# IIPP0009 assignment Candidate Number: HHYY7

*The November 2023 UK AI Safety Summit represents the opening of a "window of opportunity" to develop policy on AI safety in the UK. Identify the policy entrepreneurs and their roles leading up to this moment. What were the system boundaries implied by these policy entrepreneurs and their messaging? Critically reflect on this analysis.*

## 1 Introduction

The "AI Safety Summit" (hereafter: UKSummit), held at Bletchley Park in November 2023 brought together international leaders from government, industry, academia and civil society who were interested in the advancement and governance of artificial intelligence (AI). As was noted at the event, these "participants identified a narrow window for clear, decisive, and committed action" (UKGov, 2023). Drawing from Kingdon's (1993) Multiple Streams Approach (MSA), I will explore the nature of this "window" by first identifying the policy entrepreneurs and their roles leading up to the UKSummit and then describing how they exploited the opportunities that arose. I will widen the analysis to consider the boundaries implied by the messaging of policy entrepreneurs and the UKSummit literature itself. Finally, I will appraise the applicability of the MSA to understanding the events around the UKSummit and consider the potential for future windows of opportunity.

## 2 Framing the AI safety agenda

Kingdon (1993) used the phrase "open policy window" to describe a period of opportunity when policy on a specific issue may be brought to political attention for decision or enactment. In Kingdon's model, this moment is characterised as the converging of three separate streams of activity - the problem, the political, and the policy (solution). Such convergence occurs either when conditions have been sufficiently framed as a problem to be solved, or when political events offer space for a change of policy. Either opportunity may open the window for those *policy entrepreneurs*, with ready policy, to make their case.

For this analysis I have identified 4 groups of policy entrepreneurs: Ethics-pragmatists, Ethics-activists, X-riskers and AI-accelerators. These are characterised in Table 1 and their roles described in the following section.

### 2.1 The issue of AI safety

With a broad definition, which encompasses the prevention of a wide range of harms, AI safety tends to be expressed as two grades of problem. Firstly, concerns are expressed about the potential for the use of AI to inflict harm on individuals, groups, society or the physical and living world, such as exclusion, exposure and mis- or disinformation[1] (Arnold and Toner, 2021; Acemoglu, 2021). These latter problems are already being experienced as harms of automated systems (e.g. Eubanks, 2019; Leahey, 2022; Jemio *et al.*, 2022) and for this reason I will term this grade of problem "A-risk".

A-risk grade harms have been the main focus for a large and organised group of policy entrepreneurs, the "Ethics-pragmatists". These are mainly civic society organisations

[1] see Bird *et al.*, 2020; AIAAIC, 2023 for comprehensive lists of AI risks and harms

| | Ethics-pragmatists | Ethics-activists | X-riskers | AI-accelerators |
|---|---|---|---|---|
| Policy change | stronger governance of practice | decentralisation of power | more consideration of existential risk | no change |
| In scope | inclusion of wider perspectives; governance of data and practice | regulation; open-source; practitioner diversity | more research; regulation of "superhuman" AI | investment; scale first, fix later |
| Out of scope | narratives that associate too closely with Ethics-activists or X-riskers | narratives around X-risk | narratives that treat AI as just another technology | regulations, particularly on today's technology |
| Typically found | civic society organisations and nonprofits | campaigning and research nonprofits | technology corporations and universities | technology corporations, governments and universities |

Table 1: Characterisation of the groups of policy entrepreneurs associated with the UK-Summit

calling for more principled practice around AI (e.g. Omidyar Network and Institute for the Future, 2018; Open Data Institute, 2019; Leslie, 2019; Ada Lovelace Institute and DataKind UK, 2020; Patel *et al.*, 2021; Peppin, 2022; Davies, 2022).

With an initially similar focus on using practice to minimise harm (e.g. Mitchell *et al.*, 2019; Gebru *et al.*, 2020), the "Ethics-activists" emerged out of disintegrating[2] ethics labs in technology corporations to form independent organisations in which they could continue their research (e.g. DAIR) and remain outspoken about the harms caused by technology corporations. This disintegration seems to have been precipitated by the authoring of the now famous "Stochastic Parrots" paper (Bender *et al.*, 2021), which considered the specific harms of large language models on marginalised people and communities and remains highly pertinent today.

The second grade of AI safety problem are catastrophic events and existential "X-risk" threats. "X-riskers" are concerned with the loss of control of AI to itself (Hawking *et al.*, 2014; Müller and Bostrom, 2016) or to bad actors (NYT, 2015; Marcus, 2023; Bucknall and Dori-Hacohen, 2022), whilst others highlight the potential of today's AI to trigger catastrophe (Brundage *et al.*, 2018). They often call for technological solutions or regulation (London, 2017; Marcus, 2023), although similar to some Ethics-pragmatists there are calls for greater engagement of society in AI (Bucknall and Dori-Hacohen, 2022).

[2]Soon after the authoring of Bender *et al.* (2021), Google sacked one of the authors (Gebru, 2021), and over the following years ethics teams at both Google and Microsoft disintegrated (Ghaffary, 2021; Schiffer and Newton, 2023).

There is a further aspect of policy around AI, which has its own group of policy entrepreneurs. The "AI-accelerators" are keen to develop AI as fast as possible either for the good of humanity (e.g. Shah, 2022; Giannotti, 2023) or for profit.

## 2.2 Policy windows open

The changing agenda of UK government AI policy has resulted in the drawing in of different policy entrepreneurs at different times. Initially, AI-accelerators focused on growing UK AI industry (Hall and Pesenti, 2017; UKGov, 2017; UKGov, 2018), with AI safety an issue of perception, which could harm growth (AI Council, 2019). Gradually the agenda changed, influenced to a greater or lesser extent by Ethics-pragmatists[3]. Finally, the National AI Strategy - the UK government's "ten-year plan to make Britain a global AI super power" - was launched (Office for Artificial Intelligence, 2021). This noted both A-risks and X-risks of AI under its 3rd pillar of AI governance. Now firmly in need of policy on practical AI safety, government opened a policy window for Ethics-pragmatists to contribute on algorithmic and AI standards (CDDO, 2021; DCMS and OAI, 2022) and regulation of AI (CDEI, 2021; DRCF, 2022).

The release of ChatGPT in November 2022 (OpenAI, 2022), based on the GPT-series of foundation models[4], created something of a sensation and accelerated the flow of the AI safety problem and policy streams internationally. This "tremendous flurry of activity" (Kingdon, 1993, p41) signalled the opening of another window, this time due to the problem of AI safety becoming more pressing.

More dramatically in March 2023, a group of eminent AI researchers, industry leaders and X-riskers, called for a pause in giant AI experiments, warning about risks of mis- and dis-information, job losses, replacement of human intelligence and loss of control of civilisation (Bengio, Russell, *et al.*, 2023). Shortly afterwards, the Future of Life Institute published their policy recommendations for AI (FLI, 2023). At this moment, the X-risk framing of AI safety broke into the mainstream and the X-riskers were ready with their proposals.

The political stream was also gathering pace. In the UK ChatGPT caused a slight adjustment to policy[5]. Negotiations around the EU AI Act had been underway since 2020 and had recently become heated around the inclusion of general purpose AI (which therefore affected foundation models such as the GPTs). This was supported by, among others, X-riskers (Muller *et al.*, 2022; FLI, 2022) but opposed by AI-accelerators (Espinoza, 2022; Vincent, 2023; Ng, 2023a).The US was working on its own AI governance regime (NIST, 2023; The White House, 2023b; The White House, 2023a), for which it held its own AI summit (The White House, 2023c). The G7, China and others were also developing policy around AI governance (Foy and Pickard, 2023; Sheehan, 2023). Such international activity on AI meant that, for a country aiming to be a "global AI super power", the UK PM

[3]For instance, the emphasis on outcomes for individuals and society of two independent reports (CSPL, 2020; CDEI, 2020) was largely lost in the subsequent AI Roadmap, which continued to view regulation and standards as a means to build trust and thus grow AI markets (The AI Council, 2021).

[4]Foundation models, a general purpose AI, are models that can be deployed on a range of tasks: Jones (2023) is an excellent explainer.

[5]ChatGPT prompted a reassessment of the forthcoming regulatory framework for AI (AI Council, 2022)leading to policy recommendations for generative AI (Vallance, 2023). Unfortunately this review missed the regulatory insights about foundation models made the previous month by Küspert *et al.* (2023).

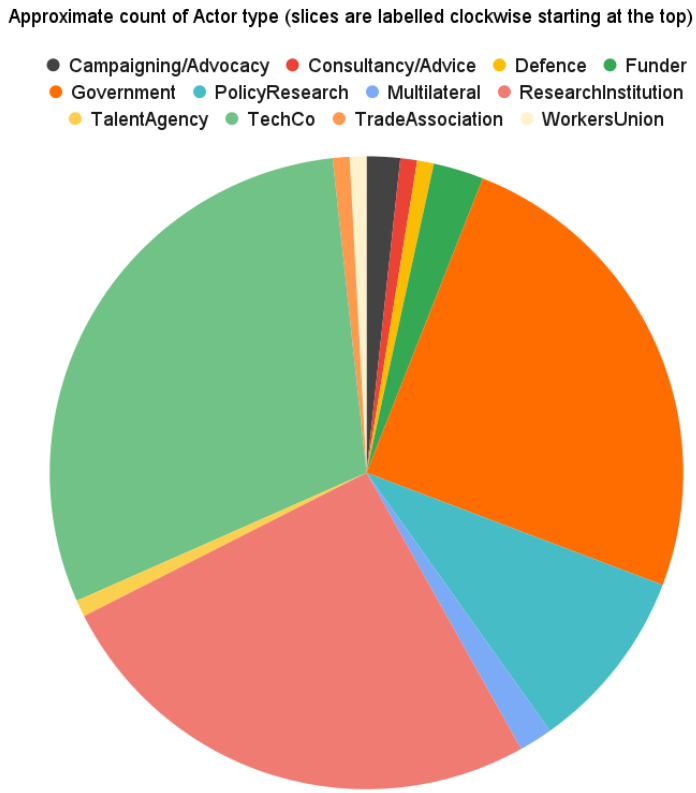


Figure 1: Approximate distribution of types or roles of attendees of the UKSummit.

needed a means to turn attention to AI activity in the UK[6]. The impact of the X-risk framing allowed Rishi Sunak to open the political window with the UKSummit.

### 2.3 Window frames

Attendees to the UKSummit (DSIT, 2023a) were mostly from technology corporations (around 30%), governments (around 25%), research institutions (around 25%) and policy research (around 10%) (see Figure 1). AI-accelerators, X-riskers and Ethics-pragmatists were all represented indicating that, in scope (Table 1) were:

- minimising regulation and maximising investment (AI-accelerators);
- maximising focus on "superhuman" AI and long-term risks (X-riskers); and
- increasing civic engagement and governance of data and practice (Ethics-pragmatists).

However, it is likely that the first two of these positions held sway owing to both the dominance of attendees with technology corporations, governments and universities, and their political influence over the host. Further, the agenda for the UKSummit was set in advance as "Misuse risks" and "Loss of control risks" (DSIT, 2023b), which aligns with the X-risk narrative. Additionally, the pre-summit literature focussed on "Frontier AI", a consciously moving target (DSIT, 2023c), and this aligns with both X-risk and AI-accelerator narratives, pushing out of scope the regulation of *current* AI systems. Figures 2

[6]It is also likely that domestic political conditions within government and regarding the forthcoming general election added further motivation for a demonstration of statecraft.

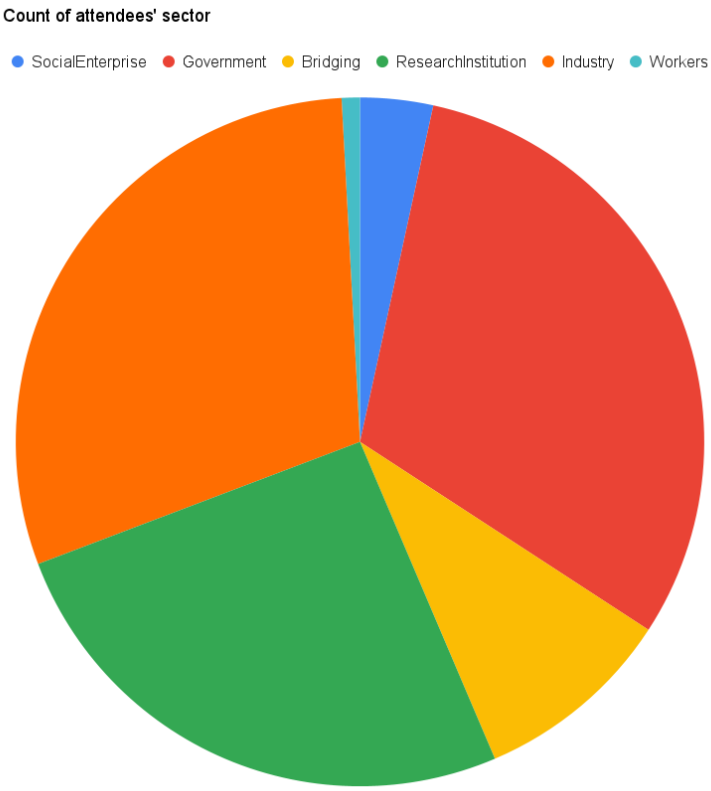


Figure 2: Approximate sector distribution of attendees to the UKSummit.

and 3 illustrate the lack of representation of civil society and workers and, with the majority of "International" attendees from technology corporations, much of the global south is excluded. This parallels observations about AI and technology (Bender, 2023; Rikap and Lundvall, 2022; Amuasi, 2022).

These boundaries did not go unnoticed. Following the announcement of the UKSummit, Ethics-pragmatists called for a greater focus on A-risks and engagement of policymakers with the public and affected people (e.g. Connected by Data, 2023; Glenster and Gilbert, 2023; Chatham House, 2023). In a rare moment of concord, both Ethics-activists and *some* AI-accelerators criticised the X-risk framing as being motivated by a fear of regulation (Paul, 2023; Ng, 2023a).

## 3 Discussion

Kingdon's characterisation has enabled the isolation of components of a complicated policy landscape. In particular, the characterisation of the two types of policy window (problem and political) are extremely pertinent in this case, with the first (roughly: "ChatGPT") precipitating the second ("UKSummit"). Also, the role of policy entrepreneurs developing the narrative around "their pet proposals or ... problems" (Kingdon, 1993, p44) and exploiting open windows is germane to the X-risk framing.

This theory emphasises the importance of timing for policy change, which points to a real-world missed opportunity: If "Stochastic Parrots" (Bender *et al.*, 2021) - with its focus

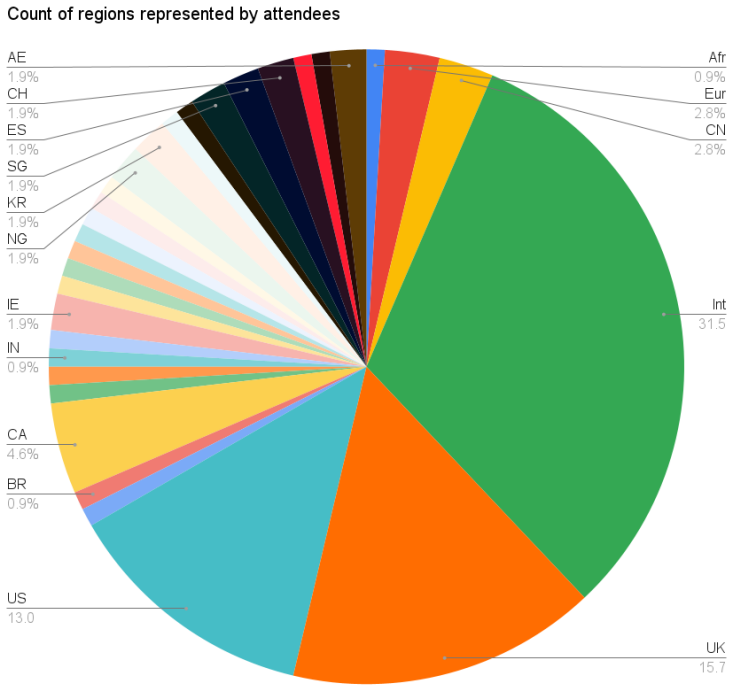


Figure 3: Approximate representation of regions by attendees of the UKSummit. ISO 3166 country codes are used. In addition, organisations that represent multiple countries are labelled "Int" (international), "Afr" (Africa) and "Eur" (Europe).

on models like the GPTs - had been published much closer to the opening of the window around ChatGPT, would the dominant problem framing have been very different? Would the UKSummit have discussed the environmental and social costs of the AI that we have today, rather than focusing on the future threats of superhuman AI? Unfortunately, this seems rather unlikely, because of the nature of the political influence that some actors possessed, as noted in Section 2.3. This is a weakness in the MSA, which proposes that power is in the idea. The MSA thus fails to represent how policy entrepreneurs can struggle to gain traction for good ideas.

The reality of policy entrepreneurs seems much messier than in Kingdon's rendering. The MSA has no sense of the divisions or coalitions between entrepreneurs, or how they change with opportunities. Yet these manifestly exist. The consolidation of the players into groups in this paper is a simplified view of the many competing policy positions and narratives around the UKSummit. Some of these are much less cohesive than "group" implies (particularly X-riskers, as described in Ng (2023a) and Johnson (2023), and AI-accelerators) and additionally there is overlap between the narratives of all the groups. A better approach to understanding the different groups' positions on AI safety would have been frameworks that focus more on the formation of coalitions, such as Dowding (2018) and Hajer (2005). In particular the latter work, with its emphasis on discourse, is highly relevant as narratives have proved central to much of the debate around AI safety and the UKSummit.

An example of the centrality of narrative to the AI safety issue is the messaging of both X-riskers and AI-accelerators that AI is fundamentally different to other technologies

(Colomé, 2023). Ethics-activists have suggested that this is a deliberate use of narrative to separate the ethical and safety issues of AI from those who create them (Gebru, 2022, 5 mins; Gebru, 2023, 8 mins 40). Bender suggests the replacement of "AI" with the term "automation" because this exposes the nature of the technology, who is impacted and how (Bender, 2023). Ever the Ethics-pragmatists, Ada Lovelace Institute demonstrated how we regulate a wide range of unique activities from which lessons can be learned for regulating AI safety (Smakman *et al.*, 2023). Interestingly, I have identified much less critique of the narratives of either the Ethics-activists (e.g. Ng, 2023b) or the Ethics-pragmatists. This suggests a conscious or unconscious lack of engagement with these narratives, and possibly a greater focus on maintaining the X-risk or AI-uniqueness messaging.

Another limitation of the MSA is that it offers little in terms of how best to exploit policy windows, beyond readiness and persistence. Both the ChatGPT and UKSummit policy windows have seen a great flurry of policy activity from Ethics-pragmatists (e.g. Davies and Birtwistle, 2023; Connected by Data, 2023; Glenster and Gilbert, 2023; Smakman *et al.*, 2023) and yet these groups remained at the Fringe of the UKSummit. An option is to embrace the X-risk narrative into a practical A-risk framework. For instance, Mitchell (2023) centres regulations on rights, demonstrating that the right to existence serves X-risk discourse.

Considering the narratives of all the policy entrepreneurs and the boundaries defined by the messaging around, and attendees of, the UKSummit, I note a potentially under-utilised leverage point. This is to better engage with technology workers, a group that will grow over coming years. There can be an assumption that the presence of industry representatives "in the room" is also representative of workers in that industry. However, as the Ethics-activists have demonstrated, work even in AI can be precarious and alienating. Engagement is needed both in terms of sharing understanding of principles and practice of safer AI (Kelley, 2022), but to also understanding the experiences of workers in AI (Strümke *et al.*, 2021; AI Now Institute, 2023). Also, tooled with the right narratives and skills, this group is well positioned to help exploit future AI safety policy windows.

## 4 Conclusion

This endeavour to digest the "policy primeval soup" related to AI safety in UK identified 3 windows of opportunity and 4 groups of policy entrepreneurs attempting to exploit them. Messaging was key to the positions of policy entrepreneurs, with debate focusing on the narratives used by different groups. A simple inference of the boundaries to the discourse "in the room" of the UKSummit suggested that regulation of current AI systems is unlikely to have been a prominent topic at the event. There is solid policy work going on behind the scenes both within government and across a diverse landscape of organisations. With the pace of AI developments not yet slowing, it is likely that new policy windows will be opening up for exploitation soon.

# Bibliography


Acemoglu, Daron (2021). *Harms of AI. NBER Working Paper No. 29247*. National Bureau of Economic Research. DOI: 10.3386/w29247. eprint: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.nber.org/system/files/working_papers/w29247/w29247.pdf.

Ada Lovelace Institute and DataKind UK (2020). *Examining the Black Box: Tools for assessing algorithmic systems*. Ada Lovelace Institute and DataKind UK. eprint: https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf.

AI Council (2019). *Artificial Intelligence Council meeting 9th September 2019*. Meeting minutes. Office for Artificial Intelligence, Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. eprint: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/836901/AI_Council_Meeting_Minutes_9_September_2019.pdf.

– (2022). *Artificial Intelligence Council meeting 6th December 2022*. Meeting minutes. Office for Artificial Intelligence, Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. eprint: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1126431/AI_Council_Meeting_Minutes_06_Dec_2022__3_.pdf.

AI Now Institute (2023). *Spotlight: Tech and Financial Capital*. AI Now Institute.

AIAAIC (2023). *AI, Algorithmic, and Automation Incidents and Controversies: Classifications and definitions*. URL: https://www.aiaaic.org/aiaaic-repository/classifications-and-definitions (visited on 12/11/2023).

Amuasi, John (2022). "Representation Learning in the Global South: Societal Considerations-Fairness, Safety and Privacy". In: *International Conference on Learning Representations*.

Arnold, Zachary and Helen Toner (2021). *AI Accidents: An Emerging Threat - What Could Happen and What to Do*. Analysis. The Center for Security and Emerging Technology. eprint: https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf.

Bender, Emily M. (2023). *Opening remarks on "AI in the Workplace: New Crisis or Longstanding Challenge"*. URL: https://medium.com/@emilymenonbender/opening-remarks-on-ai-in-the-workplace-new-crisis-or-longstanding-challenge-eb81d1bee9f.

Bender, Emily M. *et al.* (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.

Bengio, Yoshua, Stuart Russell, *et al.* (2023). *Pause Giant AI Experiments: An Open Letter*. eprint: https://futureoflife.org/fli_pause-giant-ai-experiments_an-open-letter/. URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Bird, Eleanor *et al.* (2020). *The ethics of artificial intelligence. Issues and initiatives*. Tech. rep. Panel for the Future of Science and Technology, European Parliamentary Research Service.

Brundage, Miles *et al.* (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation and OpenAI.

Bucknall, Benjamin S. and Shiri Dori-Hacohen (2022). "Current and Near-Term AI as a Potential Existential Risk Factor". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '22. ACM. DOI: 10.1145/3514094.3534146.

CDDO (2021). *UK government publishes pioneering standard for algorithmic transparency*. Central Digital and Data Office. URL: https://www.gov.uk/government/news/uk-government-publishes-pioneering-standard-for-algorithmic-transparency (visited on 12/11/2023).

CDEI (2020). *Review into bias in algorithmic decision-making*. Independent report. Centre for Data Ethics and Innovation. eprint: https://assets.publishing.service.gov.uk/media/60142096d3bf7f70ba377b20/Review_into_bias_in_algorithmic_decision-making.pdf.

– (2021). *The roadmap to an effective AI assurance ecosystem*. Independence report. Centre for Data Ethics and Innovation. eprint: https://assets.publishing.service.gov.uk/media/61b0746b8fa8f50379269eb3/The_roadmap_to_an_effective_AI_assurance_ecosystem.pdf.

Chatham House (2023). *UK AI Summit: What can it achieve?* URL: https://www.chathamhouse.org/events/all/members-event/uk-ai-summit-what-can-it-achieve (visited on 12/10/2023).

Colomé, Jordi Pérez (2023). *Why are the people who pushed for artificial intelligence now signing so many doomsday manifestos?* URL: https://english.elpais.com/science-tech/2023-06-03/why-are-the-people-who-pushed-for-artificial-intelligence-now-signing-so-many-doomsday-manifestos.html (visited on 06/06/2023).

Connected by Data (2023). *AI Safety Summit: Open Letter to the UK Prime Minister*. Connected by Data, The Trades Union Congress and Open Rights Group. URL: https://ai-summit-open-letter.info/ (visited on 12/10/2023).

CSPL (2020). *Artificial Intelligence and Public Standards*. Review. Committee on Standards in Public Life. eprint: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF.

Davies, Matt and Michael Birtwistle (2023). *Seizing the 'AI moment': making a success of the AI Safety Summit*. URL: https://www.adalovelaceinstitute.org/blog/ai-safety-summit/ (visited on 12/12/2023).

Davies, Tim (2022). *Participation pathways - designing for effective engagement*. URL: https://connectedbydata.org/blog/2022/11/21/pathways-of-participation (visited on 12/11/2023).

DCMS and OAI (2022). *New UK initiative to shape global standards for Artificial Intelligence*. Department for Digital, Culture, Media & Sport, Office for Artificial Intelligence, URL: https://www.gov.uk/government/news/new-uk-initiative-to-shape-global-standards-for-artificial-intelligence (visited on 12/11/2023).

Dowding, Keith (2018). "The Advocacy Coalition Framework". In: *Handbook on Policy, Process and Governing*. Chap. 13, pp. 220–231. DOI: 10.4337/9781784714871.00020.

DRCF (2022). *The benefits and harms of algorithms: a shared perspective from the four digital regulators*. Research and Analysis. Digital Regulation Cooperation Forum.

DSIT (2023a). *AI Safety Summit: confirmed attendees (governments and organisations)*. URL: https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-confirmed-governments-and-organisations (visited on 12/11/2023).

– (2023b). *AI Safety Summit: introduction*. URL: https://www.gov.uk/government/publications/ai-safety-summit-introduction (visited on 12/10/2023).

DSIT (2023c). *Frontier AI: capabilities and risks*. Discussion Paper. Department for Science, Innovation and Technology.
Espinoza, Javier (2022). "Google in last-ditch lobbying attempt to influence incoming EU tech rules". In: *Financial Times*.
Eubanks, Virginia (2019). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*.
FLI (2022). *General Purpose AI and the AI Act*. Future of Life Institute. eprint: https://artificialintelligenceact.eu/wp-content/uploads/2022/05/General-Purpose-AI-and-the-AI-Act.pdf.
– (2023). *Policymaking in the Pause. What can policymakers do now to combat risks from advanced AI systems?* Future of Life Institute.
Foy, Henry and Jim Pickard (2023). "G7 leaders call for 'guardrails' on development of artificial intelligence". In: *Financial Times*.
Gebru, Timnit (2021). *Moving beyond the fairness rhetoric in machine learning*. URL: https://iclr.cc/virtual/2021/invited-talk/3718 (visited on 12/12/2023).
– (2022). *Timnit Gebru talks to Al Jazeera*. URL: https://www.aljazeera.com/program/talk-to-al-jazeera/2022/8/6/timnit-gebru-is-ai-racist-and-antidemocratic (visited on 12/12/2023).
– (2023). *How to make AI systems more just with Hilary Pennington and Dr. Timnit Gebru*. URL: https://www.youtube.com/watch?v=MgcUatmPnyE (visited on 12/12/2023).
Gebru, Timnit *et al.* (2020). "Datasheets for Datasets". In: arXiv: 1803.09010 [cs.DB].
Ghaffary, Shirin (2021). *Google says it's committed to ethical AI research. Its ethical AI team isn't so sure*. URL: https://www.vox.com/recode/22465301/google-ethical-ai-timnit-gebru-research-alex-hanna-jeff-dean-marian-croak (visited on 12/09/2023).
Giannotti, Livia (2023). "Who is Sam Altman, the man who co-founded OpenAI "for the benefit of humanity"?" In: *Techmonitor*.
Glenster, Ann Kristin and Sam Gilbert (2023). *Policy Brief: Generative AI*. Minderoo Centre for Technology, Democracy, and Bennett Institute for Public Policy.
Hajer, Maarten A. (2005). "Coalitions, practices, and meaning in environmental politics: From acidrain to BSE". In: *Discourse theory in European politics: Identity, policy and governance*. Palgrave Macmillan UK, pp. 297–315.
Hall, Dame Wendy and Jérôme Pesenti (2017). *Growing the artificial intelligence industry in the UK*. eprint: https://assets.publishing.service.gov.uk/media/5a824465e5274a2e87dc2079/Growing_the_artificial_intelligence_industry_in_the_UK.pdf.
Hawking, Stephen *et al.* (2014). "Transcending Complacency on Superintelligent Machines". In: *HuffPost*.
Jemio, Diego *et al.* (2022). *The Case of the Creepy Algorithm That 'Predicted' Teen Pregnancy*. URL: https://www.wired.com/story/argentina-algorithms-pregnancy-prediction/.
Johnson, Rebecca (2023). *X-riskers think differently*. URL: https://ethicsgenai.com/x-riskers-think-differently/ (visited on 12/12/2023).
Jones, Elliot (2023). *Explainer: What is a foundation model?* Ada Lovelace Institute. URL: https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.
Kelley, Stephanie (2022). "Employee Perceptions of the Effective Adoption of AI Principles". In: *Journal of Business Ethics*.
Kingdon, John W. (1993). "How do issues get on public policy agendas?" In: *Sociology and the Public Agenda*. Vol. 8. 1. Chap. 3, pp. 40–53. DOI: 10.4135/9781483325484.

Küspert, Sabrina *et al.* (2023). *The value chain of general-purpose AI. A closer look at the implications of API and open-source accessible GPAI for the EU AI Act*. Ada Lovelace Institute. URL: https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/.

Leahey, Andrew (2022). *We Can All Learn a Thing or Two From the Dutch AI Tax Scandal*. URL: https://news.bloombergtax.com/tax-insights-and-commentary/we-can-all-learn-a-thing-or-two-from-the-dutch-ai-tax-scandal.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. Tech. rep. The Alan Turing Institute. DOI: https://doi.org/10.5281/zenodo.3240529.

London, Andrew (2017). "Elon Musk warns 'AI is a fundamental risk to the existence of human civilization'". In: *Techradar*.

Marcus, Gary (2023). *Why Are We Letting the AI Crisis Just Happen?* URL: https://www.theatlantic.com/technology/archive/2023/03/ai-chatbots-large-language-model-misinformation/673376/ (visited on 04/19/2023).

Mitchell, Margaret (2023). *The Pillars of a Rights-Based Approach to AI Development*. URL: https://www.techpolicy.press/the-pillars-of-a-rightsbased-approach-to-ai-development/ (visited on 12/09/2023).

Mitchell, Margaret *et al.* (2019). "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. ISBN: 9781450361255. DOI: 10.1145/3287560.3287596.

Muller, Catelijne *et al.* (2022). *AIA in-depth 1 Objective Scope Definition*. ALLAI. eprint: https://allai.nl/wp-content/uploads/2022/03/AIA-in-depth-Objective-Scope-and-Definition.pdf.

Müller, Vincent C. and Nick Bostrom (2016). "Future Progress in Artificial Intelligence: A Survey of Expert Opinion". In: *Fundamental Issues of Artificial Intelligence*. Ed. by Vincent C. Müller. Cham: Springer International Publishing, pp. 555–572. ISBN: 978-3-319-26485-1. DOI: 10.1007/978-3-319-26485-1_33.

Ng, Andrew (2023a). "[Editorial]". In: *The Batch*.

– (2023b). *[Editorial]*. URL: https://www.deeplearning.ai/the-batch/issue-209/.

NIST (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. DOI: 10.6028/NIST.AI.100-1. eprint: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

NYT (2015). "Artificial intelligence is marvelous – unless it's part of a killer robot: Some experts worry that the technology is dangerous as well as marvelous. Think killer robot". English. In: *The Washington Post (Online)*. Copyright - (Copyright The Washington Post Company, Feb 2, 2015; People - Musk, Elon; Last updated - 2021-04-22.

Office for Artificial Intelligence (2021). *National AI Strategy*. Guidance. Department for Science, Innovation, Technology, Office for Artificial Intelligence, Department for Digital, Culture, Media & Sport, and Department for Business, Energy & Industrial Strategy. eprint: https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National_AI_Strategy_-_PDF_version.pdf.

Omidyar Network and Institute for the Future (2018). *Ethical OS Toolkit*. URL: https://ethicalos.org/.

Open Data Institute (2019). *ODI Data Ethics Canvas*. URL: https://theodi.org/article/data-ethics-canvas (visited on 02/04/2021).

OpenAI (2022). *Introducing ChatGPT*. URL: https://openai.com/blog/chatgpt (visited on 12/11/2023).

Patel, Reema *et al.* (2021). *Participatory data stewardship: A framework for involving people in the use of data*. Ada Lovelace Institute. eprint: https://www.adalovelaceinstitute.org/wp-content/uploads/2021/11/ADA_Participatory-Data-Stewardship.pdf.

Paul, Kari (2023). *Letter signed by Elon Musk demanding AI research pause sparks controversy*. URL: https://www.theguardian.com/technology/2023/mar/31/ai-research-pause-elon-musk-chatgpt.

Peppin, Aidan (2022). *The rule of trust: Findings from citizens' juries on the good governance of data in pandemics.* Ada Lovelace Institute.

Rikap, Cecilia and Bengt-Åke Lundvall (2022). "Big tech, knowledge predation and the implications for development". In: *Innovation and Development* 12.3, pp. 389–416. DOI: 10.1080/2157930X.2020.1855825.

Schiffer, Zoe and Casey Newton (2023). *Microsoft lays off team that taught employees how to make AI tools responsibly*. URL: https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs.

Shah, Agam (2022). "AI pioneer suggests trickle-down approach to machine learning". In: *The Register*.

Sheehan, Matt (2023). *China's AI Regulations and How They Get Made*. URL: https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117 (visited on 12/10/2023).

Smakman, Julia *et al.* (2023). *Mission critical: Lessons from relevant sectors for AI safety*. Policy Briefing. Ada Lovelace Institute.

Strümke, I. *et al.* (2021). "The social dilemma in artificial intelligence development and why we have to solve it". In: *AI Ethics*. DOI: https://doi.org/10.1007/s43681-021-00120-w.

The AI Council (2021). *AI Roadmap*. Office for Artificial Intelligence, Department for Business, Energy & Industrial Strategy, and Department for Digital, Culture, Media & Sport. eprint: https://assets.publishing.service.gov.uk/media/5ff3bc6e8fa8f53b76ccee23/AI_Council_AI_Roadmap.pdf.

The White House (2023a). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (visited on 12/10/2023).

– (2023b). *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/ (visited on 12/11/2023).

– (2023c). *Readout of White House Meeting with CEOs on Advancing Responsible Artificial Intelligence Innovation*. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/readout-of-white-house-meeting-with-ceos-on-advancing-responsible-artificial-intelligence-innovation/ (visited on 12/10/2023).

UKGov (2017). *Industrial Strategy Building a Britain fit for the future*. Industrial Strategy White Paper. UK Government. eprint: https://assets.publishing.service.gov.uk/media/5b5afeffe5274a3fd124c9ba/industrial-strategy-white-paper-web-ready-a4-version.pdf.

– (2018). *AI Sector Deal*. Department for Science, Innovation et al. eprint: https://assets.publishing.service.gov.uk/media/5ae0f342e5274a0d85c1c6d5/180425_BEIS_AI_Sector_Deal__4_.pdf.

– (2023). *Chair's Summary of the AI Safety Summit 2023, Bletchley Park*. UK Government.

Vallance, Patrick (2023). *Pro-innovation Regulation of Technologies Review: Digital Technologies*. HM Government.

Vincent, James (2023). "OpenAI says it could 'cease operating' in the EU if it can't comply with future regulation". In: *The Verge*.