DEMOS

OPEN SOURCING THE AI REVOLUTION

FRAMING THE DEBATE ON OPEN SOURCE, ARTIFICIAL INTELLIGENCE, AND REGULATION

JAMES BALL
CARL MILLER

NOVEMBER 2023

The project was supported by

pwc

Published by Demos November 2023
© Demos. Some rights reserved.
15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

ACKNOWLEDGEMENTS

We undertook this project at speed, over the course of six weeks in order to deliver its insights ahead of the AI Safety Summit.

We are grateful for PwC's invaluable insights and support and would like to thank all who contributed to the ideas in this paper by attending our private discussion session and feeding back on the drafts of this document and a briefing note that preceded it.

Over 40 leading technologists, representatives from AI companies and investment firms, as well as colleagues from across civil society, have been involved in sharing ideas on this, representing a diverse range of points of view on this subject. We deliberately and thoughtfully engaged people from the widest spectrum of opinion to help inform this conversation. We are grateful to officials from the Department for Science, Innovation and Technology, as well as those from the AI Foundation Model Taskforce, for their engagement in this project.

Finally, our thanks to Demos colleagues: Mary Briggs, Polly Curtis, Sophia Knight, Naema Malik, Hannah Perry and further insights and support from researcher Flynn Devine.

Any errors remain the authors' responsibility.

James Ball, Demos Fellow
Carl Miller, Research Director, CASM at Demos

ABOUT THIS PROJECT

This project is part of the work that CASM at Demos undertakes to articulate, measure and advocate for an internet and technologies that protects and promotes democratic values and human rights. It sits within our programme looking at Emerging Technologies and their impacts on society, democracy and the economy.

At Demos we put people at the heart of policy-making and are interested in finding consensus in an era defined by division. In this project we were keen to put that into practice: taking a subject that is dividing the interested community and looking for pragmatic solutions that build on consensus for the common good, rather than fuelling division.

POLICY CONTEXT

As we publish this paper, the UK government is convening an AI Safety Summit at Bletchley Park, with governments and technology companies from across the globe gathering to consider the challenges of the future of AI governance. These companies and many others are in a frenetic race to develop ever-more advanced artificial systems intelligence models, staking the future of their companies on a technology the UK's outgoing Chief Scientific Advisor told parliament could be the most transformative since the industrial revolution.[1]

The UK government, like governments around the world, is being asked to thread the needle on AI regulation. They know they have to marry controls with innovation, safety with speed, and morality with disruption. They are under pressure from the public that see both the great benefits AI might confer, as well as the risks. AI talent and capital is global and mobile, and any regulatory regime must reflect national traditions within a context of frenetic international competition.

Governments are having to work just as quickly to try and respond to the pace of the development of AI, which in recent years has outpaced what even most of the technology's biggest advocates had expected. Legislators and regulators are keen to harvest the potential benefits to growth, business, and society – but are acutely aware of its potential risks. These range from the near term observed harms and risks, including accelerated fraud, misinformation and bias, to potential risks such as mass unemployment and even existential threats.

Regulation is a fraught and uncertain business at the best of times, full of unexpected outcomes and potentially perverse incentives. With AI it must now proceed at a pace that matches the dizzying rate of advance of the technology itself. The decisions that must be made are both extremely difficult and absolutely essential.

Showing that the government understands AI and has a plan to roll it out safely will be significant for gaining public confidence, amid considerable concerns about AI safety. Protest groups have used direct action to warn of the risks they see.[2] Content producers have led the first 'data revolts' against AI models ingesting their product.[3] In 2016, just 5% of the public named AI as a top-three risk to humanity, but the same survey repeated in June 2023 saw that rise to 17%.[4]

In the shorter-term, 62% of the public think that AI will have a net negative effect on jobs (versus 8% who think it will be positive) – and only 1% have "a great deal of confidence" that the companies working on AI will do so responsibly, and similarly just 1% have a great deal of confidence in the government's ability to regulate it.[5] The use of AI poses risks to companies' reputations. One American study found that 70% of people who had heard of AI don't trust companies to use AI responsibility.[6] These figures emphasise the importance of establishing a credible regulatory system for maintaining the social legitimacy of AI in democratic contexts.

[1] Vallance, P, 'HC1324/ Q47: Science, Innovation and Technology Committee,' House of Commons, 3rd May 2023. https://committees.parliament.uk/oralevidence/13104/html/

[2] Meaker, M, 'Meet The AI Protest Group Campaigning Against Human Extinction,' Wired, 25th June 2023. https://www.wired.co.uk/article/pause-ai-existential-risk

[3] https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html

[4] Smith, M, 'Concerns for an AI apocalypse rise in last year,' YouGov, 5th June 2023. Available at: https://yougov.co.uk/technology/articles/45782-concerns-ai-apocalypse-rise-last-year [accessed 26/10/2023]

[5] Smith, M, 'Britons think artificial intelligence will cost jobs…but not their own,' YouGov, 19th May 2023. Available at: https://yougov.co.uk/technology/articles/45730-britons-think-ai-will-cost-jobs-not-their-own [accessed 26/10/2023]

[6] Pew Research Centre, https://www.pewresearch.org/internet/2023/10/18/views-of-data-privacy-risks-personal-data-and-digital-privacy-laws/pi_2023-10-18_data-privacy_1-11/

THE OPEN SOURCE CHALLENGE

In this context, governments are being forced to confront a new regulatory challenge: whether the potential risks of AI at the frontier mean that for the first time, they need to seriously consider restricting the open source tradition that has underpinned the innovation of the internet since its inception.

This is fast becoming one of the most crucial questions related to AI regulation: How do you marry safety and control with the long established open source orthodoxy? This is where technical information regarding the models is made available to enable more people to develop, host, retrain and repurpose it. From StableLM to Dolly, Cerebras-GPT to Llama 2, across 2023 a number of powerful AI models have been released in open sourced ways (albeit to different degrees); now commentators speculate whether open sourced models will win the 'AI race' against their proprietary counterparts.

The open source movement has long voiced a series of powerful arguments for open software development. Making the code available, they argue, makes software transparent and therefore safer and also speeds up innovation. Much of the software that undergirds the internet is open source, and it is a way of working as old as computing itself. Many now extend the moral and technical arguments for open source to artificial intelligence too.

Opening up models to outside developers and even to the public at large could clearly help build confidence in their positive potential. Having the assurance of knowing that the code underpinning models is open for analysis and modification may help its wider acceptance, as well as speeding up its uptake.

But open source AI is not without risks: people can use open source AI to create models without guardrails, as we've seen with applications like WormGPT,[7] a generative model built to assist cyber-criminals in their endeavours that defines itself as an 'enemy' to ChatGPT. However, open source advocates point out that some closed AI systems are easily circumvented and also vulnerable to cyber-criminals.[8]

Existing open source AI can be used to create models specialised in causing harm or spreading misinformation, or in creating images of child sexual abuse.[9] As models become vastly more capable, the potential for accidental or deliberate harm grows commensurately and, some believe, potentially to an existential risk on a par with biological or nuclear weapons.[10] Open sourcing highly capable AI models can exacerbate that risk.

Crucially, once models are open source, enforcement of any regulation that was later deemed necessary would be more onerous for government and possibly the regulated entities alike. Few organisations are capable of creating foundational models, but people with relevant skills could download, build on and deploy one that's already made. At this stage, new regulations would have to be enforced at the user level – a much costlier and more complex proposition for all concerned.

[7] Nachiappan, A, 'WormGPT: AI tool designed to help cybercriminals will let hackers develop attacks on large scale, expert warn,' Sky News, 18th Sepetember 2023. Available at: https://news.sky.com/story/wormgpt-ai-tool-designed-to-help-cybercriminals-will-let-hackers-develop-attacks-on-large-scale-experts-warn-12964220 [accessed 26/10/2023]

[8] https://www.europol.europa.eu/media-press/newsroom/news/criminal-use-of-chatgpt-cautionary-tale-about-large-language-models

[9] Safeguarding And Child Protection Association: Available at: https://www.sacpa.org.uk/2023/06/05/why-computer-generated-child-abuse-is-the-next-crime-wave-waiting-to-happen/ [accessed 26/10/2023]

[10] Hazell, W, 'Sunak to host global AI summit and warnings of threats to human civilisation,' The Telegraph, 3rd June 2023.Available at: https://www.telegraph.co.uk/politics/2023/06/03/rishi-sunak-host-global-ai-summit-joe-biden/ [accessed 26/10/2023]

THE CURRENT DEBATE ON OPEN SOURCING AI

In the public sphere, the debate around AI regulation – and particularly whether open access to advanced models needs to be restricted – is an extremely polarised one, provoking passionate disagreement between those working on AI themselves. Marc Andreessen, the founder of Netscape (which in turn founded Mozilla, which produces open source web apps such as the Firefox browser) and general partner of the venture capitalist firm A16Z, has positioned himself as the champion of unfettered development.

His movement, "Effective Accelerationists", or "e/accs" as its adherents are identifying themselves on social media, think that any obstacle to AI development is a negative. In his recent "Techno-Optimist Manifesto", Andreessen listed ideas such as "existential risk", "tech ethics", and even "trust and safety" as "enemies".[11]

In contrast to that, Connor Leahy – CEO of the AI safety company Conjecture – has publicly called for an immediate moratorium on the riskiest forms of AI development, a movement dubbed (mostly by its critics) "Decelerationism".[12]

The stakes are high and the range of public positions are extremely wide and passionately held. The version of a debate that appears in the public sphere often looks even more polarised than it might be in reality. TV shows pick the most interesting and often extreme positions. People make the most newsworthy version of their argument when in front of the camera or writing op-eds, and newspaper editors look for conflict. Social media is fine-tuned for division, rather than nuance and consensus building.

Against that backdrop, Demos worked to convene a private discussion forum – including tech executives, venture capitalists, civil society, and figures from government – not to try to settle the issue, but to try to understand the true nature of the division, where any consensus lies and to better frame the terms of debate.

But the question sparks such strong opinions from all sides that we need to create a more constructive and pragmatic space to explore the issues, based on the realities of the emerging technologies of today and what we are learning about their risks and opportunities.

Given the pace of AI development we now see, coupled with the expectations for AI to deliver tangible value and the need to boost innovation, we believe that the debate around the openness of next-generation artificial intelligence models is an urgent one – and one which we owe it to the public to hold constructively.

[11] Andreeson, M, 'The Techno-Optimist Manifesto,' A16Z, 16th October 2023. Available at: https://a16z.com/the-techno-optimist-manifesto/ [accessed 26/10/2023]

[12] Stacey, K, and Milmo, D, 'Sunak's global AI safety summit risks achieving very little, warns tech boss,' The Guardian, 20th October 2023. Available at: https://www.theguardian.com/technology/2023/oct/20/rishi-sunak-global-ai-safety-summit-connor-leahy [accessed 26/10/2023]

HOW WE APPROACHED THE PROBLEM

At the outset of this project, we hypothesised that if key stakeholders in this debate had a structured opportunity to continue the specific discussion on open source and AI in a constructive way, there might be at least some degree more agreement than was first apparent. This informed the core of our approach.

We wanted to better understand the debate around open source and AI in a number of ways. First, what is the real nature of the disagreement and where might points of consensus be found? Second, what are the drivers of disagreement? How are different parties to this debate citing different bodies of evidence or pointing to different precedence, legal or regulatory principles, or moral priority when justifying their position? And third, what was the nature of these drivers? Are the differences empirical or ideological?

Our approach to answer these questions was through a structured and observed debate. Demos identified key stakeholders across the issue and convened a private discussion forum in London. We offered a provocation: a framework that asked discussants to nominate a level of regulatory control that they saw to be necessary for a given series of increasing AI capabilities (the framework used is included as an annex to this report). Discussants were placed in groups to provide structured feedback on the proposed framework before joining a moderated discussion to allow more ideas to be introduced.

The participants to this discussion were carefully selected to ensure that the full breadth of the debate was represented in the room. It included CEOs and public policy leads from leading technology companies, AI investors, civil society specialists and government officials and senior advisors. It operated under Chatham House rules to encourage a maximally candid and open conversation – for that reason this report contains the summaries and individual quotes, but these are unattributed to protect anonymity. At the end of the event we felt that we needed to seek out additional voices from the open source community, which we did in subsequent sessions where we gathered further inputs.

We asked participants to break into groups and discuss the framework, seeking areas of both alignment and divergence on where control should be matched to potential capability of LLMs. We gathered feedback on the framework, which we don’t include in full here, but do plan to iterate in future work. What follows is a description of the themes that emerged in the resulting debate.

THEME 1. Open source and security

Open source advocates used the traditions and examples of open source software development to make the case for open sourcing generative AI. A key claim was that software in general becomes safer and more accountable as it becomes more transparent, and that we will see the same with open source AI models. “The more open you are,” said one discussant, “the more you can tap into the academic and scientific community which means it’s more safe”.

This claim was contested, however. One discussant felt strongly that given that companies currently producing foundational models could not explain their capabilities or how they arrived at particular outcomes, further development, whether open- or closed-source, should be paused at this stage to safeguard against potential existential risks. Open-source AI models, another discussant responded, are a "big blob of numbers" that are "grown, not designed", resembling biology more than technology to some degree. They are not interpretable in the same way as an open source programme, and so you will not see the same security benefits when making it visible.

The second aspect of open source security was around the offensive and defensive balance. A vulnerability just has to be found once, and all versions of a piece of software can be patched; it enjoys a defensive advantage. It is a widely accepted and evidenced view for most software that open source and disclosure of vulnerabilities helps people defending against exploits more than those using them – but this consensus does not hold when applied to AI vulnerability or exploit disclosure.[13] "Many things don't operate in this way," said one discussant. "It's much easier to be on the offence."

THEME 2. Competition and concentrations of power

There was a risk, discussants suggested, of market concentration if open source AI is prohibited or restricted. It would mean a small number of very well capitalised companies taking their models behind closed doors, with no reasonable prospect of smaller innovators mounting real competition. This would be bad for innovation, for redistribution and for open markets. Other regulatory interventions (like GDPR) had, one participant noted, locked in incumbent advantage and there was a real risk of that happening again. Some participants warned that a regulatory regime agreed solely between government and large businesses would constitute "regulatory capture", acting against the interests of start-ups, not-for-profits, and ultimately consumers – and even potentially against public safety.

Advocates of more control noted however that it was large companies actually doing the open sourcing of the models. It wasn't right, they said, to see open source development as an organic or volunteer-driven one: the only people capable of developing transformational foundation AI capabilities were other very well capitalised companies capable of funding enormously intensive training runs.

THEME 3. How to understand risk

Risk was completely key to the way that everyone understood how emerging AI capabilities should be navigated and understood. There were, however, different basic understandings about how risk should be conceived. One important idea was 'cumulative risk', the notion of aggregating lots of different types of risks together that may appear at different times, and in different stages of AI life-cycles. A contrasting idea was of 'net benefit'; that the benefits have to also be understood and included into any moral arithmetic.

However risk was understood, a second layer of disagreement was whether the burden of proof was then on developers to show their models were safe, or on regulators (or others) to show a model was dangerous. "We shouldn't be building and sharing by default and then looking for justification", one discussant said. "We need to reverse the framework." Whereas others pointed to the UK's open regulatory culture, and a presumption of openness until levels of risk or harm are demonstrated. A key point of divergence emerged: should we start from an open or closed presumption?

THEME 4. Getting away from binaries

There was more consensus around the idea of moving away from absolutist positions, false dichotomies and forced binaries. There was no easy way, most agreed, to navigate these dichotomies or to gain clear consensus regarding them. To move ahead, regulation should be fine-grained, it needed to abandon crude distinctions between open and closed: there are many regulatory interventions that don't necessarily 'close' development, including requiring documentation, particular forms of testing, availability of training data to particular researchers, and so on, all of which were raised by different discussants during the process.

Another rejection of open vs closed being a simple binary involved multiple discussants noting model releases could be staged or staggered – essentially a model could be opened up incrementally, checking each level of openness is a net good and then continuing, rather than opening all at once.

A third binary considered and largely (though not universally) rejected by attendees was the idea that immediate risk versus existential risks acted as binary – a continuum was felt to be more realistic by some discussants.

[13] Shevlane, T and Dafoe A, 'The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse,' Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), 7-8th February 2020. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/The-Offense-Defense-Balance-of-Scientific-Knowledge.pdf [accessed 26/10/23]

THEME 5. Arguments to authority. Metaphors, precedents and comparisons

Participants discussed how in moments which are new and complex, we often reach for metaphors we can understand to relate them to. This is happening with AI. Discussants compared it to open source software of course, but also to nuclear technology, to opioids, to human gene-editing, to chemical engineering, and with biological engineering.

These different analogies represented deep drivers of disagreement between discussants. They cause people to reach for different precedents, different regulatory traditions, salutary warnings and overriding ethical principles. These metaphors are extremely powerful empirically and also emotionally, and reflect a whole range of different defaults and biases that people have.

However, there is no straightforward way to prove one metaphor is better than another. Any metaphor is debatable, contestable and replaceable, and mustn't, without evidence, be used to directly answer or conclude any of the controversies or themes laid out in this paper.

THEME 6. Despite the wide range of opinions in the room there was near consensus that there is potentially a point at which the risks outweigh the benefits of open sourcing AI

There was a very wide range of views about what the point of regulation would be, how you would go about assessing the risks and what the response might be. But there was an emerging consensus that such a point does exist. This last theme is significant, and we develop it in the premises we describe below.

THE WAY FORWARD
FOUR PREMISES TO BUILD UPON

Bringing AI participants together in a structured format drew out many useful insights, beyond what we routinely saw in the public version of the open source AI debate. Through reviewing the material from the roundtable and both formal and informal post-roundtable discussion with some discussants, we sought to build further on that progress by offering four premises to help provide a basis to further the conversation on creating an effective legislative framework to support and regulate open source artificial intelligence.

These four premises are grounded in the discussions, debates and insights from the roundtable, but do not necessarily reflect the views of everyone (or even necessarily a majority) of participants in the roundtable. Instead, they are propositions that we believe each reflect the current state of debate, which if they are taken as premises for its next stage could help reach consensus, or something closer to it.

PREMISE ONE: Generative AI is a very specialised form of software, for which open source may not bring the same beneficial effects as it does to most other forms of software

One reason the debate on open source and AI models evokes strong feelings among the open source community is it echoes for them another debate that has roiled since the 1990s. In that decade, the US government sought to restrict access to strong encryption – through means such as inserting backdoors, requiring export licences, and other similar interventions.[14] This was justified by noting the high potential for misuse inherent with encryption, which was for several decades regulated as if it were weaponry.[15] This was strongly resisted by the open source community, who argued that cryptography would work better if it was in the open, where it could be tested and audited by anyone.

For secure encryption – now essential to the operation of the modern internet – the open source community was largely validated. In this arena at least, "security through obscurity" did not emerge as the best response: the ability to deploy strong encryption, test it, and audit it, proved to be an effective system for building the security system upon which the internet now relies. The 1990s was

[14] Aurora, F, 'What were the CryptoWars?' F-Secure, 22nd March 2018. Available at: https://blog.f-secure.com/what-were-the-cryptowars/ [accessed 26/10/2023]

[15] Reibe, T, Kühn, P, Imperatori, P and Reuter, C, 'U.S. Security Policy: The Dual-Use Regulation of Cryptography and its Effects on Surveillance,' European Journal for Security Research, 26th February 2022. Available at: https://link.springer.com/article/10.1007/s41125-022-00080-0 [accessed 26/10/2023]

also a decade that saw proprietary software released, again and again, with significant vulnerabilities and flaws. This led to new norms around transparency and disclosure – norms of openness – which have undoubtedly made software safer and more secure.

The parallels with the current debate on regulating AI models are apparent, but there are good reasons to argue that they may not entirely hold true. There are generally two main categories of arguments for the benefits of openness.

The first of those is that opening up models allows for hundreds or thousands of different people or organisations to try them out for different use cases, potentially finding beneficial uses that would never have occurred to the model's creators. This is possible to an extent with models that are open for access without being open source (such as ChatGPT), but the potential for adapting models through fine-tuning or even modification or removal of 'guardrails' means that it is much more significant for open source models than it is for models that are simply open access. A company can alter a model available through open access at any time. An open source model that has been downloaded and deployed independently could be a very different proposition.

Those at the more wary end of the spectrum do tend to note, though, that not all of those uses will necessarily be good for society – and once they have been found they cannot be undiscovered. Wherever one falls in this particular debate, there seems to be consensus that this argument holds up as strongly for AI models as it does for encryption software, or software in general.

The second argument is that open source code is available for audit and testing, and so bugs and vulnerabilities can be found, as can unintended interactions. Open audits can test that code is as bulletproof as it can be. Several roundtable participants noted that this simply isn't the case for modern AI models, which operate as "black boxes" even to their creators. There is much that users and researchers can do to test whether they can be used in unintended ways – often through a process known as "jailbreaking"[16] – and internal and external testing of this sort is a crucial part of the development process of current models.

However, as one participant noted, unlike with encryption or static code, these methods only find "the lower limits of what a model can or will do". There is no reliable way to know what a model will be capable of, through accident or design, after several rounds of fine-tuning and once it has handled several years of real-world usage and data input.

In other words, there is no automatic reason to think that the full range of benefits of open source software will be applicable to open sourcing AI models. This did not change the general consensus within the roundtable's participants that open sourcing of AI so far had been beneficial – everyone agreed that Meta's decision to open source Llama 2[17] had been a net positive, though one person said that they still would have wanted to know that it had been through a governance process to release it. Another noted that it might be too early to tell. Social media seemed an unambiguous good for several years, one noted, and the internal combustion engine seemed an unproblematic improvement for most of a century.

This premise is not intended to say that AI models should be less frequently open sourced than conventional software. Instead, it is a contention that AI models and traditional software seem to be quite distinct things, and so will have different considerations and potential benefits (and risks) from being open – what was right in the 1990s may not be applicable now.

PREMISE TWO: Neither closed nor open AI models are unalloyed goods nor unalloyed evils and so any regulatory position, including being entirely laissez-faire, involves trade-offs – this debate is not an exception to that norm

Effective Accelerationists, by dint of their very manifesto, reject the concept that any kind of regulation – including that which is much less restrictive than moratoria or bans – could be net beneficial in the development of artificial intelligence. That is a rejection of a trade-off that is widely accepted in decision-making across the rest of society. Few of us would disagree that if we banned all driving we would see fewer road accidents, but equally, few of us would think that is enough to recommend a full driving ban as a good policy idea.

Accepting the idea that restricting openness involves trade-offs also allows us to honestly consider the incentives of different groups participating in the debate – without negating their contributions to the debate.

This is significant given concerns cited by several participants around the risks of "regulatory capture", just as it is around the risks of over-regulation. Effects

[16] Learn Prompting, 'Jailbreaking,' Learn Prompting. Available at: https://learnprompting.org/docs/prompt_hacking/jailbreaking [accessed 26/10/2023]

[17] Meta, 'Introducing Llama,' Meta, 2023. Available at: https://ai.meta.com/llama/ [accessed 26/10/2023]

that can be positive for some actors in the AI space will be negative for others – considering trade-offs doesn't just mean considering the total potential benefit or loss on a societal basis, but should include to whom the benefits would flow, or who would be impacted by the losses.

A framework in which we accept there are trade-offs is one in which we can accept that different actors are working with different incentives, and weigh up how each of us consider those contributions with those in mind.

It also means that we can accept that there is – at least to some degree – a societal consideration of how fast technology emerges versus how able we are to mitigate potential negative effects, and how this might affect where AI companies choose to launch or invest.

Significantly, multiple discussants said they saw value in the approach of "cumulative risk" when it came to AI models, an approach which considers and combines the aggregate risks of each development – as new capabilities are released and used together (possibly interactively), this approach tries to look at that whole picture, rather than just individually assessing the risk of each separately.

This seems a promising approach, though it is made more complicated not just by a different assessment of risk, but also by the difficulties of actually measuring many of these impacts. "There's a question where the cumulative risk is so high that the government steps in. That's the precedent," said one participant. "There's an empirical question of whether we're there yet. [In my view] we're not there yet, but we also don't have robust ways of knowing if we're there yet. A lot of the conversations are finger in the air."

PREMISE THREE: There is a broad consensus that there will be a level of AI capability that would merit restrictions on its openness, though not what that level would be, nor how soon that might arise

Our drive to find analogies to help explain new technologies like artificial intelligence can be helpful – we've used plenty in this paper so far already. But they can sometimes serve as a crutch – the right answer to "is AI software is or is it nuclear weaponry" is that it is neither. Reducing the debate on how to treat it to picking between a number of imperfect similes just leads to arguments.

Instead, we can consider what artificial intelligence models can do now, what they are likely to be able to do in the next few years, and where we think they might go beyond that. We can, for example, consider whether or not we think artificial general intelligence could be made open source without restriction whether we think it could happen in the next five years, or we think it is likely never to happen.

This seems to us a helpful way forward. If we are trying to draw up a framework that is forward-looking, we should look forward – if we've broadly agreed what we should do if certain capabilities arise, does it matter if they arrive next year or in 20 years time? To the extent that we make sure regulatory measures, should they be deemed necessary, are in place in time, it matters. If we pass a law that would govern a capability that doesn't then materialise for 10 years (or ever) there's very little in the way of opportunity cost provided it doesn't say "don't do any research that might find this" but instead "get it validated and approved before open-sourcing it".

Which risks of AI policymakers should focus on is itself a divisive topic: the focus on 'existential' risks is often promoted by those driving the development of next-generation AI, and is seen by some as either a deliberate or a well-meaning but misguided distraction from concrete harms that AI is causing today and in the near future – from enabling greater mass surveillance of workers, to amplifying racial discrimination in policing.

We believe that policymakers need to be developing flexible and iterative policy frameworks that can respond to immediate and emerging harms as well as long-term risks: and that regulatory principles to minimise harms to citizens now and in the future can be co-constructed rather than competing. We further argue that this distinction between immediate and existential risk is somewhat arbitrary, as one may easily turn into another. By looking forward, we don't need to restrict ourselves to one or the other.

Participants at the roundtable – which included people who are generally strongly against regulating AI in the context of open source – almost universally agreed that there is a point in the development of artificial intelligence where some form of restriction on openness would be merited. There was however, no consensus on where that point would be, how soon it might be reached, or what form that action should take.

Several discussants on the open side of the debate noted that they would expect and encourage regulatory actions that didn't restrict openness for some potential risks. One cited the case of tackling an AI that might suggest means of creating biological or radiological weapons may be more easily tackled through existing restrictions on who can access the precursor materials to make such things, for example.

This degree of agreement is still significant, however: the overwhelming majority of actors in this space accept that considering to what degree open source in AI models should be restricted is necessarily a matter of considering trade-offs, as discussed in the previous axiom.

PREMISE FOUR: Given that it is currently impractical to curb the use of a model that has been made fully open, regulation of an AI model of a certain capability level would need to be in place before that breakthrough was made

It is currently taken as a given that the creation and training of highly capable foundational models will be restricted in the foreseeable future to organisations with significant resources and deep pockets because of the costs of development – meaning that these would be the best targets for regulation, as they would need to be legally established, have physical assets, and so on.[18]

Legitimate businesses and other organisations can still be effectively targeted by retrospective regulation if they are using an open source model – either through changes to the licences or regulations as to how they can use a model. Once the weight of a model is in the public domain, however, it is available to actors who regulators or law enforcement regimes may find it difficult if not impossible to reach.

This suggests that as a basic principle – especially when it comes to managing risks around criminal or malicious misuse of AI models – effective regulation would look to target the large entities that might choose to open source a model with particular capabilities, rather than those who might use it, though such users would likely be pursued through other means if their uses broke existing laws.

For this approach to be effective, it must be proactive. Retrospectively changing licensing rules by law could work to prevent legitimate actors using models in certain ways (though with significant, potentially business-ending, disruption) but would do little to stop malicious misuse by criminal or state actors.

As a result, for many potential harms, once a model with certain capabilities is out in the world, the menu of regulatory options has already largely sold out – it is too late. This favours consideration ahead of the development of such capabilities. If, for example, we are worried that an AI capable of not just accurately citing law but suggesting a defence strategy for a client based on political factors, the presiding judge, and so on, should not be openly available, we need to have considered that before it is released.

This does not mean banning companies from doing research either deliberately aiming to reach such capabilities, or doing research that might accidentally lead to their development. Interventions could be as light as a voluntary code under which signatories agree to certain discussions or audits before releasing a model, to legislation requiring a certain period of closed-door testing, or any number of other means.

This kind of forward-facing model can also be adapted as we learn more. Considering how AI models might negatively impact us can only be evidence-led to a certain extent – the evidence doesn't exist yet, but we can adapt our thinking as we go.

Being able to consider cumulative risk as we consider forward-looking policy is one helpful approach, but we must also remember we are not just dealing with risk, which is measurable, but also uncertainty, which is not. The further forward we look, the more a risk-based framework is actually just looking at our appetite for risk, rather than our ability to measure it.

[18] There are those who argue that open, iterative development may soon be able to keep pace or outpace big tech company foundational models, however. See: Milmo, D. Google engineer warns it could lose out to open source technology in AI race. The Guardian, 5 May 2023. Available at www.theguardian.com/technology/2023/may/05/google-engineer-open source-technology-ai-openai-chatgpt [accessed 26/10/2023]

CONCLUSION

This project has been an experiment in finding consensus and mapping the true nature of the divisions that exist about how to regulate frontier AI in the context of open source ways of working. We have created a platform for this discussion, chartered the conversation and offered some ways of understanding the debate as the UK government moves forward with the regulatory discussion.

Having presented the findings of this process, we now offer our own sense of how we think the debate can move forwards. Considering everything we have heard, we see that it is in the interests of the UK's government, tech sector and civil society to continue the conversation on the effective regulation of artificial intelligence and its openness, and to try to agree on a forward-looking, permissive, regulatory framework with as much urgency as possible.

We think that an appropriate framework of controls for future capabilities can provide the public with both protection and assurance on next-generation AI models, while giving developers certainty as to the future direction of travel.

There are considerable benefits to openness, where that can be shown to have considered risks and safety, and knowing at which point some forms of audit or restriction might be mandated gives businesses and investors a degree of certainty.

Effective Accelerationists may mandate unfettered development, but such an approach risks losing public support and could lean towards the eventual sudden introduction of much stricter measures if the release of models leads to unexpected negative consequences.

Going as fast as you can while bringing as many people as possible along may prove a more effective course of development than sprinting ahead alone. We hope that these premises can help shape this conversation towards positive outcomes – an approach that we think of as "Pragmatic Accelerationism": a belief in the positive potential of artificial intelligence models, tempered by an acceptance that risk needs to be managed and public support will be necessary.

Indeed when we convened this conversation, what we heard was closer to this Pragmatic Accelerationism than we had anticipated. Most people do think there is a point at which the risks become so high that some form of regulation of open sourced AI is necessary. The divergence is on what this point looks like, how to assess the risks and whether the starting point should be a presumption of open- or closed-sourced development. But this glimmer of consensus is a platform for pragmatists to explore further.

NEXT STEPS

This project was developed at speed, in the six weeks preceding the AI Safety Summit in London on November 1, 2023, after we identified the need for a different kind of discussion about open source in the context of AI regulation. As such we offer it as a provocation and a starting point for a wider programme of activities over the course of coming months to convene different parts of this debate and to take this experiment further. If you are interested in being involved, please email james.ball@demos.co.uk.

ANNEX 1
GLOSSARY OF TERMS

The glossary below is intended primarily as a guide to the reader to explain terms which may be unfamiliar, or in some cases to explain how we have used particular terminology within this paper – it is not intended to suggest these are definitive or settled.

In general, when we refer to "AI models", we are thinking about general-purpose AI models with capabilities beyond those available today. In the short-to-medium term, these are likely to be large language models – though this is not a given. Some, but not all, of the issues raised within this paper will apply to advanced specialised models, but those were not given specific consideration on this occasion.

Existential risk: a possible consequence of artificial intelligence that could result in large-scale loss of life, or even extinction. This could include AIs sabotaging power or transport grids, developing novel bioweapons, or similar catastrophic outcomes

Fine-tuning: Further training or tweaks to an open source model that can be used to make it more specialised, or to try to bypass its guardrails. This requires much less time and computing power than the initial training of a model. Despite the name "fine-tuning", it is believed that repeated fine-training of a model could lead to substantial advancement and divergence from the foundational model

Foundational model: A core AI model released either to be built upon in a closed way (through API or similar) or an open source way, in which the code is also released. Training a foundational model requires huge quantities of data and processing power, and so the number of entities capable of working on these are limited. GPT-4 is an example of a closed foundational model, Llama 2 is an example of an open one.

Generative AI: An AI (typically an LLM) that is primarily used to generate new 'original' content, whether that is imagery, audio, video, text, or something else. It is primarily generative AIs that have sparked the huge recent uptick of interest in artificial intelligence.

Guard rails: No-one understands fully how an LLM arrives at a particular output for a given prompt, which means engineering in restrictions to avoid offensive answers or plagiarism cannot yet be made intrinsic to the model. "Guard rails" are fine-tuning specifically aimed at safety features – moderation for offensive content, portrayals of living people, and so on.

Immediate risk: This term covers risks from existing deployment of algorithms and artificial intelligence, which can be significant in their own right: job displacement, entrenching biases, boosting misinformation and fueling fraud are all examples.

LLM/large language model: This term applies to most current cutting-edge AI models, which are trained on huge corpuses of natural language data. From this point, they essentially work like an extremely advanced autocomplete, working out which word/pixel/sound best follows another to generate new content.

Open source: Open source software loosely means software which anyone can use, modify, and re-release – the code is open to anyone and can be changed by anyone. There is no set definition for open source as it applies to AI models. It is agreed that simply being able to use a model (anyone can use ChatGPT 4, for example) does not qualify as open source, but whether a model's code being accessible is enough, or if its training data and weights must also be available, are open questions. In this paper we have defined open source as anything from the code being available and the model being downloadable to be run by outside users as open source, primarily for convenience.

Training: Training is the process of feeding huge quantities of data into a foundational model so it can be deployed and can function. These are well beyond the scope of any casual user or small business: GPT-3 was trained on 45 terabytes of data, while GPT-4's training data set was at least one petabyte (one million gigabytes).

ANNEX 2
THE CAPABILITY VS CONTROL FRAMEWORK

To help focus and structure the discussion at the roundtable, we presented a very early draft of a potential framework in which to think about the regulation or restriction of open source capabilities.

The prototype consisted of a grid with two axes. Along one axis was a series of seven stages of AI 'capabilities', ranging from models simpler than those in use today, through better-developed and widely-adopted LLMs, to something resembling artificial general intelligence. It was not intended to suggest any of these stages are inevitable, or linear, but to concentrate thinking on each possible scenario.

The other axis looked at regulator or legal mechanisms to exert control over technology, ranging from purely voluntary codes, to use of existing regulatory mechanisms, the creation of new ones, or even taking steps to delay or ban research which could lead to the emergence of these capabilities.

We have included the prototype grid below for reference, and if possible would like to continue on developing this approach from this embryonic stage. We received significant feedback on the approach through the roundtable, including whether the capability access stages were distinct enough, whether there was a more consistent way to delineate step changes, and whether the control access should focus more on the level of restriction than the regulatory mechanism used. Should we research this further, we intend to consider all of that as we examine its potential uses.

ANNEX 2 (CONTINUED)

Grid for the capability versus control framework - discussed at the private discussion forum

	A. Laissez-faire and/or policy makers encourage or invest in open sourcing	B. Policy makers encourage and/or facilitate voluntary codes of conduct	C. Policy makers encourage the use of existing laws and powers against new uses	D. Policy makers pass new 'hard' law granting new powers to existing regulators or enforcement bodies	E. Policy makers pass new 'hard' law creating new regulatory or enforcement bodies	F. Policy makers require licencing to further develop and/or use AI beyond a given level of capabilities	G. Policy makers introduce a moratorium against development of AI beyond a given level of capabilities	H. Policy makers ban the development and/or use of AI beyond a given level of capabilities
1. Standalone, purpose-specific rules-based systems and machine learning								
2. Advanced machine learning and neural networks (WE ARE HERE/GPT-4)								
3. AI-driven automation in public and private sectors								
4. AI moving towards artificial general intelligence								
5. Decentralisation / ubiquity of AI								
6. AI / human symbiosis								
7. Towards the singularity								

for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

a By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

> i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;
>
> ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

a This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

b If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS NOVEMBER 2023

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK