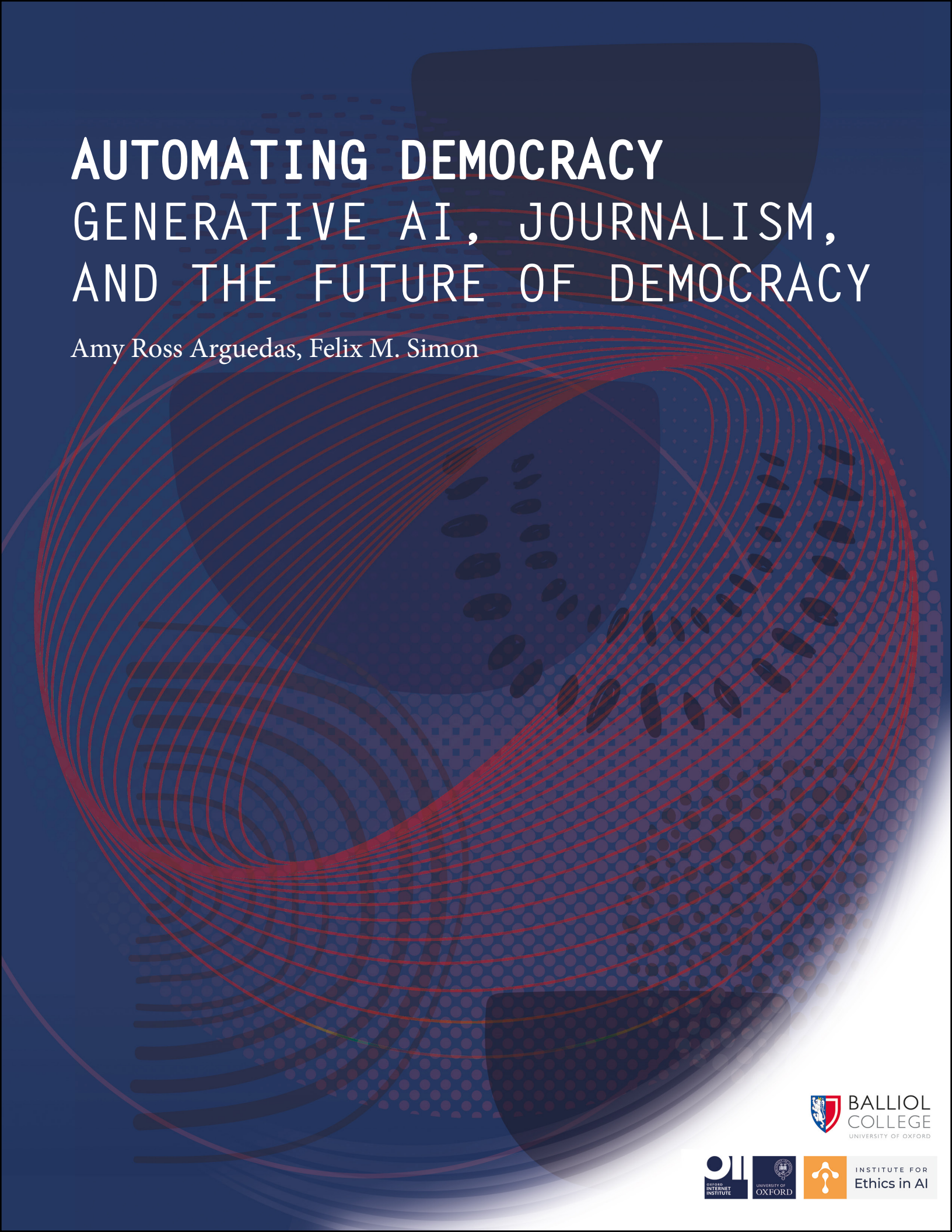
AUTOMATING DEMOCRACY
GENERATIVE AI, JOURNALISM, AND THE FUTURE OF DEMOCRACY

Amy Ross Arguedas, Felix M. Simon



BALLIOL COLLEGE
UNIVERSITY OF OXFORD

OXFORD INTERNET INSTITUTE
UNIVERSITY OF OXFORD
INSTITUTE FOR Ethics in AI

TABLE OF CONTENTS

This report has been funded by the Balliol Interdisciplinary Institute (BII) and received support from Balliol College, the Institute for Ethics in AI, and the Oxford Internet Institute.

Keywords: Artificial Intelligence, AI, LLMs, Generative AI, News, Journalism, Democracy, AI Governance

All graphics in this report were created with Open AI's Dall-E 2

EXECUTIVE SUMMARY

- **'Generative AI' is an umbrella term used for AI systems that can generate new forms of data**. Often by applying variations of machine learning to large quantities of training data. This output can be multimodal and include text, visuals, and audio. Often, this is done indistinguishable from content created by other, or human, means. Large Language Models (LLMs) are the most prominent form of generative AIs.

- **The creation, implementation, and regulation of LLMs raises important questions about power and inequalities.** Power is central to decisions about whose worldviews are included and prioritised in models, who gets to use and capitalise on these technologies, and who gets to decide how they are regulated. Helping counter inequalities and bias requires careful and proactive strategies from regulators and companies at all stages.

- **The application of LLMs developed largely in the Global North to other contexts poses important technical, ethical, and legal challenges.** To the extent that the data sets on which LLMs are trained disproportionately represent some worldviews above others, their outputs may fail to provide adequate sociocultural sensitivity and specificity in different contexts. This has important implications not only for how different populations perceive these technologies or may be impacted by them but also raises important questions around national sovereignty, especially to the extent that countries become dependent on technologies created by private companies in a handful of countries.

- **The potential of personalisation through LLMs can be a double-edged sword.** Generative AI could potentially be a real game changer for accessibility, tailoring outputs to the needs of individual users. Yet the ability to create extremist LLMs – which has already been reported – or more generally tailor responses to individual preferences could limit users to output that only reinforces their world views, which reanimates ongoing debates about new technologies and echo chambers, including how they may impact democratic dialogue.

- **While automation has been used for news recommendation and distribution for some time, media organisations are increasingly experimenting with it for news production.** Large organisations such as the BBC have been investing in building models based on their own data sets, even as most always keep a human in the

loop. Understanding how audiences feel about the use of AI in newsrooms, and communicating these uses transparently, will likely be important for rolling out new technologies without damaging trust.

- **Generative AI creates opportunities and risks for news organisations.** An optimistic approach suggests generative AI can help journalism more effectively fulfil its duty of informing the public in a way that is more personalised, relevant, accessible, and interesting. Yet, becoming overly reliant on these tools also carries important risks, especially for organisations who depend on tools created by others, and which they may not fully understand.

- **The regulation of generative AI needs to be done in a manner that is both democratic and global in scale.** Leaving decisions about how to regulate AI in hands of private tech companies and regulators in the Global North carries some real risks. Finding ways to include the public in the deliberation and decision-making around how to implement and regulate these technologies – at local, national, and international levels – will be crucial to achieving a truly democratic outcome.

- **Looking to the past can help us understand the present and prepare for the future.** While hype around tools like ChatGPT4 and DALL-E can make them seem entirely novel and transformative, it is important to recognise they are part of an incremental process. Looking to examples of past technological innovation can help put the hype around generative AI in perspective, while also helping us anticipate future challenges and how we can address them more effectively.

INTRODUCTION

Sophisticated AI systems are increasingly everywhere. In many ways, we have already been affected by the rollout of AI systems into more and more areas of life, from insurance and law to healthcare and the media – often without really noticing. However, 2023 will likely prove to be a particularly critical moment in the history of AI. Ever since the public release of ChatGPT, a so-called Large Language Model (LLM), in December 2022 by the US start-up OpenAI, we are witnessing a proliferation of a form of AI that has been labelled 'generative AI' due to the ability of these systems to create seemingly everything from realistic text to images. ChatGPT reached 100 million users in just two months and has now been built into Microsoft's Bing search engine. Various applications rely on the system, which is increasingly integrated into other software, too. Meanwhile, the 'AI race' is heating up, with Google releasing its own chatbot and other technology companies vying to get a piece of the cake by building and releasing their own models.

Powerful and technologically impressive as some of these developments are, they also raise important questions about their democratic impact. Up until now, we could take for granted humans' central role in shaping democratic deliberation and culture. But what does it mean for the future of democracy, if humans are increasingly side-lined by AI? Does it matter if news articles, policy briefs, lobbying pieces, and entertainment are no longer created solely by humans? How will an increasingly automated journalism and media culture affect democratic participation and deliberation? How can we protect democratic values, like public deliberation and self-governance, in societies which stand to be reshaped through AI? And how might these new technologies be used to promote democratic values?

To investigate this situation and to gauge the opinions of experts and academics, the Balliol Interdisciplinary Institute project 'Automating Democracy: Generative AI, Journalism, and the Future of Democracy' convened a group of experts for a public symposium at Balliol College Oxford, in collaboration with the Institute for Ethics in AI and the Oxford Internet Institute.[1] The aim of the symposium, organised jointly by Dr Linda Eggert,[2] an Early Career Fellow in Philosophy, and Felix M. Simon,[3] a communication researcher and DPhil student at the Oxford Internet Institute, was to identify key issues in this space and start a conversation among academics, industry experts, and the public about the questions outlined above. The symposium featured three panel discussions on 'The Technology, Context, and Socioeconomics of LLMs,' 'How Generative AI is Impacting the News Media,' and on 'Regulating Generative AI Democratically and Globally.'

---

[1] Simon, F. M. (2023, June 17). "No Technology Is Going To Take Over All Of Society": Exploring LLMs, Generative AI, and their Societal Implications. Medium. Retrieved from https://felixsimon.medium.com/no-technology-is-going-to-take-over-all-of-society-c70efd017b9d

[2] https://www.oxford-aiethics.ox.ac.uk/linda-eggert

[3] https://www.oii.ox.ac.uk/people/profiles/felix-simon/

Speakers included leading experts on AI, the news, and democratic theory: Hannah Kirk, an AI researcher and DPhil student at the Oxford Internet Institute; Hal Hodson, a special projects writer and technology journalist at The Economist; Laura Ellis, the BBC's Head of Technology Forecasting; Gary Rogers, co-founder of news agency RADAR and Senior Newsroom Strategy Consultant at Fathm; Dr Gemma Newlands, Departmental Research Lecturer in AI and Work at the Oxford Internet Institute; Polly Curtis, the Chief Executive of think tank Demos; Prof John Tasioulas, Director of the Institute for Ethics in AI and Professor of Ethics and Legal Philosophy at the University of Oxford; and Prof Hélène Landemore, Professor of Political Science at Yale University.

After briefly introducing and defining LLMs and generative AI, this report provides a summary of the main themes that emerged during the symposium and outlines a list of open questions to be addressed in future research and discussions.

An oil painting by Henry Matisse of an ancient greek temple with rolled newspapers as columns. [detailed, oil painting, colourful, on canvas]

BACKGROUND: LARGE LANGUAGE MODELS AND GENERATIVE AI

So-called Large Language Models (LLMs) are currently seen as one of the most advanced forms of artificial intelligence. Many LLMs can handle multiple language-related tasks (e.g., text-generation, translation), and often also show 'emergent abilities' that they were not explicitly programmed for. Some of these models are also multimodal and can process and generate videos, images, and audio, in addition to text.

A particularly comprehensive definition and explanation can be found on Wikipedia where a large language model is defined as:

> a computerized language model, embodied by an artificial neural network using an enormous amount of 'parameters' (i.e. 'neurons' in its layers with up to tens of millions to billions 'weights' between them), that are (pre-)trained on many GPUs in relatively short time due to massive parallel processing of vast amounts of unlabeled texts containing up to trillions of tokens (i.e. parts of words) provided by corpora such as Wikipedia Corpus and Common Crawl, using self-supervised learning or semi-supervised learning, resulting in a tokenized vocabulary with a probability distribution.[4]

There is disagreement among experts about whether Large Language Models 'understand'[5] and, therefore, whether they have an internal model of the world and 'an actual conception of what they are talking about.'[6] However, the majority consensus currently seems to be that this is not the case and that these models merely 'create things which look like things in their training sets; [but that] they have no sense of a world beyond the texts and images on which they are trained."[7]

Large Language Models are often used synonymously with the term 'generative AI.' They are, however, not the same. 'Generative AI' is an umbrella term used for AI systems that can generate new forms of data, often by applying machine learning to large quantities of training data. This output can be multimodal and include text, visuals, and audio. Large Language Models are the most prominent form of generative AIs. The output that can be produced with these is, depending on the instructions, sufficiently sophisticated that humans can perceive it as indistinguishable from human-generated content.

---

[4] https://en.wikipedia.org/wiki/Large_language_model

[5] Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models.* http://arxiv.org/abs/2102.02503

[6] Newport, C. (2023, April 13). What Kind of Mind Does ChatGPT Have? *The New Yorker.* https://www.newyorker.com/science/annals-of-artificial-intelligence/what-kind-of-mind-does-chatgpt-have

[7] Siegele, L. (2023, April 20). How AI could change computing, culture and the course of history. *The Economist.* https://www.economist.com/essay/2023/04/20/how-ai-could-change-computing-culture-and-the-course-of-history

SYMPOSIUM SUMMARY

The Technology, Context, and Socioeconomics of LLMs

Speakers: Hannah Kirk (Oxford Internet Institute) and Hal Hodson (The Economist)

The first panel provided a brief overview of LLMs as technologies, how they work, and how they have evolved over time, particularly the role of feedback learning in programming and shaping the outputs of these models. The panel also addressed some of the potentials and technical limitations and risks of LLMs, in addition to common misunderstandings about them. The speakers also discussed the social economics of technologies such as ChatGPT, which have relied on what is publicly available on the web, and concerns around copyright that are likely to shape the regulation and implementation of these models in years to come.

In opening the first panel, Kirk discussed the fundamental role feedback learning plays in optimising models at major industry labs, such as OpenAI, the organisation behind ChatGPT. Feedback learning or reinforcement learning from human feedback (RLHF) involves human reviewers in programming the model, who iteratively provide feedback to fine-tune and improve the quality of the responses. However, as Kirk explained, this technique raises concerns because the data are often collected from the voices of a very small number of people and under the specifications of Silicon Valley technocrats. As a result, very few humans concentrate the power to shape the models that are used around the globe.

Especially troubling is their homogeneity across demographic lines, resulting in what Kirk referred to as 'the tyranny of the crowd worker,' or the lack of diversity and undemocratic process whereby models are adapted to humans. As she noted, this is consequential because of the false assumption of universality of these models: the problematic idea that training a model on the preferences and feedback of some 50 workers is somehow going to be generalisable to the diversity of humans that will be using the technology. Given that such few voices are at the table, it is unsurprising that these models have been found to display bias and cultural hegemony.

That said, Kirk also underscored that these biases are not a product of the architecture itself but of the data on which it is trained – e.g., the digitally written human text. A common misconception around LLMs, she pointed out, is attributing more to the model than just being a statistical probability distribution. She noted how biases in LLMs mirror those that already exist in society itself, even as these models may amplify them. As such, when curating and pre-training the training data, it is crucial to reflect on what worldview we want our models to have (i.e., whose voices are included and prioritised) and how models can be fine-tuned through approaches such as human preference learning. Kirk advocated for tweaking models

in pursuit of making them harmless, helpful, and honest.[8] She emphasised the importance of incorporating more diverse voices and worldviews into the models and making their training more democratic so that they benefit the many and not the few.

Meanwhile, Hodson emphasised the importance of understanding the historical trajectory of these technologies and suggested that there is no reason to think this new version of AI is an exception to the hype curves we have historically seen with the emergence of new technologies. Despite the sudden buzz around ChatGPT, he argued that LLMs are but the latest iteration of an incremental process that has been underway for quite some time. While in a sense this technology has exploded onto the scene, he noted, what is new is its public availability. What we have seen is not a massive spike but a more gradual upward trend in capabilities: what was machine learning became deep learning, which became AI, which became foundation models, etc. This rebranding, Hodson suggested, has not been an entirely innocent process, either: Producing catchy new names helps build excitement around technologies and draw new money from investors. Yet it is clear that a large portion of these efforts will not only be expensive but will ultimately fail.

Despite this, Hodson made clear he is not an AI 'doomer' and is also not particularly fearful of what the LLMs can do as statistical representations of language. Unless you get very philosophical about it, he argued, at present, developers can still essentially see and understand what is happening with these models, which at the end of the day are computer systems. As such, Hodson maintains that no one who thinks about this issue really believes these are anything but machines, even as they may worry the machines may make mistakes. Thus, while he acknowledged that it is not unreasonable to worry about long-term risks, he also maintained that this is still a very narrow form of intelligence. All language, even statistical representations of language, is just a tiny slice of what humans are capable of. Along these lines, one common misunderstanding Hodson sees around LLMs is that people often do not think enough about or understand the scale of large computer systems and the massive amounts of data required for them to work. That is, it takes an enormous amount of energy to enable an even narrow bit of intelligence, and in his view, this backdrop should contextualise worries about things going seriously awry.

Hodson also showed himself particularly interested in the social economics of these technologies, especially who gets the rewards for producing them, given that most LLMs are built on large databases created by scraping the entire public web. This data is used to pre-train models before the RLHF happens. Efforts to regulate the creation of these databases could have a profound impact on what these technologies are allowed to do. This, Hodson noted, has played out in different ways over history as disruptive technologies have, time and again, troubled previous copyright regimes, for example, with player pianos in the 1800s that relied on existing sheet music.[9] More

[8] Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... & Kaplan, J. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*. https://arxiv.org/abs/2112.00861

[9] Greenberg, B. (2017, May 22). ***Copyright Law and New Technologies: A Long and Complex Relationship.*** Library of Congress Blogs. Retrieved from https://blogs.loc.gov/copyright/2017/05/copyright-law-and-new-technologies-a-long-and-complex-relationship/

recently, he recalled, the Author's Guild in the United States sued Google for digitising copyrighted books for an online database. What has often happened in such cases is that a concord is reached between copyright holders and creators, and Hodson anticipates an agreement like that will have to be reached with LLMs as well.

Regarding AI specifically, he pointed out that the third bullet point of the forthcoming European Union AI Act dictates that models must publish all copyrightable data that went into their training[10] – a demand whose compliance could require more labour than went into making many of these models in the first place. He explained that alongside new companies experimenting with new capabilities, and in parallel with businesses trying to be fast and get an edge, there are multitudes of lawyers making claims about why these companies need to be careful. In Hodson's view, such trends indicate that considerable challenges may precede the consolidation of many of these technologies in the long run.

Personalisation, plurality, and democracy. There are also a variety of avenues that could be explored to give users greater control over their interactions with LLMs. The potential for personalisation is one possibility Kirk finds both scary and exciting. For instance, personalisation could be used to adapt language or style in presenting information to people of different ages or backgrounds. However, she was quick to acknowledge that personalisation can also come with trade-offs, and the risks-benefits need to be carefully evaluated and thought through. One significant concern Kirk affirmed is that LLMs can be used to train more radical language models, such as so-called 'anti-woke' language models, demanded by some figures on the political right.[11] Personalisation also raises questions about the implications of these technologies for democracy to the degree that they may inhibit people from encountering views different from their own, extending longstanding conversations about whether new media technologies are conducive to echo chambers and filter bubbles,[12] or that they may be ill equipped to handle thorny issues around bias.

Kirk described one promising technique for consensus building in LLMs, which rewards the model for reflecting a representative sample of a given country in its responses, akin to a democratic electoral process; that is, it reflects a distribution of what the population would prefer. However, Kirk was most enthusiastic about narrower approaches to personalisation – what she called 'personalisation within bounds' – which focus less on elements that are broad and value-based or contentious and more on the style, tone, or other attributes of the text. That said, she acknowledged it was almost inevitable for there to be some ideological splintering and underscored the need for policy to stay ahead of the curve in this regard. In her view, any efforts to address bias should focus on the largest unfair advantage, in other words,

---

[10] Mukherjee, S., Chee, F. Y., & Coulter, M. (2023, April 28). *EU proposes new copyright rules for generative AI.* Reuters. Retrieved from https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/

[11] Knight, W. (2023, April 27). *Meet ChatGPT's Right-Wing Alter Ego.* Wired. Retrieved from https://www.wired.com/story/fast-forward-meet-chatgpts-right-wing-alter-ego/

[12] But see: Ross Arguedas, A. , Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2022, January 19). *Echo chambers, filter bubbles, and polarisation: a literature review.* Reuters Institute for the Study of Journalism. Retrieved from https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review

pulling extreme points closer to the middle distribution. However, definitions of fairness and bias can vary across communities, making these approaches fraught with difficulty.

Meanwhile, Hodson pointed out that companies using RLHF are often optimising against controversy and scandal. He also cautioned against alarmist narratives rooted in assumptions that all human conversations would suddenly be held within the confines of a ChatGPT box. For him, the largest concern in this sense pertains to the consent process for LLM supply chains. Thinking through these matters more carefully could help pave the road towards models built by and for communities.[13] Hodson suggested that the fundamental question for nation states, in this regard, is what data they should be collecting and what datasets they should be generating. As he pointed out, creating a plurality of datasets and models may help address concerns about ideological isolation. However, he also underscored the need to protect 'democracy' from becoming a tech company buzzword, as for-profit corporations are not in the business of democracy. More specifically, Hodson noted that while making money is a legitimate objective, it is not the same as and should not be conflated with democratisation, and how we use language when discussing these matters should be precise enough to reflect the difference.

Along these lines, grappling with generative AI also requires a deep reflection on how much power technology companies have in controlling models and outcomes, especially across diverse national contexts.[14] If you speak to a model in a different language, can it also notice shifts in the cultural context, such as topics that are controversial in some countries but not in others? Kirk suggested there was evidence suggesting models did shift worldviews when speaking, for example, in Danish versus American English about topics such as gun control or abortion. One trend she believes we might see in this regard is a further push on part of nation states wanting their own sovereign language models.[15] However, much still needs to be drawn out in terms of boundaries and privacy, among other issues.

How Generative AI is Impacting the News Media

Speakers: Laura Ellis (BBC), Gary Rogers (Fathm), and Dr Gemma Newlands (Oxford Internet Institute)

Participants in the second panel discussed the impact of generative AI on newsrooms and journalism more broadly, including examples of how these technologies are already being implemented and promising avenues for using them in the future. The speakers also addressed important concerns around how these technologies may complicate efforts to counter misinformation and disinformation, and how a growing reliance on LLMs might create new vulnerabilities for newsrooms. Moving forward, it will be crucial for organisations to carefully evaluate risks and set in place necessary rules and safeguards.

[13] Brown, I. (2023, June 29). *Expert explainer: Allocating accountability in AI supply chains.* Ada Lovelace Institute. Retrieved from https://www.adalovelaceinstitute.org/resource/ai-supply-chains/

[14] See e.g.: Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science, 379*(6635), 884–886. https://doi.org/10.1126/science.ade2420 & Simon, F. M. (2022). Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy. *Digital Journalism, 10*(10), 1823–1854. https://doi.org/10.1080/21670811.2022.2063150

[15] Titcomb, J., & Field, M. (2023, February 22). *ChatGB: Tony Blair backs push for taxpayer-funded 'sovereign AI' to rival ChatGPT.* The Telegraph. Retrieved from https://www.telegraph.co.uk/business/2023/02/22/chatgb-tony-blair-backs-push-taxpayer-funded-sovereign-ai-rival/

Ellis began the second session describing the BBC's first incursions into generative AI, dating back to 2018 and 2019, when they started training people to look at ChatGPT2 and anticipate what impact this technology might have. As has been the case with many newsrooms, the process at the BBC has also, at times, created anxiety about how (generative) AI might impact jobs in an industry already deeply affected by the uptake of new technologies and which continues to grapple with mass layoffs. She noted that there have also been ongoing conversations around issues like safety, copyright, accuracy, and privacy, in addition to the development of foundational work on AI ethics.[16]

Ellis acknowledged that the BBC is already implementing generative AI in some spaces, using a model trained internally on BBC content to summarise information. While some news organisations have already begun releasing AI generated content directly to audiences, most have chosen not to do so for now, including the BBC, where Ellis explained the decision has been to always keep a human in the loop. As such, any outputs produced with AI must be assessed by a human, who is fully responsible for what is published. She also underscored her conviction in the importance of the 'human soul' for news and rejected any suggestion of journalism being reduced to a social scraping exercise, which she believes has a role but should not be the primary role. That said, she argued that AI can also create new capacities and stimulate creativity.

While acknowledging the need to guard against some of the real risks posed by AI, Rogers invited thinking about AI from a more positive framework, including how AI tools can help journalism fulfil its role of informing the electorate and serve democracy.[17] He proposed that a useful starting point could be acknowledging that AI in news is not entirely novel, as it has been used in recommendation engines and news distribution for some time. There is now a growing focus on whether AI can be used in content creation too, but as he pointed out, this is not entirely new either.

Rogers described his own experience founding a local news agency based on natural language generation in 2017 – referred to as AI then as well – intended to help people make sense of the news in their communities, rather than a random series of events, which most people do not have time or skills to interrogate on their own. The agency, called RADAR AI, now generates around 165,000 localised news stories a year with a staff of only five people. Rogers suggested AI could be enlisted to help address some of the current challenges confronting journalism, such as growing news avoidance, which audiences often attribute to news being too depressing or boring.[18] As such, he proposed that journalists may want to think about how generative AI could be used to offer news in a variety of styles, languages, and formats, perhaps even helping make news more palatable to more people. Some may even be interested in different styles or moods at different times of the day.

---

[16] For example, in 2021 the BBC published its internal framework that outlines key principles for the responsible use of Machine Learning Engines: https://www.bbc.co.uk/rd/publications/responsible-ai-at-the-bbc-our-machine-learning-engine-principles

[17] See e.g.: Lin, B., & Lewis, S. C. (2022). The One Thing Journalistic AI Just Might Do for Democracy. *Digital Journalism, 10*(10), 1627–1649. https://doi.org/10.1080/21670811.2022.2084131

[18] Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., & Nielsen, R. K. (2023). *Reuters Institute Digital News Report 2023* (Digital News Report). Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

Meanwhile, Newlands invited thinking more about AI in relation to newsrooms as a place of work and, more specifically, how organisational policies may shape outcomes. While conversations about AI often focus on the use of these technologies as a matter of personal choice, Newlands argued that in few cases do employees get to choose which technologies to use on a daily basis, as organisational rules and protocols frequently determine these things. Simply assuming that soon everyone will be using tools such as ChatGPT misses the larger question of whether organisations will allow them to. She suggested that we may also see tensions arise between individual and organisational interests. For example, organisations could require the use of automation to speed up work while negatively impacting employees. From her vantage point, the interesting questions are around who has the agency to determine what is used, who can refuse, and in benefit of whom.

However, Rogers underscored that technological change in the context of newsrooms is often led not by management but by staff, as they identify and experiment with tools that make their jobs easier. In this regard, he suggested that the uptake of AI tools is already taking place in many newsrooms even as few have policies in place. This, he pointed out, puts additional pressure on newsrooms to get their ducks in a row and implement policies that give people the space to work safely within the bounds of organisations' ethical and editorial policies.

On misinformation, truth, and trust.

One of the anxieties about generative AI is that it might further complicate efforts to counter misinformation and disinformation, an issue that has drawn growing attention with the rise of digital platforms such as social media, where gatekeeping is much more limited than it was in the past, in addition to increasingly sophisticated tools to alter, and now, create images and video, such as deepfakes. This is something the BBC worries about, according to Ellis, who recounted past examples of false information being delivered with their branding on it. However, it is not only audiences who can be fooled by digitally manipulated content – journalists, too, can fall prey. Some recent examples, such as the widely circulated image of the Pope in a puffer coat, fooled even experienced internet users, Ellis included.[19] While she appreciates ongoing efforts to improve deepfake detection, success rates remain relatively low (about 65% maximum according to Ellis).

So, what can be done about is? One strategy the BBC is currently exploring is how to place signals into content that that can be detected and used by platforms when making decisions about what to elevate.[20] Ellis suggested that news media have an important role to play in this context of heightened uncertainty. She also proposed that media literacy will be crucial to helping people understand how AI can change the game when it comes to assessing the credibility of what they encounter online. However, news organisations especially need to remain more vigilant than ever in keeping a gold standard of verification. Rogers agreed on the important role to be played by media literacy, although he pointed out that people already have a toolkit of coping

[19] See, for example, Golby, J. (2023, March 27). *I thought I was immune to being fooled online. Then I saw the pope in a coat.* The Guardian. Retrieved from https://www.theguardian.com/commentisfree/2023/mar/27/pope-coat-ai-image-baby-boomers

[20] https://contentauthenticity.org/

strategies, including tech solutions, but also a reasonably developed skill base at detecting deepfakes.[21] This is an area that will almost certainly continue to evolve in light of ongoing technological changes.

For Newlands, organisations concerned about the degradation of truth need to think more about how to avoid the 'infrastructuralisation' of LLMs; that is, preventing LLMs from becoming part of the core infrastructures where truth is generated. She noted that there are real risks involved in allowing LLMs to replace human layers in information technologies that cannot be 'unreplaced' later. As such, becoming overly dependent on these technologies can be risky for any business or organisation, news media included.

One aspect Ellis expressed interest in knowing more about is how members of the public feel about the prospect of increasing the role of AI in news production, especially in the context of decreasing trust in news. The Alan Turing Institute published one survey on the topic, but this preceded generative AI's burst onto the scene.[22] For news organisations invested in cultivating trust with audiences, she suggested, it will be important to keep an open dialogue with audiences to gauge their level of comfort with the incorporation of these tools. Ellis and Newlands agreed that transparency around how AI technologies are being implemented in newsrooms – and what for – will be an important piece of the puzzle.

For Rogers, there may be other avenues, unconventional ones even, worth exploring with regards to AI and trust. For example, if an individual journalist or anchor is known to be well regarded and trusted by an important audience segment, they have a huge value in passing information through to people in a way that they will trust. He proposed that one possible functional use of AI in this sense would be having all information read to those audiences in the voice they like to hear their news from. However, Ellis maintained a more conservative position about such possibilities, saying she would personally feel uncomfortable using AI simulations of well-regarded correspondents or reporters, as there are certain foundations that need to be protected even if AI makes these practices feasible.

Accessibility and disparities. Thinking about AI in journalism also means contemplating issues around access. Many news outlets, already financially strained, lack the resources of organisations like the BBC to build their own tools, finding themselves limited to generic, ready-made options.[23] This also raises questions about how to use AI responsibly. As Rogers acknowledged, there are very real risks for news organisations of becoming reliant on technologies that they do not fully understand (e.g., ChatGPT) but also a financial risk that the adoption of these technologies may further complicate their business models. In his view, part of what makes generative AI so disruptive is that the tools of content creation are moving toward the audience, which could fundamentally change

[21] See, for example, Mercier, H. (2020). *Not Born Yesterday: The Science of Who We Trust and What We Believe.* Princeton University Press.

[22] https://www.adalovelaceinstitute.org/report/public-attitudes-ai/ Ada Lovelace Institute & The Alan Turing Institute. (2023). *How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain.* Retrieved from https://adalovelaceinstitute.org/report/public-attitudes-ai

[23] Simon, F. M. (2022). Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy. *Digital Journalism,* 10(10), 1823–1854. https://doi.org/10.1080/21670811.2022.2063150

the role of news organisations and their ability to monetise on their content. If we all have generative AI tools at home, he argued, the role of news organisations may be to create the information people can build on, shifting from news production to news gathering. Amidst all AI hype, news organisations small and large alike, need to reflect carefully on the value of going down that route.

Newlands agreed that one key issue to pay attention to are skill gaps, especially in low-tech newsrooms that may not know how to implement AI in a way that is functionally useful or for whom it does not fit into their business models.[24] She also raised concerns about the risks of newsrooms becoming overly reliant on very few companies with monopolistic tendencies for the same LLMs, which, echoing the point made in the previous panel, also raises important questions about what the infrastructuralisation of LLMs could mean for sovereignty. Newlands noted that global power structures and disparities are an important part of this conversation, as countries in the Global South are often disproportionately impacted by new technologies run by a handful of companies – or even a single company – in the Global North. At the same time, she pointed out how rarely we talk about the fact that these technologies are reliant on rare earth mining, which often takes place in countries with troubling histories of colonialism and extractivism. Yet to the extent that the development of these technologies is reliant on what is, quite literally, in the ground in these countries, this is one way she believes they may be able to draw lines and demand a voice in the conversation.

Regulating Generative AI Democratically and Globally

Speakers: Prof Hélène Landemore (Yale University), Prof John Tasioulas (Institute for Ethics in AI), Polly Curtis (Demos)

The speakers in the third panel reflected on the urgency of regulating AI, especially on a global scale, and the risks of leaving such an important task in the hands of technocrats and politicians. They discussed what democratic regulation should fundamentally achieve, what it might look like in the context of AI, and how to bring diverse voices to the table.

How might we regulate AI in democratic fashion? This question was at the core of the third and final panel, in which the speakers envisioned both best- and worst-case scenarios of how AI regulation could play out. Landemore first described what she called a technocratic path, in which technologists converse with hapless officials, many of whom will misunderstand the technology, and tell them what to do, whether at the level of nation-states or a global institution. An alternative and more appealing path to deal with the threats and promises of AI, she argued, would be a truly democratic, representative, and inclusive process. While it may sound utopian, it aligns with the perspective of deliberative democrats like herself, and she believes that it is not only feasible but something that should be pushed for. She also noted that there are already hundreds of examples of randomly selected bodies of citizens – so-called citizen assemblies – that have been convened to address a variety of topics (e.g., climate change) at the local,

[24] Rinehart, A., & Kung, E. (2022). *Artificial Intelligence in Local News: A survey of US newsrooms' AI readiness* (p. 56). The Associated Press.

regional, and national levels.[25] These past experiences, she suggested, provide the modules and elements to build a global deliberative process, drawing from rich traditions all over the world, which could be combined and structured around a global assembly to gather and think about AI regulation.

Tasioulas proposed another compelling narrative about AI regulation as a tussle between technocracy, on the one hand – the idea that such decisions should be left in hands of experts, and where the crowd has no role to play – and populism, on the other. Excluding the public from these decisions, he warned, risks creating a backlash, which often takes the form of an authoritarian claiming to stand in for the public. For this reason, he urged, democrats need to pursue a compromise that allows moving beyond these two poles. From his perspective, technocratic input is necessary for regulatory decisions, but its role should be one of advising rather than usurping democratic publics. Doing so may help diffuse concerns among those who are deeply impacted by issues they often have no say in. Tasioulas suggested that universities are important sites for these bottom-up movements to take place, but that they have largely failed to play a part because they have often fallen in with technocrats. He also underscored the importance of focusing on the common good and whether these systems are generated for truly valuable ends, beyond economic growth.

Looking to recent history for some insight, Curtis described feeling a strong déjà vu to the utopian aspiration of the early internet, when there was hope for true democratic change. She argued that we are at a similar moment in history but are now carrying the baggage of 20 years of poorly handled disruption and change. She suggested that if we manage AI correctly, we could envision a world in which citizens can access information in better ways, leading to improved decision-making, and helping repair the relationship between the state and citizens, from how we get our public services to how we interact democratically. Meanwhile, a more pessimistic prediction would see technologies baking in the biases and discrimination we already see, with very dark consequences. In this doomsday scenario, she proposed, democracy would no longer be dependent on the truth and would be more of a gamer battle of who can wield disinformation more successfully.

However, what Curtis sees as the most likely outcome is a gradual but fundamental rewiring of the way we work, which would likely happen so slowly that we almost do not notice it until it changes everything. This will not necessarily involve direct impacts on democracy but on the tangential things that come in hand with the shifts of industrial revolutions. Curtis described the present moment as an investment arms race powered by greed, and she underscored the urgency of learning from the lessons of the last two decades where we failed to intervene. Particularly telling, in her view, is that none of the four pieces of legislation in the UK on the matter are explicitly about democracy, which should be the fundamental driver for these changes. In order to make the technology work for the public, Curtis argued that ethics needs to be baked into the design and regulation of the systems, and the common good needs to be made explicit.

---

[25] See, for example, Heller, N. (2020, February 19). *Politics Without Politicians*. The New Yorker. Retrieved from https://www.newyorker.com/news/the-future-of-democracy/politics-without-politicians and Giraudet, L. G., Apouey, B., Arab, H., & et al. (2022). "Co-construction" in deliberative democracy: lessons from the French Citizens' Convention for Climate. *Nature Humanities and Social Sciences Communications*, 9(207), https://doi.org/10.1057/s41599-022-01212-6

But who is best equipped to ensure these past lessons inform future decisions? While Landemore appreciates that people are trying to get involved and ask questions, she believes those at the top are ultimately responsible for asking the right questions and reforming themselves. She posited that if there are segments of the public that want to participate, they should, but part of that is changing the narrative of what democracy is. In her view, ideas of democracy boiling down to elections are insufficient, especially given the number of empirical studies showing how electoral systems cater to the 'top' 10% of the population – and even worse in emerging countries. She sees this bar as being too low and believes the public needs to push it and claim more power. That said, she also acknowledged that those in power are very reluctant to share it, and one risk is that deliberative democratic processes can be co-opted and turned into 'participation washing' without truly enabling more participation.

Tsaioulas was sceptical with regards to the ability of politicians and the technology sector to guide the conversation in the right direction, acknowledging a combination of factors, such as financial interests, limited regulatory powers, and self-interest that should make us view their contributions with caution. In his view, what the public must think about is how to engage with others on AI in a meaningful way, in a manner analogous to how ordinary citizens have come to see climate change as a serious problem and exert pressure on corporations and governments. Curtis agreed that one of the biggest challenges now is that politicians feel like they are losing power and clinging onto what they have. For democracies to rise to the challenge, it is necessary to create a system that listens, rather than putting the responsibility on the individual.

Imagining a global regulation. What might AI regulation at a global scale look like? Landemore maintains that we need a global 'demos' that goes beyond the level of nation-states. One possibility she advocated for was thinking about a federal structure that also takes into consideration national and regional regulation, etc., given that some decisions are inevitably going to be culturally coloured. However, in her view, it may be necessary to think beyond a federation of nations given that nation-states are less relevant for certain questions. She also espoused thinking about a structure that is based on random selection, which would be less costly and more representative. Otherwise, she warned, there is a real risk of convening the same kind of people who always participate in this kind of decision-making (particularly, rich, white men).

Tasioulas agreed that some of the regulation will have to take place on a global scale, which is typically not democratic but regulated by treaties and agreements. The question, then, is whether we can use AI to propagate democratic values. One of the challenges, he proposed, is that we often use democracy as a placeholder for all values we hold dear to us, and there is no way to get all the countries we need on board when taking such an expansive approach. Rather, he argued that we need a more basic and minimalist understanding of democracy that does not automatically imply liberalism and instead focuses on freedom of speech.[26] Meanwhile, Curtis emphasised the need for any global bodies to be mirrored locally as well.

[26] Ober, J. (2017). *Demopolis: Democracy before Liberalism in Theory and Practice.* Cambridge University Press.

CONCLUSION

Generative AI – and Large Language Models – have emerged as a key trend in the development of AI systems. However, various decisions around their creation, implementation, and regulation play a pivotal role in determining whose worldviews are prioritised in their output, who benefits from and controls these technologies, and who governs their regulations. This symposium and summary report have tried to highlight some of these questions and issues.

Instead of summarising the main themes that emerged from the symposium – something for which we refer the reader to the executive summary at the beginning of this document – we would like to use this final section to draw attention to some of the questions in need of further discussion, scrutiny, and research that emerged during and after the conversations at the symposium.

On Large Language Models:

1. What are some common misunderstandings about LLMs? How can these be best addressed and dispelled?

---

2. What role do other layers of training (e.g., reinforcement learning from human feedback [RLHF]) play a role in the optimisation of LLMs? How do these approaches shape these models differently from the underlying training data?
3. Where does the concentration of power and lack of diversity in LLMs matter and how can both be best addressed?

---

4. How can policy address bias, fairness, plurality, and factuality in LLMs, considering the definitions and cultural understanding of these concepts vary across communities?

---

5. How have previous copyright regimes dealt with disruptive technologies, and how might this apply to LLMs and generative AI?

---

6. How can LLM supply chains be best interrogated, audited, and improved to ensure community involvement and ownership, adherence to human rights, fair working conditions, environmental sustainability, etc.?

On the use of generative AI and Large Language Models in the news:

1. What is the impact of generative AI on newsrooms and journalism, and how are these technologies being implemented?

---

2. What issues arise regarding misinformation, disinformation, and the potential vulnerabilities created by the reliance on and increasing use of LLMs in communication infrastructures?

---

3. Can generative AI be used to address some of journalism's current challenges – changing audience habits, declining business models, news avoidance – while

preserving the role of human involvement and the credibility of news?

4. What role do organisational policies and structures play in shaping the use of AI in news organisations, and what tensions may arise between individual and organisational interests?

5. What are the skill gaps and challenges faced by news organisations in implementing AI effectively? What issues arise around safety, copyright, accuracy, and privacy?

6. How do global power structures, disparities, and sovereignty issues intersect with the development and implementation of AI technologies in news organisations?

On the governance and regulation of AI systems:

1. How can AI be regulated in a democratic fashion, and what are the best - and worst - case scenarios for AI regulation?

2. What are the risks and consequences of leaving AI regulation in the hands of technocrats or politicians?

3. How can a truly democratic, representative, and inclusive process be implemented for AI regulation?

4. How can universities, grassroots movements, and citizen assemblies play a role in shaping AI regulation for the common good?

5. What are the potential outcomes of the gradual rewiring of the way we live, work, and interact due to AI, and how can ethics and the common good be integrated into AI design and regulation?

6. What might global AI regulation look like, considering federal structures, national and regional regulation, different representation regimes, and cultural considerations?

7. Can AI be used to propagate democratic values, and how can a basic understanding of democracy be applied in the context of AI regulation?

Variation: An oil painting by Henry Matisse of an ancient greek temple with rolled newspapers as columns. [detailed, oil painting, colourful, on canvas]

BIOGRAPHIES

The BII project team consisted of the following people:

Dr Linda Eggert *(Principal Investigator)*: Linda is an Early Career Fellow in Philosophy, who works primarily in moral and political philosophy. Most of her work concerns issues in normative and practical ethics, and theories of justice. At the Institue for Ethics in AI, her work focuses on the relationship between human rights, democracy, and AI. Currently, she is especially preoccupied with the ethics of delegating to AI, autonomous weapon systems, and the right to a human decision. Before returning to Oxford after her DPhil, Linda held fellowships at the McCoy Center for Ethics in Society at Stanford University, the Edmond and Lily Safra Center for Ethics at Harvard University, and the Carr Center for Human Rights Policy at Harvard's Kennedy School of Government, where she was among the inaugural cohort of Technology & Human Rights fellows. Linda holds a DPhil and MPhil in Political Theory from Oxford and a BA in Humanities, the Arts, and Social Thought from Bard College Berlin, and occasionally teaches for Apple University.

Dr Amy Ross Arguedas *(Rapporteur and lead author)*: Amy is a Postdoctoral Research Fellow at the Reuters Institute for the Study of Journalism (RISJ), where she works on the Trust in News Project. Her research interests include journalism, media, and health, especially as they relate to new technologies. She holds a PhD from the Media, Technology, and Society program at Northwestern University and a BSc in Communication Studies with a concentration in Journalism from the Universidad de Costa Rica. Before pursuing her doctorate, she worked as a journalist for five years at the Costa Rican newspaper, La Nación.

Felix M. Simon *(Principal Investigator and co-author)*: Felix is a communication researcher and doctoral student at the Oxford Internet Institute (OII), a Knight News Innovation Fellow at Columbia University's Tow Center for Digital Journalism, and an affiliate at the Center for Information, Technology, and Public Life (CITAP) at the University of North Carolina at Chapel Hill. He also works as a research assistant at the Reuters Institute for the Study of Journalism (RISJ) and regularly writes and comments on technology, media, and politics for various international outlets. His past and current research focuses on AI in the news, political communication in the digital age as well as the changing nature of journalism and the media in the 21st century. He also takes an active interest in the future of mis- and disinformation. Felix holds degrees from Goethe-University Frankfurt and the University of Oxford. He is currently a fellow at the Salzburg Global Seminar and an Associate Fellow of the UK Higher Education Academy. Before returning to the OII for his doctoral studies, Felix worked as a journalist, editor, and researcher in London. Past work experience also includes the BBC and Olympic Broadcasting Services (OBS) in London and Innsbruck.

ACKNOWLEDGMENTS

We would like to express our gratitude to the team at the Balliol Interdisciplinary Institute (BII), especially Elinor Richardson, and Balliol College for their generous funding, which enabled us to carry out this project. The Institute for Ethics in AI, especially Marie Watson and Lauren Czerniawska, provided invaluable administrative support, without which the symposium would not have been possible. We are also grateful to Dave Barker and his team at Balliol College, thanks to whom everything went smoothly on the day of the event.

We are indebted to Dr Amy Ross Arguedas, both for her extensive and exceptional notetaking during the symposium, and for putting this detailed report together. We are also grateful to Daniel Patiño for the terrific graphic design and the layout.

Finally, we would like to express our sincere thanks to all the panellists and interviewees, whose valuable perspectives and insights shaped the symposium and, subsequently, this summary report. Their input and participation were instrumental in informing the findings and recommendations presented herein. The usual disclaimers apply.

Dr Linda Eggert
Felix M. Simon

Oxford, July 2023