Committee on Standards in Public Life

Artificial Intelligence Seminar

Summary Note

Thursday 21 March 2024
2pm - 5:30pm

The Institution of Mechanical Engineers
One Birdcage Walk
London, SW1H 9JJ

Attendees

Committee on Standards in Public Life

- Doug Chalmers CB DSO OBE (Chair)
- Rt Hon Lady Arden of Heswall DBE (Independent member)
- Rt Hon Dame Margaret Beckett GBE MP (Labour Party member)
- Baroness (Simone) Finn (Conservative Party member)
- John Henderson CB (Independent member)
- Professor Gillian Peele (Independent member)
- Ewen Fergusson (Independent member)
- Professor Mark Philp (Chair, CSPL Research Advisory Board)

External Participants

- Stephen Almond, Executive Director for Regulatory Risk, Information Commissioner's Office
- Felicity Burch, Executive Director, Responsible Technology Adoption Unit, Department for Science, Innovation and Technology
- Mark Durkee, Head of Data and Technology, Responsible Technology Adoption Unit, Department for Science, Innovation and Technology
- Professor Edward Harcourt MBE, Professor of Philosophy, University of Oxford
- Elliot Jones, Senior Researcher on Foundation Models in the Public Sector, Ada Lovelace Institute
- Professor Neil Lawrence, DeepMind Professor of Machine Learning, University of Cambridge
- Professor David Leslie, Director of Ethics and Responsible Innovation Research, The Alan Turing Institute, and Professor of Ethics, Technology and Society, Queen Mary University
- Jessica Montgomery, Executive Director, AI@Cam, University of Cambridge
- Professor Gina Neff, Executive Director, Minderoo Centre for Technology and Democracy, University of Cambridge
- Mike Potter, Government Chief Digital Officer, Cabinet Office
- Helena Quinn, Principal Policy Adviser for AI and Data Science, Information Commissioner's Office
- Dr Jat Singh, Head of the Compliant and Accountable Systems Research Group, University of Cambridge
- Professor Karen Yeung, Interdisciplinary Professorial Fellow in Law, Ethics and Informatics, University of Birmingham

CSPL Secretariat

- Lesley Bainsfair, Head of Secretariat
- Peter Kelleher, Senior Policy Advisor
- Maggie O’Boyle, Press Officer

Introduction

In 2020, the Committee on Standards in Public Life (CSPL) published its report, *Artificial Intelligence and Public Standards*[1], looking at whether the then regulatory and governance framework for AI was sufficient to ensure that public standards would continue to be upheld as AI is adopted more widely across the public sector.

Since the report was published, we have seen extraordinary advancements in AI capability with the emergence of progressively sophisticated foundation models, which may already be in use in the public sector. Countries all over the world are grappling with how to regulate this fast-moving technology.

The purpose of the seminar was to hear from experts in the field on some of the issues raised in the 2020 report, in particular the assurances required to enable public office holders to be comfortably accountable for advice and decisions derived from, or made by, AI.

The CSPL is very grateful to all those who attended and were willing to give up their time and share their expertise with us.

The following summary note, prepared by the CSPL Secretariat, does not attribute comments or views to any particular individual or organisation.

Summary of discussion

Session 1: Alignment of AI tools with the public interest

Q1. The government's response to its AI White Paper defines "alignment" as "the process of ensuring an AI system's goals and behaviours are in line with human values and intentions". What does that mean and how is that achieved?

Q2. How do we assure ourselves and the public that AI in the public sector is 'aligned' with the right parameters and policies, is fair and delivers the outcomes it intends to, and checks, then revises outcomes if appropriate, over time?

The following points and themes emerged from this session's discussion.

Language

- It was important to think carefully about the term 'alignment', the terms we are using and how we describe what is going on, and the ability of our own language to exacerbate or attribute certain kinds of risks. Is the term 'alignment' too general; do we need to be more specific when considering a particular use of AI? 'Alignment' in AI terms has a specific technical meaning; there is a long tradition of technical AI alignment research, which means something quite specific. Perhaps a different term

[1] https://www.gov.uk/government/publications/artificial-intelligence-and-public-standards-report

should be used to describe the principle of aligning AI goals and behaviours with human values. There was no agreement on what that term or word might be.

Principles and values

- The question of principles was discussed. We already have principles rooted in the basis of democratic governance that ensure that public power should be used for proper purposes. The group felt that these are capable of being operationalised in relation to digital systems, including AI systems. The challenge is how we operationalise them in specific contexts with respect to specific systems that are in place. How do we flesh out the general principles to ensure proper accountability of the systems themselves so that they ultimately have accountability to the body politic.
- Some of those principles might need to be worked on to make them more applicable to civil servants' use of and encounters with AI. Leadership and good practice were needed so that these fundamental principles are embedded in the design of AI systems and not sacrificed to other principles of e.g. cost reduction or efficiency. That said, the principles will often need to be more focused and directed to particular cases. It is also clear that there is not always concordance between commercial principles and the interests of the wider public. There is, as a result, a need for 'toolboxes' of some kind to ensure desired principles and values are practically and transparently put in place.
- It may be helpful to think about how we align the values of the public sector with the goals of those who are designing, developing and deploying AI tools. It is a capacity building question of making sure that the public sector keeps a very specific and focused lens on the choices and challenges around this to ensure that the ethical agenda remains 'front and centre'.
- The role of CSPL was discussed - could CSPL think about principles of shared culture? People using these systems are the ones who are going to be pivoting between what those AI principles are, what their goals and aims are, what the law is, what the Nolan principles are, and the AI results that they are seeing over time. Making sure that the correct organisational culture is in place within public entities using AI, is critical for technical safety and security.

Risk of 'sleepwalking' into a world without human involvement in the process

- There was the risk of 'sleepwalking' into a world where machines are effectively making decisions. The automation of tasks means we could end up with a group of people who can no longer do these things manually themselves, and so are unable to spot when something is wrong and, therefore, are disabled from learning from mistakes. To prevent this 'sleepwalking', human involvement must be incorporated in steering or reviewing AI activity. Machines should not be making significant consequential decisions without there being some serious safeguards. It was noted that there is already quite a large body of regulations that prevent automated consequential decision-making for individuals, but it is not well known and, therefore, not well understood.

Assuring the public - pre and post AI design

- The question of assuring the public was discussed: how we set up AI - the ‘pre tool kit’ - the values and risk assessment, transparency; and then the ‘post tool kit’, considering audit, review, right to view and redress. To reassure the public, we need to prove that AI is beneficial. The point was made that the use of AI was heavily context specific, with some low risk and low impact areas; others where the AI might be quite simple in terms of e.g. risk assessment of benefit tools, but where the context in which it is applied means the impact risk is significant and the way in which it is governed, therefore, is key. When we talk about explainability, for example, understanding what is relevant in those contexts really matters.
- We need to evolve the thinking of public sector leaders and develop governance approaches for the changes in the way risk might manifest itself as AI tools are introduced. Research has shown that “people's appetite and for the acceptability of the use of AI tools, does tend to be heavily context-specific”.
- The public has consistently said for the past decade that AI should provide increases in efficiency and access; it should enhance public services; and reduce the burden on public officials, freeing them up for more human facing roles. The public does not want to see diminished authority, to see decision-making handed completely over to machines, services depersonalised, their privacy violated, or to feel at risk of replacement. Over time, there have been calls for a stronger regulatory response because there is a sense that AI is not being developed for the public’s benefit.
- Recent reviews, however, have suggested that people were not aware of the guidance already accessible and available tools, such as risk assessments, audit, recording standards and the like, were, therefore, not being used. There was a sense that we have not used the tools we already have to their best effect. We need to do more to raise awareness of existing solutions. We need a longer conversation about what education, or ‘convening’, is required, where it is most required, and in relation to what.

Transparency

- Transparency is a critical issue. It was suggested that government itself does not know all the ways AI is already being used, either by government or by their suppliers. What products are government buying that already have AI built in them? We need to consider the wider picture. We should ask not only “does this functionality work?” but “does this functionality work in that particular business process?”. An algorithm register, or rather a register of use cases, can help in instances where an AI product has been developed for a particular use, but then is reused to solve another problem.

Involving the public

- The point was made that we should be looking at allowing public conversation to occur, to enable government to get the best evidence. We need to be more demanding at the top level with regards to evidence and proof and testing, with collective redress and post-implementation vigilance.

No need to reinvent the wheel

- It was not necessarily the case that the wheel needed to be reinvented for AI systems but there are places where the risks of using AI are considerably greater, and there are challenges. This was not necessarily about creating something new, but thinking about how we deploy those techniques that we already know how to use for these specific AI contexts. There were examples from other areas, simple fixes, that would ensure that AI was used effectively.
- We have to think harder about how we create a world in which the end to end chains of responsibility remains a plausible ambition, given the uncertainties, due to the sources of data that will only increase with time.

Summary of this section

- *Firstly, we should think harder about alignment and whether that is the right thing we should be thinking about. It seems to be too general, and human values are fine as a logo on your family shield, but for a particular use of AI we need to be more specific.*
- *Secondly, we do have lots of good principles around, the Nolan principles amongst others, but some of those might need to be worked on in order to make them more applicable to civil servants' use of and encounters with AI. We might need to expect them to expect themselves to have a better grasp of the systems they use.*
- *There were, thirdly, noises about sleepwalking. It is clear that things can go wrong. It is clear that there is not always concordance between commercial principles and the interests of the wider public. It is clear that we need toolboxes of some kind to make sure those alignments are put together.*
- *Fourthly, we heard two separate sets of claims about the public. One is that they know a lot more than we think they know, and the other is they do not know where to go or what to do. Regarding comments about focus groups, one of the issues with focus groups is how much of what you are doing is educating people versus actually learning about what they really think. In this area, quite a lot of education gets done when you talk to people. You make them think about things they do not normally think about; you make them think about the relationships between principles and systems and so on. We need a longer conversation about what education is required, where it is most required, and in relation to what.*

- *To go back to the principles question, it is absolutely clear that this needs to be more focused and directed to particular cases. That seems to be the message across the board.*
- *The final point was the point about thinking much harder about the description of what is going on, and the ability of our own language to exacerbate certain kinds of risks and to give certain kinds of agency when that agency is, in fact, wholly attributive rather than actually there. Those points seem to be absolutely right and we should keep those in mind.*

Session 2: Responsibility in practice

Q1. In public life, the responsibility for any AI recommendation remains with the decision-maker, which means the relevant public office holder. How should this work in practice?

Q2. How do we monitor responsibility for AI advice when the decision-maker changes? What might institutional responsibility for AI systems look like?

The following points and themes were raised during this session.

Existing regulations

- It was noted that there exists already a body of excellent work done by academics, and others in the field, in identifying in quite concrete ways, how we can help to ensure responsible development and deployment.

Human involvement at all stages

- There is pre-vetting and post-vetting, but it is important that the middle stage, the development and the design phase, involves normative, human choices that are made by technical developers building the tool.

Government's procurement leveraging power

- Reference was made to the desirability of government making more use of its procurement power as leverage. The UK government is the biggest global buyer of software from Microsoft and Microsoft is currently the biggest provider of AI technology. Should procurement include riders such as 'If you cannot explain to us, you cannot sell to us', or, 'If you cannot give us these guarantees, we are not going to buy it'? A lot of smaller SMEs will not have that opportunity, so it is not straightforward, but arguably there is a unique opportunity here to push harder for better, from those higher up the supply chain and those involved in developing these complex systems.

Behaviours, values and accountability of AI professionals

- One of the problems suggested was that while many ‘traditional’ professions (such as doctors, accountants, lawyers) typically have an understanding of what they do not know, there is an unfortunate tendency for computer scientists to not know what they do not know and yet to deploy solutions under the premise of “try fast and fail” to improve. Computer scientists may be disconnected from the consequences of those solutions, in such a way that dramatic effects can occur. Should there be a duty of candour for people to be able to raise those things in a public domain and have them responded to? Should consideration be given as to whether that becomes a legal obligation in the future?
- The legal profession, the medical profession etc, have a very clear professional basis and shared ethical principles. There is an accountability system when dealing with e.g. doctors, who have been trained and have the skills; do the purchasers of AI need to have that kind of knowledge? The software development culture is different. In fact, rather than ‘first do no harm’, the basic principle is ‘move fast and break things’. This is deep in the character, motivation and attitude of software design and development. We have to try and work towards changing that ethic so that transparency of how these tools develop their advice is explainable.
- How do we maintain the vocational and professional knowledge bases in humans when they are relying on outcomes from machines? Tasks can be part-automated, someone who is a domain expert can work with these machines, and they do 80% of the work and the expert can fix the 20%. That is where we are at. AI tools cannot do 100% of the job. That last 20% is very difficult.
- One unintended consequence is that the use of AI will change the training pathway. The routine tasks that are now being done by AI, used to be the way that trainees in professions such as surgeons, the law and audit, ‘cut their teeth’ in the profession and developed intuition through experience. However, we will need to rely on people that have not had these opportunities in the future, to provide assurance that AI systems are working properly.

Governance regime

- Government is seeking to build an AI governance regime but it is not there yet. Part of the uncertainty of the next 10 years will be how effective that governance is and what safeguards might be needed in addition to our usual processes to avoid some of these unsafe products being used in unsafe ways.

Summary of this section

- *The most frightening question we heard people raise was, 'Do they know what they are doing?'. That speaks to an experience that most of us have had when engaging with software engineers and so on. Yes, they do know some of what they are doing, but they do not know all of it. If the ambition is end-to-end chains of human responsibility then we have problems. We have to think harder about how we create a world in which that is a plausible ambition, given that there are those kinds of uncertainties built in. The other problem is that not only do the software engineers not altogether know what they are doing, but certainly the people commissioning them do not know what they are doing. That means that there has to be upskilling and often very serious upskilling.*
- *The positive case is that we did it with pharma, we did it with the financial sector – apart from the occasional every-20-year crash of the market – we have done these kinds of things and we can do it in this. One worrying suggestion is that software engineers are just different. They do not have a professional body which sets ethical standards. They are attracted by destroying things to make things better, and that does seem to be a challenge.*
- *If we are not dealing with people like doctors, where you know they have had the training, they have the skills, there is an accountability system, do the purchasers, the people buying these kinds of products, need to be the doctors? Do they need to have that kind of knowledge skill? It seems that we are not in a good place in the current situation.*
- *There is a consistent ethical question that needs to be asked in these kinds of systems of holding people responsible, and that is whether they asked everybody affected. The 'everybody affected' principle can get you quite a long way, but somebody needs to keep asking that and seeing what comes out of it.*

Session 3: Accountability to the Public

Q1. How can public office holders be held to account for decisions based on advice derived from or provided by AI?

Q2. What other safeguards are needed to ensure the public continues to have trust and confidence in public decisions when AI is used in the decision-making process?

Opaque decision making

- It was noted that CSPL's current review was on accountability within public bodies, specifically trying to provide best practice and advice which will help organisations get better at holding themselves to account for the effective delivery of public services. The point was made that transparency was fundamentally important and yet there was a view that a lot of decision-making is opaque at the moment. Bringing machines into the decision making process was adding another layer of complexity and obscurity. This could actually make the whole process of trying to hold people accountable for decisions much more difficult.

Clarity about the use of AI

- It was important to understand when the technology is actually already being used. The work around the Algorithmic Transparency Recording Standard was welcomed but use of AI across public services is currently very opaque. It is very hard to understand when it is happening, which then makes it hard to apply some of the other tools that already exist in this space e.g. FOIA. What is needed strategically is something that brings together the legal, technical and cultural sides and adds that next level of strategic thinking and implementation to underpin the vision we have for AI in public services, and the vision we have for the future of the civil service.

Keeping humans in the loop

- Humans need to be kept in the loop to ensure accountability and redress for wrong decisions, but also to preserve a human relationship in a transaction: e.g. if a machine reaches the wrong decision, it cannot say 'sorry'; you cannot say 'thank you' if it gets it right. There is a whole ethical dimension to the relationship between the consumer and the service provider which is absent if the human is not in the loop, and which is a separate issue from the factual point about whether or not the machine-driven decisions are correct or incorrect. The challenge in 'building in' accountability is also around giving that "human-affecting principle" a voice at various stages: design, development, procurement, operation and review/redress.

- There is a risk that the people who have the authority to make the decisions about the system get further removed from the lived experience of the people who are directly affected. There is an important message, when we come to systems that are making consequential decisions about people, about having chains of accountability where we keep in mind the person about whom the decision is made.

Engaging the public from the start

- It is important to ensure that you are engaging the public when you decide to build a tool. Is it actually achieving what they want it to achieve? Is there proper continuous evaluation of it actually meeting those goals?

- (It was noted that AI can obfuscate the choices that are being made. Accountability for the disquiet about the A level results in 2020 was offloaded to the algorithm. However human decision makers were told what the algorithm would do and that choice was made. The AI algorithm worked fine but then the outcome was disliked and then 'thrown under the bus'. The algorithm was a distraction for the choice made.)

- The question of how to communicate with the public was raised. There may be people who know nothing about AI and do not really understand the terms or principles being used, and who may feel alienated by the process of introducing this more anonymous mechanism into the decision-making process. It was important to engage the public directly and simply but also in a more focused way. The need for a bigger piece on public education and engagement was discussed.

Challenges for accountability

- Challenges for accountability in the future were discussed. Some argued that explainability is a hard problem for generative AI systems given models' complexity, that tracking back, or tracing any access to patterns of bias or discrimination, will become increasingly difficult; others were more positive, taking wisdom from the mechanisms we use at a large scale to govern bureaucracy and hold it to account. The reason bureaucracy functions to the extent that it does is that people have tuned it. There have been changes, but over time it evolves slowly to adapt. The same can be true of AI.

- AI gives us the capacity to personalise which is very powerful. In many kinds of AI processes, there is the ability to give people individualised feedback; the information around which decisions were made can be personalised and AI can give reasons for decisions, e.g. why a particular benefit was denied; here are the reasons our systems have been set up to evaluate in general; here are the specific factors that were at play in your case. However, the decision to personalise and give proper reasons has to be built into the design so is a technical design decision.

- Building up skills within government was deemed as very important. This would involve improving the expertise and changing the culture of the civil service. Always contracting out and being beholden to the company which provides you with the maintenance of the AI tool was not seen as a good long-term solution for the public interest. There were processes in place, but it seems they are not joined up in quite the way that they might be. We have to think about what the expectations will be of public sector bodies in relation to these kinds of responsibilities.

Summary of this section

- *Two quite big things need to be expressed. One is that it is clear that the need to know when it is being used should drive quite a lot of activity. It sounds like we have various things in place, but they are not joined up in quite the way that they might be. The joining up seems to imply the existence of expertise in the civil service that is not necessarily a recognisable picture of the civil service that we currently have. We have to think about what the expectations will be of that body in relation to these kinds of responsibilities, and that seems to be something that we can address. The second point arises from a comment that although you do have accountability to avoid bad decisions, that is not the only reason you have accountability. That sounds like it is bringing back the idea of aligning AI with human values.*

- *We have some more detailed principles that we could appeal to there, such as the 'person-affecting' principle. If it affects somebody, they need to know that it is affecting them, how it is affecting them, etc. This aligns with the point that it connects with the idea of human accountability within a moral community. It is as a moral community that we respect the 'person-affecting' principle. The difficulty we have is that at the very least we have three systems, two of which we have been talking about. We have AI, bureaucracy, and the market. They are all different ways of* *thinking. They all shape our expectations, skilling and de-skilling us in equal measure. They all encourage us to be bad at worrying about the 'person-affecting' principle. 'The market is just the market. Bureaucracy. Well, that was the decision. We made the decision. AI. That is just how it works.' The challenge in building in accountability seems to be giving that human-affecting principle a voice at various stages in the process, both pre and post-mortem.*