demand for new solutions. The system also has wider public functions in setting and enforcing norms, and regulating behaviour and activity and it is critical to maintain public trust in the processes and outcomes.

Administrative justice systems manage several stages within one case process, through case identification, preparation, process to judicial decision-making. There may be stages of these processes, including support for applicants to develop their cases, and distilling contested issues from a case file, where AI could help without being involved in judicial decision-making. Other stages may be a poor fit for application of AI because of risks, uncertainties, or potential damage to trust, confidence or transparency.

Certain processes in administrative justice processes may make them comparatively promising areas to apply AI. The decision processes are rules-based, and there are large numbers of previous cases that could be used to train decision-making systems if made available. There are already end-to-end online systems in use, so there are digital and data foundations to work with.

Acceptability of applying AI in a process may be different when the context is the initial decision-making system, from in a subsequent mechanism of scrutiny or appeal.

In administrative justice, a successful appeal implies a mistake in the initial decision-making process, that has caused harm and incurred costs. It may also imply poor feedback within the system, if there is a recurrent tendency to this kind of error that might have been corrected. AI might improve the use of feedback to re-engineer systems and decision-making.

Administrative justice systems aim to limit inconsistency and uncertainty, which may give them a particularly low appetite for the risks which come with using a new technology application.

Research ideas and questions proposed

Current use of AI: We need a better evidence base about where AI is already used in administrative justice, and in related and comparable areas including welfare and social protection. What have the impacts and results been to date? Ideally this would be international in scope.

Research on the current state of play may well show that it is in use already in areas like case-preparation by different parties in administrative justice processes. AI may not be used in decision-making yet, but if it is used to modify the inputs into decisions, it is already part of the picture.

Sandboxes: In this context, a sandbox is a testing environment where a software application can be trialled safely because it does not have real world impacts. Making effective use of AI might involve changing established structure and processes, and that could be hard in justice systems that put emphasis on precedent and custom. Space to experiment is constrained, so it would be helpful to know of any sandboxes run to test AI in legal and administrative decisions. Some other jurisdictions may be particularly relevant, including Australia, New Zealand and Canada given their similarities to our justice system. It could be instructive to run tests using models trained on previous case data, and tested on separate previous case data, to explore the outcomes of decision-making by a machine learning system to help identify potential benefits and challenges.

Reviewing duties: It may be advantageous for administrative justice organisations to review exactly how new uses of data could help deliver their statutory duties, and any ways that novel techniques might risk diverging from those duties, or going beyond the legal basis they establish. This kind of analysis should aim for more than compliance, to develop organisational understanding of what is involved in aligning AI to serve specific duties and objectives.

It would be helpful to work through what kind of public interest obligations could support controlled access to data in pursuit of better access and more fairness for future users of the system. The public is a party in disputes, and has demonstrable interests. Research could explore the experience of other jurisdictions: the US allows access to judicial data analytics, while France prohibits it.

Access to information: How might AI improve access to information, without having to change any law, regulations or other fundamentals? The data is not available or not well collated and curated in

some areas, including local government decisions. Research could identify opportunities where informational improvements could be delivered without changes to law or systems.

Trust and Legitimacy: How would trust and legitimacy (as understood by all participants) be affected by increased use of automation, across different stages and domains in administrative justice? Automation could be perceived as undermining judicial independence. There may be concerns about a lack of transparency around AI-based processes and decisions. Perceptions might not be in common across the public: research elsewhere has shown that some groups are less likely than others to trust machines in comparison with people. That said, there may be decision processes in less critical, lower stakes contexts, where desire for speedy resolution and a limit on negative impacts may make users relatively more accepting of increased automation. There may be differences between expressed and revealed preferences, so research on perceptions where a decision process has been automated could be particularly valuable.

The recent report by *Justice, Beyond the Blame Game* (June 2024) looked at the risks to individual rights posed by the outsourcing of public services and recommended a more hands-on and rights-based approach from public organisations to the management of their digital services. The same organisation's newly-published report, *AI in our justice system* provides a framework for assessing the applications of AI in justice more broadly and whether they weaken or strengthen that system.

It may also be possible to learn more from previous stages of technology adoption within administrative justice contexts. Previously, appeals processes and hearings have been moved to take place remotely, online. Clearly, that shift is not the same automating decision-making, but there may still be lessons to learn about user expectations and experiences, and perhaps about what has worked in terms of supporting and empowering participants.

Insurance: It may be helpful to look for comparative experiences in insurance sectors, where providers have been using algorithms for longer, to understand how their use is managed and circumscribed, including in relation to appeals, explainability and the powers and role of any ombudsman. Trust may also depend on outcomes for the individual: it would not surprise if people were happier to be exonerated by a machine, than to lose from a decision by one. There is ongoing research about trust which could be brought in.

Workshop 2: Civic AI for Place-based Solutions

Primer

By Professor Pauline Leonard, Professor Matt Ryan and Dr Richard Gomer, Web Science Institute, University of Southampton

Most citizens meet the state not in grand corridors of power, but in the places where they live, work and play. Our interactions with locally delivered public services are not just routine; they are vital threads that contribute to our wellbeing, quality of life, and sometimes, our very survival. We rely on refuse collectors, accessible public transport and quality highways, welfare, social care and education support as well as parks and green spaces for recreation. Our lives are significantly changed by the way we interact with teachers, police, social housing landlords and carers as well as policymakers in local authorities.

This workshop explored how AI could be best applied to local services to improve access, enhance outcomes or deliver services more efficiently and justly. We discussed the constraints that will be on AI in civic place-based contexts by the need to maintain democratic accountability, legitimacy and accessibility, and what could be the right technical, organisational, and governance frameworks for AI in civic contexts.

Opportunity: Local authorities and other civic organisations deliver myriad services to their communities. AI affords faster access to tailored information, process automation, and new data-driven insights. Where, within everything that those organisations do, are the biggest opportunities for AI enabled improvement? The pandemic demonstrated that local data can help solve localised issues better than national data. How can communities work with local authorities to improve how non-personal data is collected and used to make decisions about areas?

Accountability: AI can optimise behind-the-scenes operations and enhance citizen-facing interfaces using chatbots and virtual assistants, but AI can also systematically deliver discrimination through unjust, unexplainable, or uninformed decisions. Much of the recent debate on AI has been led by private companies who are subject to market forces but do not receive democratic oversight in the same way as public bodies. There are important differences between public services and commercial services that motivate a specific focus on AI in civic settings.

AI deployment in public services requires ethical and evidence-based decisions, ensuring that these technologies serve the public interest, uphold democratic values of accountability, and enhance the quality of life for all. How can we ensure that the deployment of AI technologies in civic settings preserves and enhances the democratic relationship between public services and the people they serve in all their diversity?

Locality: Public services can benefit from economies of scale, but also need to be attentive to local variations, maximising the creativity and opportunities afforded by different places and regional geographies. AI requires technical expertise in its deployment, but many local authorities currently lack the resources and skills to develop, incorporate, and especially maintain systems that benefit from advanced technologies. What does a modern public service, and the associated technical support required to deliver theses services need to look like to serve varying place-shaped local needs? And what is needed to ensure that all areas and communities within the UK are able to benefit from AI in public services – so no one is left behind?

Capability: The skills to marshal data, feed it into new AI systems, and integrate those new capabilities into existing or re-designed processes are still emerging, but remain the preserve of relatively few experts. How can we best resource and optimise local public service delivery, build skills and capacity to maintain democratic oversight while harnessing new technologies to serve the public good?

Discussion

Local authorities, public services, and other publicly funded organisations including third-sector charities and arts-based organisations have unique needs and constraints in relation to AI. The pandemic demonstrated that local data can help solve localised issues better than national data, and that local organisations have the capability to adapt and adopt new ways of working.

AI deployment can and should be adapted to local needs and resources. Conversely, where approaches are successful, there will be benefits in spreading that success. So there is a role for mechanisms or organisations that research “what works” in local AI and offer lessons – not fixed models – to other places. In local government in the UK to date, this has arguably not yet been fixed: how to spread successful digital delivery through a dispersed system (of local authorities and other local public bodies and service providers) in ways that enable, and support responsiveness to local

conditions and do not centralise. AI is arriving in a local government context which has not resolved how to spread digital delivery best practice, but is well aware of the problem.

AI Localism is a term coined by Stefaan Verhulst and Mona Sloane, that describes actions by local decision-makers to govern use of AI within a place or community, generally because it has been determined that national or global governance frameworks have not provided sufficient or optimal tools. There are repositories of these actions and tools, which could be much better known.

Local services: AI could optimise behind-the-scenes operations and enhance citizen-facing interfaces using chatbots and virtual assistants. AI could also systematically impose discrimination through unjust, unexplainable, or uninformed decisions.

Inclusion: Applying AI could result in the same failures to include some communities that have been seen in the past, if existing datasets are relied on uncritically. However, AI may also offer new ways to assess data and systems to improve accountability, fairness and the way a service meets the particular needs of individuals and communities. The challenge will be in developing AI and data-sharing mechanisms that can deliver all these results. There is work going on around the world on local data systems and locally focused AI applications and frameworks to govern them, that can be learned from. AI could help to map the service providers working in an area, improving information and transparency.

AI analysis to address bias in service provision: Racism and other social biases can affect provision of support services, in particular to young people. Bias is not just vertical, and needs to be addressed across all these domains and audits. AI-enabled analysis could map and address imbalances using a range of methods and inputs, combining data from macro levels and from lived experience, diversity of sampling and participation; co-creation in design, and interdisciplinary approaches and knowledge.

Representation and deliberation: People living in places know what goes wrong with services in those places, and why, and they can come up with new approaches that work in local conditions. AI-enabled platforms could support debate and provide information about the depth and richness of local opinions by offering new ways for decision-makers to elicit opinions, insights and ideas from communities. Different communities, groups and generations have different interests and voices which should be heard. Young people will see more of an AI-enabled future, but older people

might meet more of it in their lives sooner, in social care. Young people may routinely be underserved because they cannot sign up for services where authorisation is restricted to over 18s.

Environment and community action: AI could support dynamic mapping of environmental circumstances and conditions, while involving and empowering communities. Digital mapping of natural capital works and technologies are improving, with local and international best-practice. Access to more data does not necessarily make decisions simpler, but can help decisions to be better informed. AI-enabled communications could use real-time environmental reporting to provide local and personalised advice.

Research ideas and questions proposed

Skills: Most local authorities currently lack the resources and skills to develop and maintain systems (in particular data systems) that would support testing, evaluating and applying AI to core services. Local government leaders have many pressures to manage. While many UK local authorities have come a long way in digital delivery over the last decade, and often with relatively little central government support, few local government leaders are digital experts. In relation to AI, many are waiting until others have demonstrated clear and replicable success, and that is a rational position, in particular in relation to sensitive data and services for vulnerable citizens, children and health.

In the longer term, local bodies might need to develop skills in several categories. What local leaders may most need now is guidance on which skills to build next. This is not simple: for instance there is currently very lively debate about how much generative AI will remove or change the need for coding skills. If technology companies do not know the answer, it is unreasonable to expect local government leaders to. Research with the sector could identify the skills they feel they most lack in this space, and which skills would most enable decision-making. Research could explore how international countries, cities and regions are approaching this.

Agency and digital public assets: Citizens might expect that the local public sector could take the lead, if they had the right skills and resources, but it is not clear how confident local organisations and leaders are about building AI capabilities. This may be experienced as a step-change in their activity that needs clear direction and perhaps new legal basis.

Research could improve understanding of how local leaders view the development of public digital assets. It could identify what they need to make decisions in the medium term, including to take on new responsibilities like owning and developing an AI model over time, and what approaches might enable a local public sector organisation to avoid losing control by contracting this out to private entities. Research could explore where any non-profits or collectives have successfully trained AI on local data for local public interests, and how they have set objectives.

Deliberation and legitimacy: There are related questions around using democratic deliberation tools. Using them to surface opinions is relatively uncontroversial, but using them to make significant decisions can potentially clash with established local democratic processes including local elections and representative processes. Public sector bodies or community organisations may need a civic sociotechnical framework for using AI in support of local deliberation and democracy, and means to ensure that it does not introduce new challenges.

Local data as community asset: There can be cultural and institutional barriers if public sector bodies are willing to use and share their own data, but not to recognise data collected by other bodies. Mechanisms to bring in citizen data can increase representation and legitimacy, and are potentially a good in themselves for increasing agency and ownership. Many people have a reasonable sense of what happens with data they hand over now, for personalised services, but there is less understanding of what could happen when that feeds general purpose AI. These questions are relevant globally, and research could identify successful approaches where public bodies have gained the trust of citizens for their use of data about them.

“Data sovereignty” often describes national policies on location of data and processing, but it can also describe the capacity to control and access local data. EU law now has data sovereignty clauses to help local governments to compel companies to share data of public interest. Local businesses may hold a great deal of data like that, which they could share (without necessary safeguards and processes) without harming their own interests. Research could uncover what mechanisms are in use by local governments to access local data held by local companies, and what works. Jersey has a function to collect cycling data, and trust law that is more supportive. This is a potential case-study to develop.

Environment and participation: Citizens and community groups can become more active in local data initiatives to build information resources and improve connectedness to nature and

places. Local people and organisations will be aware of the stressors in their places, like flooding, fly-tipping, and noise. Data can lead to changes in priorities, as it has in relation to air quality. Data can also help predict where known issues could become more critical in the future, for example with urban heat islands. Some challenges (water management, traffic) need a group of organisations and the public, to contribute data and contribute to solutions.

People increasingly want to know what environmental impacts are made locally, whether locally specific (water quality) or local instances of broader factors (carbon emissions). Research could identify what metrics are relevant, achievable and explainable through a place lens, how these scale up and down (street to region), and at what scales and time horizons they can deliver actionable information.

AI agents may be the next busy area of AI development. Research could explore how a local AI agent might interact with national and local bodies (in terms of institutional relationships, access to data, and legal basis and compliance) to deliver more efficient and personal services.

Spatial data: Maps can be liberating for people and communities, but have also been instruments of extractive practices, oppression and exclusion. Increased access to spatial data can make planning more efficient, but that may not benefit everyone equally. Legal rights in location data tend to be limited, personal and negative, not positive and collective. AI applications using spatial data could help consultation, add insights to inform decisions, and aid in explaining decisions and spatial planning. Where there are conflicting interests – for instance around a proposed traffic measure – using local spatial that data in the collective public interest is complicated. Research could improve understanding of skills and models that would enable more effective use of spatial data in community interests and in relation to contested decisions.

AI for local planning consultations: Citizen capacity is limited by time, bandwidth, exclusion and perceptions of exclusion. Better understanding of barriers to good consultation could enable better design. Research could collate what works, in using digital tools to improve access to information and participation and reducing admin burdens.

Participants discussed interacting generative AI applications that could elicit more and better inputs from citizens while reducing the burden on them, and minimising risks to privacy. An LLM could distil perspectives from consultation data, and read council documents and meeting minutes,

transcriptions, maps, plans, and local demographic data. A model would learn, and could increasingly reflect the type and range of questions and concerns citizens have. AI visualisation services might help to show citizens the impacts and trade-offs of different options from a development or traffic measure in a consultation, illustrating environmental costs and benefits and other impacts, using common reporting standards.

There is already some activity in this space. Research could evaluate what works, and where risks and benefits are emerging. That could include exploring and mapping categories of local data, to illuminate which datasets can be used with AI to deliver new value to citizens and services.

Targeted funding could enable communities to develop models for specific places, and explore what could be a sustainable funding model for maintaining and developing a local AI model over time, and what incentives could bring in the full range of participants.

Trialling AI to address bias in service provision and outcomes: Design a process to develop an AI tool with a community to explore health disparities among groups of young people. Work with civic institutions, government bodies, delivery organisations, local communities and technology developers. Explore how young people can be successfully involved in reducing structural discrimination, uneven representation and unequal outcomes.

Workshop 3: AI for Public Health

Primer

By Professor Dame Wendy Hall and the Web Science Institute

Understanding the relationships between people's mental and physical health and their social and economic circumstances can inform policy and practice interventions. Nuffield Foundation supports research into these relationships, as part of its wider portfolio addressing inequalities, disadvantage, discrimination and vulnerabilities that people face in education, justice and welfare. Previous grant funding has delivered insights on health inequalities in later life, health impacts of early years education interventions, and social effects of pandemic isolation measures. This workshop focused on how AI could be applied to help tackle the causes and consequences of ill health in communities, and on public service co-ordination as it affects public health outcomes.

The hypothesis is that AI applications could help to improve public health in some contexts. To test this, it is necessary to identify specific opportunities and areas of risk, including how to identify issues before they become problems, and situations in which badly designed or applied AI tools could cause harms, undermine the performance of public health systems, or damage trust in them.

Understanding lives for prediction and prevention: AI applications might predict where public health challenges (for instance obesity, heart disease, or mental ill-health) are likely to develop in specific areas and among age cohorts and other patient groups. Applying AI to deliver this kind of insight could anticipate and prevent poor outcomes, and enable better targeting of resources.

Training AI models to predict future public health demands could involve combining data from clinical sources, GPs and social care with data on housing, benefits, environmental conditions and other place-based factors. The need to protect privacy and ensure security has resulted in an environment where these domains are, in terms of data, walled gardens isolated from each other. We want to identify the critical questions related to managing trust, legitimacy and risk, in using data from multiple sources with AI to guide preventive measures.

There are already tools for managing access securely, to support data flows across institutions, including data trusts and privacy protecting applications. Data trusts are mechanisms that make it possible for individuals and organisations to provide access to their data collectively, with access in the control of trustees. These mechanisms rely on dedicated frameworks. We want to understand the regulatory and legal frameworks that would support an ecosystem in which public health researchers can work with data from this wide ranges of sources.

For that ecosystem to function well, it would also need the confidence and support of individuals, citizen groups and wider society. We want to understand the factors that would enable informed consent to use AI across a wide range of data sources to be nurtured and deserved.

Coordination and fragmentation: The workshop also considered how AI could help address coordination and fragmentation in public health, in and between the institutions, professions and people involved in planning, delivering and receiving healthcare. Fragmentation is part of the daily experience of many healthcare professionals and patients, but there is a lack of analysis on exactly why it is so prevalent and how it could be reduced. AI applications might deliver new insights to improve the quality of interactions, communication and trust between healthcare professionals and between them and patients.

Recognising that there will be pressure for AI to deliver savings in public health, we will also examine what principles and practices would ensure that AI is used to deliver better and more cost-effective healthcare. That could include assessing the risks of using AI at scale and in many functions simultaneously across public health.

Questions:

- Where do you see opportunities to use AI to help prevention of illness and disease, and improve public health delivery and outcomes?
- How could AI improve understanding of the relationships between health and social, economic and environmental circumstances?
- What data and regulatory and legal frameworks are needed to apply AI optimally in public health?
- What institutional arrangements and accountability mechanisms will enable AI and data to be used across the different responsible institutions in public health?
- Where are the critical gaps in knowledge about fragmentation in public health, that AI could help to fill?

Discussion

New sources of insight: Technological devices provide opportunities for health-related monitoring, and across organisations far more data about our lives is being generated than was the case a decade ago. Health outcomes are influenced by complex factors including education, housing, and life events, which are not well integrated into clinical data. A richer perspective would recognise the interplay between multiple factors including genes, physiology, behaviour, physical environment, and social relationships and context. Data within healthcare can be supplemented by data from additional sources (Non-Traditional Data, NTD), and by qualitative research data that explores aspects of patients' experiences, to fill out understanding of Social Determinants of Health (SDOH).

Integrating data: Existing health data platforms are not well-equipped for sharing SDOH data across different sectors. There is a need for federated data platforms to aggregate data securely and bridge gaps between healthcare and social factors. Policies for data collection may not be co-designed with end-users, limiting practical application. There is fragmentation of data controllers and liabilities across multiple datasets. Bringing together datasets will need authorisation and skills in the same place. In some areas there may be a need to simplify law.

Additional analytical techniques: Individuals' social and environmental contexts evolve over time, and traditional models struggle to account for this complexity, hindering accurate predictions and recommendations. Over time, a person's health may emerge as the result of causal loops spun out over extended periods of time, that cross disciplinary boundaries, with causal interactions spanning the genetic, physiological, psychological, and sociological realms.

Bringing in wider datasets and new analytical tools could help develop a better understanding of how shifts in certain causal factors might affect future health-related outcomes. There is a distinction between macro- and micro-simulation. The macro-simulations can be used as the basis for policy-related decisions, while the micro-simulations can be used for tailored interventions and recommendations based on an individual's current context and history.

There may be a strong case for more use of generative AI models that have been used to model health-related phenomena, including those used to model affective states and bio-psycho-social interactions. AI systems that shed light on human mental phenomena may prove helpful in identifying social and emotional factors affecting health.

Resolving fragmentation: A more effective public health ecosystem would also have better coordination and less fragmentation in and between the institutions, professions and people involved in planning, delivering and receiving healthcare. Fragmentation is part of the daily

experience of many healthcare professionals and patients, but there is a lack of analysis on exactly why it is so prevalent and how it could be reduced. At present the relevant environments (home, surgery, hospital) and data from them are not integrated. There could be uses for different AI applications in collecting, integrating, managing and applying data across the spectrum from lifestyle context through to interventions. This might include better collective intelligence at local levels. People in communities know about causes and opportunities, but that knowledge can be poorly connected.

Preventive health enabled by more and richer data might help to counter the medical prioritisation of length of life over quality. New and broader perspectives could also act a counterweight to the focus on the NHS that undervalues other health and social care that is outside it.

AI tools might support improved community segmentation and personalization. There is potential for AI to provide tailored recommendations, bridging the gap between population-level and individual-level health modelling. AI may also offer ways to better assess the outcomes of policy and healthcare interventions. There should be better connections between national guidance and enabling tools, and bottom-up community projects. If AI could help improve prediction of health problems, as well as helping to head those off with preventive measures, it should also improve prediction of demand for healthcare.

Communication: Poor communication contributes to adverse healthcare outcomes. AI applications might help to improve the quality of interactions, communication and trust between healthcare professionals and between them and patients. At one level, that could be using AI-enabled communication services to give personalised information on activity and volunteering opportunities to members of the public, potentially expanding the depth, range and reach of preventive health advice.

AI could automate personal access to information, guidance and health administration. Building on that, it may offer better ways of using personal, communal and environmental factors to give advice to individuals, including holistic advice towards better health, as well as reducing risks of specific conditions. A personalised AI service could advise different individuals in different ways of according to their preferences and capacity to use advice and make lifestyle or other changes.

Misinformation about health is already resulting in poor health outcomes, and affects some groups more than others. AI analysis of social media and other information sources could identify what misinformation is circulating, the channels involved, and groups likely to be particularly affected. AI applications could be used to target corrections to misinformation.

More data and emerging applications could connect up elements, provide insights and manage data in a 'P4' ecosystem: predictive, preventive, personalised and participatory. However, the supporting structures are not in place. AI for public health is difficult to invest in, given uncertainties around agency, authority and priorities, data privacy and rights, and liability.

Again, there should be no assumption here that AI is a default solution. Misuse of resources is already a public health problem. AI in healthcare in general shows a gap between promise and delivery. The wealth of data now generated could lead to new statistical insights even without using AI models, which could be missed if AI is over-emphasised. On principle, using data to develop AI should reinforce rather than complicate ongoing work to improve data sharing.

Research ideas and questions proposed

Guidance and decision-support for leaders: Compared to administrative justice (for example), the horizon of data that might be relevant to public health is broad, and so is the spectrum of unknowns. The medium term focus should be on what additional preventive applications are achievable with currently accessible data, technology and institutional arrangements, rather than seeking to answer too many questions at once. Preventive health AI could interface with other emerging technology areas including genomics.

In this context there is a need for practical means for assessing whether a specific use-case of AI in preventive health warrants resourcing. This would include the fit of AI tools to needs in the preventive health space, and what additionality they might offer, as a basis for low-risk trials and a roadmap. Case-studies would make it easier to map opportunities as high / low value and risk, and as near / long-term opportunities. Research with decision-makers could clarify what factors most influence their decision-making on novel applications of technology, and what information they need most, or feel the lack of most.

Data: Research should develop a broader and developing view of what data is relevant to societal well-being, communal health in places and contexts. Understanding of relevance can sometimes change quite fast, as has been seen in relation to causes of air pollution and harms caused by it.

Training AI models to predict future public health demands could involve combining data from clinical sources, GPs and social care with data on housing, benefits, environmental conditions and other place-based factors. In terms of data, these are walled gardens isolated from each other. Research could identify comparable contexts where techniques have been trialled to combine observational data, contextual data and clinical data into machine readable datasets and streams.

Research could surface regulatory and legal frameworks that would support an ecosystem in which public health researchers could work with data from this wide ranges of sources, beginning by identifying high priority and achievable integration.

People at different economic levels have different types of data created about them, for reasons to do with the incentives of the organisations collecting the data, rather than because of their needs, healthcare and otherwise. That data may over- and under-represent people in different ways. This may require additional research to deliver fairness across cohorts and communities, and avoid exacerbating social inequities.

A lot has been learned within medicine about building trust for data-sharing, so it will be useful to determine what is different, and why, when using data from more sources, and for prevention, to be clearer about what can be adapted from more traditional and developed healthcare contexts.

Trust: Several factors influence trust, including past history, evidence, appreciable protections against misuse, availability of redress, and additional measures for high-risk areas and/or low trust populations. Fear of data breaches can cause excessive risk avoidance, on all sides. The public do not want health data sold to or shared with Big Tech. A risk with preventive health is that in seeking to be open-minded in taking in data from many sources (food and drink purchases, movement, financial) it would be possible to fall into supporting excessive surveillance, and also to trying to use unmanageable quantities and variety of data. There is a need for ways to identify relevance and prioritise. The focus could be on need, and new ways to reveal need, rather than data about previous service use, or historical data that reinforces biases and omissions.

What would it mean for a data-driven AI system in preventive health to be trustworthy? For example, where new insights might create a case for a local authority or other agency to do something differently, trust would be conditioned by existing trust in that agency. Research could develop understanding of how trust functions in preventive health, which organisations are trusted more and less in this context, and why.

Communicating in preventive health: Behavioural research should guide the use of AI systems that advise individuals on risks and ways to manage them, by providing insights into what kind of information people can use, what they respond to, and how to maintain more advantageous lifestyle choices rather than reverting to normalised habits.

There may be lessons from genomics. The genomic promise is that genetic information can indicate higher risk of developing a condition, but over-reliance on it can distract from lifestyle and environmental factors. Genomic counsellors have worked in this area to provide realistic,

proportionate and actionable advice to individuals, including on the interaction of individual and collective needs and interests.

Some communication around data in preventive health should make overt the interactions between personal benefits and collective benefits: “how will this benefit me, how this will benefit us.” Suppose that by participating in data sharing, one could help alleviate the suffering of the next generation by (let us say) the development of effective treatments for Alzheimer’s disease. In this case, the costs of data sharing (whatever they are) remain the same, but the nature of the problem is transformed into one with moral implications. A preventive health data social contract could help individuals use data from the community, and contribute to community-led knowledge and informed choices. Kings Fund uses the term “shared responsibility”. Research could explore means to shape, represent and communicate specifically collective benefits, exploring any examples where comparable familiarisation and negotiation processes have nurtured pragmatic and successful agreements.

Legal and regulatory issues: Guidelines are emerging for liability in responsible AI and AI for health, but it will take work to embed them into practice across the organisations who could be involved in preventive health, which may also cross regulatory domains and need work on consolidation. Insurance already uses lifestyle data, so it may be possible to learn from that sector what is acceptable, how data can be used, and where sensitivities and inequities can occur.

There are important concerns about how to align commercial interests with public health outcomes, whether that is the commercial interests of companies in health or the wider commercial world and its impacts on public health. One aim of data-enabled public health should be to develop incentives to support companies to promote health. Research could also explore liability for outcomes from preventive measures or providing preventive information, including nudges.

Workshop 4: AI and market failures: what will Silicon Valley not do for us?

Primer

By Professor Diane Coyle, Bennett Institute

As participants in markets, we - the public - can in theory decide not to purchase products, but the way digital markets are structured means these are often not genuine choices. Featuring large economies of scale, non-rival products and network effects, there is no reason to believe market forces will deliver good outcomes for society, and every reason to expect the concentrations of market power that have indeed emerged. So there is strong public interest in the way AI evolves as a set of commercial products.

As the public, we are also funding AI research and have a collective interest in whether it is developed responsibly and regulated appropriately. Leaving AI "to the market", when there are significant market failures, is therefore multiply flawed. What's more, the technology is evolving rapidly, so there is a need for creative thinking about shaping it in the interests of society.

The aim of this workshop is to generate thinking about moving AI beyond its current limitations (for which Silicon Valley is a useful shorthand). If we consider society and technologists as co-creating what Tim Berners-Lee named 'social machines', how can the direction of travel be influenced? Is it through the character and culture of the technology and innovation community, the incentive structures created by financing models and legal or regulatory frameworks, the ways AI is being adopted and used given the context of immense market power and Silicon Valley values?

The identity of researchers and innovators matters because individuals' experiences shape how they understand society. Research questions or the development of AI products and services – the direction of innovation – will be distorted if the social and cognitive make-up of the technical community remains so narrow. What actions – and by whom – can address this, and dilute the influence of the more extreme tech philosophy (such as the TESCREAL ideology or, less dramatically, its limited intellectual range). TESCREAL is an acronym neologism bringing together a set of overlapping ideologies ascribed to groups of Silicon Valley AI evangelists and developers:

Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalism, Effective altruism, and Longtermism.3

How should AI decision-making be shaped? The technology is naively utilitarian, assuming needs can be encoded in a regret or objective function. Yet – as Alan Turing observed in his famous 1936 paper On Computable Numbers – most problems in society cannot be solved by a finite set of procedures a computer can calculate. Any decision will involve a conflict of interest (shareholder or consumer benefit? taxpayer or benefit recipient?). Given this, how should government, public bodies and businesses avoid the naïve use of AI decision-making in their activities? Does the technical community understand the limitations when it comes to the development of practical applications?

What is the role of the financing structures (research grants, VC investment) in shaping the direction of AI development; how do these contribute to the dysfunctional outcomes – such as mass surveillance, theft of creative ideas and substituting for creative humans, monopolised markets?

What other models could create different incentive structures and market dynamics? What AI technologies has the for-profit model prevented? Why is tax on online advertising (not just services) not more widely levied? Could a public sector or non-profit competitor change the dynamics? What might give governments and regulators the public permission and confidence to doubt AI hype and prevent AI harms?

Are there actions that could create an alternative set of markets for AI that serves underserved markets (analogous to how advance market commitments encourage the development of vaccines or medications for markets too small or poor to be commercially viable, or how standard-setting increases the scale of markets to make them viable as was the case with mobile telephony and the GSM standard). What needs are not being addressed by the AI community? How can they be served?

[3] The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence, Timnit Gebru, Émile P. Torres 2024.

Discussion

There was agreement with the proposition that the digital technology market characterised by huge economies of scale, market power and network effects, will not necessarily deliver good social results. The private sector will focus on profits, increasingly through lobbying and rent-seeking activities, and under-provide public goods. Internet platforms are public information and communication spaces that can condition how people think and interact with each other, and they can have negative impacts on individuals and on society. Now the same companies dominating internet platform markets are leaders in developing AI, which may embed their market advantage even further.

On its current trajectory, AI will not be applied to some public interest purposes, but it is not clear yet on where the AI-shaped holes in provision and development for social good really are. Novel public interest questions will emerge. Digital services use AI to nudge behaviour in ways that still look like choice. A generative AI personal assistant that knows our habits and history could make recommendations almost invisibly. It is not clear what portability – the ability to switch providers – could look like in this space, or what kind of information would need to be retrievable. Ideally a recommendation system should maximise each person's welfare, but that is not a simple thing to define and set as an objective. It is much easier to see when it fails.

Governments' fear of holding back innovation by rich companies with ample funds to invest makes them unwilling to do more to direct private initiative towards social good or to constrain tech companies. The speed of growth and change in digital and now AI markets can support the case either for special regulatory treatment or for holding back from regulation (as the companies argue), which is a different form of special treatment. States fear losing tech investment through regulatory arbitrage. Governments and regulators can be wrong-footed by hype and novelty, so that they forget lessons from successful past regulation (for instance of electricity markets). This may be self-defeating: if no one steers innovation in that direction, it will, paradoxically, only deliver more limited results overall, and under-deliver the potential of AI.

Regulation only has certain tools, points of purchase and traction, and windows of opportunity to act impactfully. Some levers only work in relation to an entrenched market position. It is not simple to determine how many AI companies are needed to make a market competitive, when it is hard to delineate the market. Digital technologies have long tended to wear down traditional barriers between sectors, and AI is continuing that trajectory, which challenges sector-based regulation. The

UK now has a strong competition law in place that can be applied to strategically important tech companies, but this is already looking vulnerable to political lobbying by the companies. Public policy and regulation are hard in this space because they require broad understanding that is not easy to build, nor easy to ensure is shared across the worlds of politics and policymaking.

The strategies of the major technology corporations have been shaped around cost, law and regulation. They have trained AI on the data that they have been able to amass or access by legal means, or at least means not yet judged to be illegal or otherwise restricted. Holding and accessing volumes of data built market power for internet platforms, and they are betting the same happens with AI. While data is the source of market power, it is arguably not well described or addressed by regulatory concepts. It is recognised in the Digital Markets, Competition and Consumers Act 2024, in measures to force Google to share data with other search engines, but generally regulation has little traction on data as a source of market power. Laws intended to protect privacy and intellectual property can apply some constraints, but are unsatisfactory instruments because they are not devised for that purpose. Data is difficult to value because the value is relational, cumulative and dependent on use. Now data trains AI, and so has a new and additional kind of value.

What different incentive structures and market dynamics could deliver a wider set of social goods and benefits? One alternative to the current market landscape of dominant AI developing corporations might be decentralised AI. AI is already being distributed as a service. It is possible for anyone who has the necessary computing resources to build a model, it just may not be as good as one paid for in the market. Individuals and companies will want their own, so services may develop to address that. The third sector is underdeveloped and underexplored as a source of AI applications that could potentially be very different than those advanced by the major technology corporations.

Concerns about the dynamics of digital markets are not new, but risks and opportunities are both developing quickly. The speed of change could easily encourage governments and public bodies towards commitments that later prove to be a poor fit for the needs of citizens, and lose control of or fail to develop collective digital and data assets. This is a good moment to build evidence for a vision of national and communal strengths in public interest AI to solve society's problems, including public computing capacity, collective data assets and data commons, public procurement, audit, energy, and public research on public value problems.

The government has declared optimism about public sector AI, but more realistic principles and practices will be needed to turn that into a movement that can work, and that citizens and people working in delivery of public services can justifiably put their confidence in. The proposed National Data Library and the recently renewed Government Digital Service present opportunities to focus

thinking and evidence about public digital and data assets, and to bring solutions and positive examples. Technology companies will seek to influence the priorities and policies of both. In that context, clarifying and presenting what public interest AI can be, and what public bodies need to make the best decisions about AI for citizens, may become increasingly valuable.

Research ideas and questions proposed

Liability: Increased automation can make liability more difficult to track and attribute, as the contribution of individuals and companies in a complex supply chain become hard to track or isolate in the operational system that combines their inputs. In relation to an AI model, relevant actors can include those who created and selected datasets, developers, organisational users, audit and compliance functions, and individual users including employees and customers or members of the public who interact with the model. This is compounded when models are built on other models A proliferation of AI agents would add to that complexity. How do insurers currently manage the use of new AI applications by corporate clients? This could cover autonomous vehicles, but also looking more widely into different application areas, including how liability is managed around use of LLMs by third parties. What might an insurance market for digital liabilities look like?

Public interest technology: What duties, powers, objectives and skills do public bodies need to own, grow and use digital assets and tools that develop over time? What additional new duties might help public bodies build capability and confidence in using their data? Changes here might require Cabinet Office guidance or a legislative change creating duties to capture public value from data.

Research could identify whether any national government has created positive duties on public sector organisations to gain public value from data, what those duties are, and what legal basis they build on. More broadly, research could deliver an overview of international approaches and concepts to public value in data. Research could explore concepts and mechanisms in use internationally. There are examples (France) where there is a more advanced concept and mechanisms for capturing public value from data. Mapping data industries by objectives and beneficiaries could show up where public interests exist and tools do not, or are not being used.

Private data and public interest: Research could identify what models have proved successful in accessing private data for public interest, and how well they support training AI models. It may be informative to work through the characteristics of different hybrid models including public investment on terms, public-private partnership, and others such as the BBC, NASA, and other national initiatives which have launched spin-offs with social and economic benefits.

Data institutions: Alternative data institutions could manage sensitive data for collective public benefit. These might be new institutions or new roles and responsibilities for existing ones, for instance universities. At present there seems to be approval of the principle of data institutions, but much less support in practice in many key sectors. Gap analysis could determine whether existing duties provide sufficient basis for what could become a substantially different set of ways of using data in the public interest.

Market dynamics: Research could identify whether any regulatory tools bear directly on data-holdings, as a source of power in markets. Research could explore historical examples to understand whether public sector or nonprofit competitors – 'public options' – in AI could change market dynamics.

Cross-cutting observations and research questions

As we progressed through the workshops, we found that some observations and questions recurred across the different themes. Here we summarise common reasons why AI could offer benefits to public services and the public good, common reasons why applying AI could fail or fall short, and cross-cutting research questions.

Why applying AI could improve services and outcomes for citizens

Consistency: Using machine learning and other AI systems to automate parts of decision-making processes could standardise decisions, reducing unfair variations in outcomes that due to human biases and other incidental factors that should in principle not have any bearing on decisions, but do in practice.

Efficiency: Decision processes might be run more efficiently, in terms of cost, of the need for domain expertise, and of burdens on users. Decisions might be reached more quickly, removing some of the delays that add to negative experiences.

Prevention and coordination: AI-supported prediction could improve preventive action and coordination (supply and demand) of services.

Data management and access: AI could enable readier access to multiple data sets to inform decision-making. Poor connections across data silos can have damaging consequences. Connecting data can enable new insights at personal and population levels. Data privacy and security will continue to be vital when using sensitive data to develop AI models, but AI applications might help fill these gaps where data currently does not get connected efficiently.

Communication: Better and more interactive tracking and communication could provide advice and reduce uncertainty for users of services. Applications might better support users in preparation and through stages of processes, reducing the imbalance of power and resources that is built into systems. AI information services could find and counter misinformation in public interest domains. AI will be used to spread misinformation, and that growing threat demands counter-measures.

Accountability: AI may offer new ways to improve accountability, fairness and the way a service meets the particular needs of individuals and communities. It may also offer ways to better assess the outcomes of policies and healthcare interventions.

Personalisation: AI could offer personalisation at scale, improving outcomes, effectiveness, efficiency and individual experiences.

Why applying AI might cause or amplify harms

Misapplication: Organisations will first want solutions for the problems which most need fixing. With providers promoting AI solutions, there is a risk that public bodies will procure them without considering sufficiently, and as informed customers, the uses that AI applications would best fit. AI tools might be used to manage persistent problems in systems, when it would be more effective to tackle the systemic problems. Focussing primarily on efficiency can lead to trying to do the same things but quicker and more cheaply, whereas reaping full benefits from AI could demand reorganisation of systems or parts of systems.

Perpetuating bias: Training a system on data from past cases will tend to replicate biases and prejudices that influenced data collection and decisions in the past.

Legitimacy: At present, the social expectation is often that decisions on complex questions and in complex contexts that significantly affect individuals are made by people. The risk of being affected by a mistaken automated decision may conflict with the expectation of human justice or fair treatment, undermine confidence in a system, and dissuade people from using to it. Already, users of systems sometimes feel that they are dealing with an inhuman machine. If more decisions were made by computation, that alienation could be made worse. Large language models effectively make informed guesses, which are sometimes factually wrong.

Accountability and explainability: In some systems users can request an adequate and comprehensible explanation of how a decision was made. Using a “black box” machine learning model could make that more difficult to deliver.

Data security: Sensitive personal data is held and used in these systems. Involving third party companies in providing AI services could increase risks of privacy violations. Fear of losing privacy could deter potential users.

Lack of long-term strategy: Part of the appeal of using machine learning applications in a public service context is that the performance of a model should improve over time. Management of AI in these contexts has to include ensuring that those continuing improvements are realised for the benefit of the public, and are not disproportionally captured by a private provider, or result in vendor lock-in.

Constraints in testing: Online platforms run A/B tests in consumer services, providing slightly different services to different users, and refining the service based on the outcomes. While past data can be used, it may not be legitimate or ethical to provide different services to different applicants in systems to deliver public services.

Cross-cutting challenges and research questions

Supporting decisions about applying AI: A decision to trial an AI application could involve the answers to (at least) these questions: seriousness / priority of the problem; characteristics of the problem and whether an AI application is an appropriate tool; availability of relevant data; availability of skills; explainability, transparency and accountability; trust; permissions, legal basis and social licence; opportunity costs.

Research could develop a better view of the capabilities needed to be an intelligent customer. It could classify the kind of problems AI can currently address in public service contexts to support leaders in making technology choices.

Research could improve understanding of cultures that have nurtured successful uptake of technology innovations in public sector and public interest contexts.

AI applications are raising new questions and reframing known ones, but not everything here is new. There is extensive experience of what can go wrong with public sector technology programmes, not least in reports by the National Audit Office. Research could summarise recurrent characteristic types and causes of failure as part of guidance for decision-makers engaging with AI.

Using AI could require creating new job roles, including in explaining AI. Research could develop understanding of what levels of detail are meaningful to citizens and service users, staff and leaders in different domains.

What kind of capabilities and guidance do leaders in government, public bodies and third sector organisations need, and currently lack, to avoid the naïve or harmful use of AI in decision-making? These organisations are under continual pressure to improve efficiency in terms both of time and of cost. Companies promoting AI services and optimistic political advocates will promise cost-cutting as a headline benefit of AI. Decision-makers will need to assess potential benefits, risks, probability of success, opportunity costs and trade-offs.

Collaboration and collective learning for co-design: There are gaps in understanding between AI developers and people (at all levels) who deliver and use public services. To improve a service, it is necessary to know how it is under-performing. AI developers know AI, but not what elements of

services it could most effectively support (highest need, risk, most potential added value, readiness of application). People who deliver public services know their challenges, but most will not have informed perspectives on how and where AI might potentially help. The public have views about how they would prefer services to be delivered. Trials could bring together these groups in one or two key sectors to explore what works in transferring learning between them, to progress to co-design applications of AI in services that are currently under-performing.

Appropriate uses of automation: It may be difficult to reach definitive general statements about where automation should be permitted, not least because that may change as technologies develop and are tested, and knowledge and skills develop. It may be more useful to identify types of process or decision where automation is unlikely to improve outcomes and should not be used, or where risk outweighs potential benefits, and why. Organisations should be clear and open about why they decide to automate any function or process.

There is some guidance on using AI in the public sector, on procuring AI services, and on using AI within procurement processes. Regulators have worked together on what AI could change within and across their activities. Professional bodies, including in the legal sectors, have developed guidance on AI and professional ethics and practices. Mapping principles across these would improve shared understanding of how to maintain governing principles in systems that use more automation.

Understanding relevant current use of AI: Clearly, better knowledge of where AI is already used in these domains, including in other countries, and in different domains that have similarities, would support decision-making. What are the emerging consequences of increased AI use in practice? Failed trials and deployments may be comparatively more difficult to find out about, unless they have become notorious.

There are mechanisms for reporting use of AI in the public sector and for decision-making in particular. The Algorithmic Transparency Recording Standard, which was developed following a public engagement study, will be mandated for central government bodies first, then the aspiration is to roll that out to all public bodies. To date only a few public authorities have used it. The specifications for ATRS, including what counts as using an algorithm, might be used in mapping.

GDPR requires reporting, but only when the decision is wholly automated, which allows loopholes and creates uncertainty. An assessment of how the EU AI Act works here would give useful context. There will be a Brussels effect, because companies will meet the AI Act. The UK might legislate to give UK citizens the same protections.

Global tracking of applications of AI in public services and public interest contexts could build an evidence base and improve understanding of what successful approaches are replicable between contexts, what characteristic challenges arise and how to address them, and how public AI assets can be grown over time.

Data management and trust: Better understanding of the importance of data assets held by the public sector and how they can be used to generate public value, and what kind of value, would help leaders make more confident and effective decisions. Without that understanding, decision-makers may find it difficult to set priorities and objectives, and assign resources. Research could explore what kind of knowledge about the value of data leaders need to make decisions about new uses of data assets.

Effective data sharing in these domains relies on public trust and transparent communication about data use. Data sharing needs to be trustworthy, or it will not be trusted for long. As data subjects, the public are increasingly aware that data about them creates value for technology companies. In relation to consumer digital services, they often accept that in exchange for using the service. They may be less likely to accept that in relation to social and civic services unless the benefits are mapped and realised.

Research could explore how users feel about their data in these systems:

What kind of value can be realised for it, and by whom: what is the data value chain?

What, indeed, is the value of data held about the public

How can non-profit and public bodies understand the value of the data they hold and turn it to public value?

Understanding legitimacy: These are social contexts, with interlinked expectations of treatment between users, staff and wider publics. Automation may undermine the expectation of shared expectations and social value in ways that are not yet fully understood. Context conditions agency: a patient in a medical consultation may well feel under pressure to consent to use of a technology

application, perhaps particularly when a clinician believes this will deliver better results and is advocating for it. Better understanding is needed of experiences of agency and consent in all these domains, and how legitimacy would be affected by increased use of automation. Lessons about user empowerment from previous stages of digital transformation could throw more light on trust and confidence.

Mapping could seek responses from professionals, users and others impacted, the key institutions and the general public on what kind of contexts (with related rights and ethics) would be more and less appropriate for more automation. There are process rights (to be heard, to challenge) and outcome rights (to be treated fairly), and automation may affect these in different ways.

These sectors have developed professional and social ethics, and work may be needed to further develop those to support well informed use of AI. In these organisations, control can be dispersed, often for strong reasons of oversight, balance and correction. Professionals working at different stages might need additional training to understand how AI was being used in other parts of the process and how that might affect their roles.

Digital identity: The public opposition to ID systems of 20+ years ago may not reflect where public sentiment is now. Smartphone users share their data many times a day to get different services. People are frustrated when they repeatedly have to give the same information, but data about a troubled family should not follow them in every context.

How ready is the UK for digital identity, or for how that could be enabling in interactions that use AI?

Should the UK develop a full digital public infrastructure with identity and data layers?

Public procurement: Public procurement rules and practice determine much of what it possible, and may limit the use of some AI tools. Procuring new AI services also comes with new uncertainty, and there is inevitably a lack of experience. In particular they should avoid being drawn to solutions that adopt what a technology provider offers, rather than what addresses an organisation's priorities.

There is research in this space (including by the Ada Lovelace Institute on local government). But the longstanding challenges of using public procurement strategically are amplified in the context of AI:

How can procurement practices and skills be updated?

Intellectual Property: There is uncertainty over the long-term stability and sustainability of licensing terms and service support for AI models and services. Arguably the responsible public body or third sector service provider should own a model that is developed with data about people who use the service, rather than the model being owned by a technology provider. Research could identify contracting models to enable public bodies to frame appropriate terms for building and improving dedicated models.

Audit and Evaluation: Audit should be central and essential from initiation of any application of AI in these contexts, to evaluate performance and outcomes for social and environmental impacts, trust and fairness, cost-efficiency and efficacy. Some jurisdictions and institutions will only work with providers that allow sight of code and functioning. It would be valuable to have a taxonomy of such conditions and clauses.

If more AI is applied, there is a case for broadening the skills base for evaluating it. The Food Standards Agency's regional food safety testing network, an association of public analysts, is a model for broadening the skills base for evaluation. A publicly subsidised network could be supported by the AI Safety Institute. Involving users in evaluation could improve it and support legitimacy. There are already AI Assurance toolkits (including a government one), but not much obligation on public bodies to use it. This could create a new commercial services sector, or open up new services to existing sectors including insurance.

Endnote

A lot of the development of AI is happening within a relatively small number of large companies with great resources and market power. However, as has been shown recently by the performance of smaller and open-source generative AI models, the future of AI is not fixed. AI could be developed and applied to new objectives that better support the majority of people to live free, healthy and satisfying lives in the future.

People expect to be treated fairly and equally by providers of public goods and services, and in that respect they hold these organisations to a different and higher standard than they do private companies. The challenge is to meet and nurture those expectations while gaining benefits from new insights and from automation. Currently, it is not clear that the technical community developing AI understand the limitations, restrictions and implications of designing for inclusive service provision and empowerment of users, rather than for profit. But that can change.

In these workshops we explored broad questions about how much control is handed over to private companies, and how to reclaim control to service public interests. We also explored more specific questions about how public bodies can make effective and informed decisions about new tools, and what resources and capabilities might need to be developed to fully make use of the potential of AI to serve public interests.

The workshops repeatedly generated opposition to any blanket assumption that applying AI across public challenges in the near future will be successful or deliver quick results. New solutions are needed to meet demand in all these domains, but a naïve belief that AI is “the” answer will be damaging. AI may support some functions more than others, and in some stages or even whole systems there may be no decisive case for using AI tools currently available. Some sources of friction in systems should be there, for safety, fair treatment, monitoring and evaluation.

Engagement with AI in public interest domains needs to be careful and critical, characterised by investigation, testing and learning, and proportionate aversion to risk. Decision-makers need to know what questions to ask, and what principles and capabilities are needed to engage with AI. Ideally, technology experts (internal and external) will work with decision-makers, operational staff and service users to establish virtuous circles of collective learning.

There is a good argument for looking outwards as well, because AI is happening in much of the world at once. There is a great deal of activity worldwide in experimentation with AI, some in public services and some in functions that public services could learn from. International outreach could

help collate and compare experiences, and grow communities for reporting and evaluation of public interest AI.

We hope this report will help the research community develop actionable projects to inform the use of AI for the benefit of society.