Expert explainer

Ian Brown

Visiting Professor, Centre for Technology & Society,
Fundaçao Getulio Vargas

Allocating accountability in AI supply chains:

a UK-centred regulatory perspective

Ada Lovelace Institute

June 2023

Executive summary

Creating an artificial intelligence (AI) system is a collaborative effort that involves many actors and sources of knowledge. Whether simple or complex, built in-house or by an external developer, AI systems often rely on complex supply chains, each involving a network of actors responsible for various aspects of the system's training and development.

As policymakers seek to develop a regulatory framework for AI technologies, it will be crucial for them to understand how these different supply chains work, and how to assign relevant, distinct responsibilities to the appropriate actor in each supply chain. Policymakers must also recognise that not all actors in supply chains will be equally resourced, and regulation will need to take account of these realities.

Depending on the supply chain, some companies (perhaps UK small businesses) supplying services directly to customers will not have the power, access or capability to address or mitigate all risks or harms that may arise.

This paper aims to help policymakers and regulators explore the challenges and nuances of different AI supply chains, and provides a conceptual framework for how they might apply different responsibilities in the regulation of AI systems. The paper seeks to address the following:

1. Set out what is or is not distinctive about AI supply chains compared with other technologies.
2. Examine high-level examples of different kinds of AI supply chains. Examples include:

    a. systems built in-house
    b. systems relying on another application programming interface (API)
    c. systems built for a customer (or fine-tuned for one).

3. Provide the components for a general conceptual framework for how regulators could apply relevant, distinctive responsibilities to different actors in an AI supply chain.

4. Explore the unique complexities that 'foundation models' raise for assigning responsibilities to different actors in the supply chain, and how different mechanisms for releasing these models may complicate allocations of responsibility.

In this explainer we use the term 'foundation models' – which are also known as 'general-purpose AI' or 'GPAI'. Definitions of GPAI and foundation models are similar and sometimes overlapping. We have chosen to use 'foundation models' as the core term to describe these technologies. We use the term 'GPAI' in quoted material, and where it's necessary for a particular explanation.

Key findings

- Our evidence review suggests that AI system supply chains have many commonalities with other types of digital technologies, for example raw material mining for smart device hardware. However, there are some significant differences in the novelty, complexity and speed of ongoing change and adaptation of AI models, which make it difficult to standardise or even precisely specify their features. The scale and wide range of potential uses of AI systems can also make it more challenging to attribute responsibility (and legal liability) for harms resulting from complex supply chains.

- After discussing various types of AI supply chains, we describe a conceptual framework for assigning responsibility that focuses on principles of transparency, incentivisation, efficacy and accountability.

- To support this framework, regulators should mandate the use of various transparency mechanisms that enable a flow of critical information. These mechanisms should also enable modes of redress up and down an AI system's supply chain and identify new ways to incentivise these practices in supply chains.

- The advent of foundation models (such as OpenAI's GPT-4) complicate the challenge of allocating responsibility. These systems enable a single model to act as a 'foundation' for a wide range of uses. We discuss how various aspects of these nascent systems

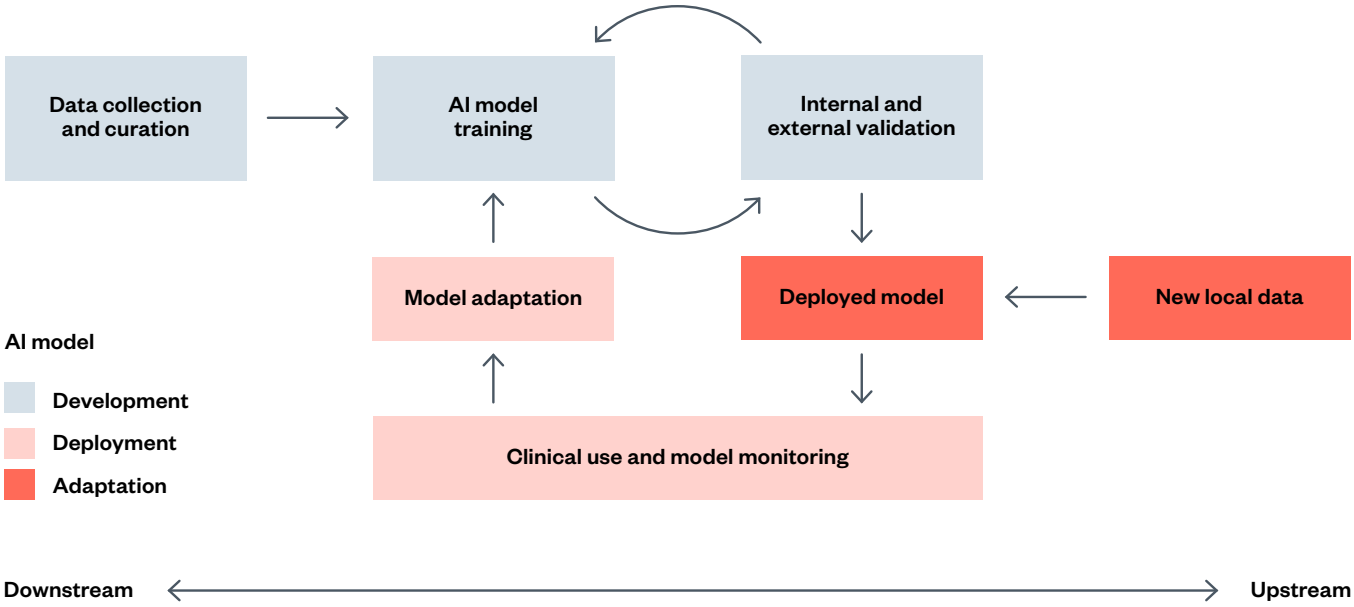
(including who is designing them, how they are released and what information is made available about them) may impact the allocation of responsibilities for addressing potential risks.

- Finally, we discuss some of the challenges that open-source technologies raise for AI supply chains. We suggest policymakers focus on how AI systems are released into public use, which can help inform the allocation of responsibilities for addressing harms throughout an identified supply chain.

Introduction

Developers and deployers of artificial intelligence (AI) systems have a variety of distinct responsibilities for addressing risks throughout the lifecycle of the system. These responsibilities range from problem definition to data collection, labelling, cleaning, model training and 'fine-tuning', through to testing and deployment (shown in Figure 1). While some developers undertake the entire process in-house, these activities are often carried out in part by different actors in a supply chain.[1] To ensure AI systems are safe and fit for purpose, those within the supply chains must be accountable for evaluating and mitigating these different risks.

Figure 1: An abstract example of an AI system's lifecycle (based on a system used in the COVID-19 pandemic)[2]



1 Engler and Renda, (2022), *Reconciling the AI Value Chain with the EU's Artificial Intelligence Act*, Centre for European Policy Studies, pp. 2–3, https://www.ceps.eu/ceps-publications/reconciling-the-ai-value-chain-with-the-eus-artificial-intelligence-act/

2 Nature Machine Intelligence, (2022), 'Translating AI models', https://www.nature.com/articles/s42256-020-0185-2/figures/1 accessed: 31 March 2023

This paper discusses how to identify who should be primarily responsible for identifying and addressing different risks in an AI supply chain

Every AI system will have a different supply chain, with variations depending on the sector, the use case, whether the system is developed in-house or procured, and how the system is made available to those who use it (for example, via an application programming interface (API), or made available via a hosted platform).

Actors within each supply chain will have differing but overlapping obligations to assess and mitigate these risks, and some will have more responsibility than others. This makes developing a single framework for accountability along supply chains for AI systems challenging.

The UK's approach to AI regulation is largely focused on companies supplying products and services incorporating AI systems. It relies on existing statutory frameworks and independent, sector-specific regulators to mitigate risks, as they are judged best-placed to understand the context and apply proportionate risk-management measures.[3]

Many companies or public sector bodies deploying AI systems will, however, need information about the practices and policies behind its development from further up the supply chain to comply with their legal responsibilities. When issues are spotted, they will also need to have mechanisms in place to communicate those problems back up the supply chain to the supplier who is best placed to fix the problems.

Based on a rapid review of academic and grey literature (including preprints, reflecting how fast the field is moving), this paper discusses how to identify who should be primarily responsible for identifying and addressing different risks in AI supply chains. It also explores the mechanisms that may allow downstream actors to reach back up through the supply chain to flag issues that they cannot deal with in isolation. We aim to cover these four areas:

1. Set out what is or is not distinctive about AI supply chains compared with other technologies.

2. Examine examples of different kinds of AI supply chains (these are theoretical, and in practice many products or organisations will deal with multiple overlapping supply chains). Examples include:

3 Department for Science, Innovation and Technology, 'A Pro-Innovation Approach to AI Regulation' https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper accessed 15 May 2023.

   a. Systems built in-house.
   b. Systems relying on another API.
   c. Systems built for a customer (or fine-tuned for one).

3. Provide the components for a general conceptual framework for how regulators could apply relevant, distinctive responsibilities to different actors in an AI supply chain.

4. Explore the complexities that foundation models raise in terms of assigning responsibilities within the supply chain, and how different mechanisms for releasing these models may complicate allocations of responsibility.

There is a limited number of companies with the resources required to produce and sustain AI systems, leading to a potentially concentrated market

What is or is not distinctive about AI supply?

Similarities between AI and other technology supply chains

Similarities between AI and other kinds of supply chains in the technology industry that are potentially of interest to regulators include:

- capital intensity and high returns to scale for production/training
- the need in some cases for access to specific, scarce inputs
- reliance on third-party software components and libraries
- questions relating to data protection and copyright law, where personal data and creative works are used for training and application of models
- a range of other fundamental or human rights issues, such as equality.

Capital intensity and high returns to scale

Like mining and large-scale industrial manufacturing, key processes in supplying state-of-the-art AI systems (particularly training large models and especially with foundation models) are likely to be capital-intensive and produce high returns to scale – that is, where outputs increase by more than the proportional change in all inputs.

This means there is a market limitation on the number of companies who can produce and sustain these systems, and this can lead to a potentially concentrated market (as we see with the cloud computing market, which is already intertwined with the training and operation of models), and hence competition law issues.

The literature corroborates this, suggesting that rapid advances in AI techniques over the last decade were primarily due to 'significantly concentrated data and compute resources that reside in the hands of

a few large tech corporations.'[4]

Other scholars point out that maintaining and developing these models will require sustained long-term rather than one-off investment:

'Due to the benefits of scaling deep learning models, continuous improvements have been made to deep learning GPAI models driven by ever-larger investment to increase model size, computational resources for training and underlying dataset size, as well as advances in research.'[5]

This has potential ramifications for the AI research ecosystem. Some analysts have noted that the increased trend towards the use of deep learning models, requiring large training datasets and computational capabilities, benefits industrial over academic research, meaning 'public interest alternatives for important AI tools may become increasingly scarce'.[6]

Scarce inputs

As with access to geographically sparse minerals such as antimony, baryte and rare earth elements,[7] many AI systems will depend on proprietary data inputs (such as large volumes of specific training data) that have been captured through access to existing specialised resources.

They may also depend on data collection practices that are hard for smaller companies to replicate ('potentially from multiple sources and labelled or moderated, relating to many use-cases, contexts, and subjects'),[8] and analysed with compute resources (with 'scarce expertise in model training, testing and deployment'[9]) that only a handful of major companies may have.

4 David Gray Widder and Dawn Nafus, 'Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers' Notions of Responsibility' [2023] Big Data & Society http://arxiv.org/abs/2209.09780 accessed 17 January 2023.

5 Engler and Renda (n 1) 1

6 Nur Ahmed, Muntasir Wahed and Neil C Thompson, 'The Growing Influence of Industry in AI Research' (2023) 379 Science 884.

7 Michel Penke, 'China's Dominance of Strategic Resources' Deutsche Welle (13 April 2021) https://www.dw.com/en/how-chinas-mines-rule-the-market-of-critical-raw-materials/a-57148375 accessed 28 February 2023.

8 Jennifer Cobbe, Michael Veale and Jatinder Singh, 'Moving beyond "Many Hands": Accountability in Algorithmic Supply Chains', Proceedings of Fairness, Accountability and Transparency '23 (ACM 2023) 4.

9 ibid.

Again, in concentrated markets, there may be competition law questions about access to high-quality inputs and outputs, including specific datasets and high-end computation capability for training the largest models.

Example: illegally mined gold in hardware supply chains

The use of illegally mined gold from Brazil in technology manufacturing is an example of a supply chain with harmful rule breaking. This can happen despite (as with AI) the existence of supplier codes of conduct and audit processes.

A 2021 Brazilian federal police investigation found that 'companies such as Chimet had been extracting illegal gold from the Kayapó indigenous land since 2015', which potentially 'ended up being used in the manufacture of tablets, phones, accessories for digital devices and even Xbox consoles'.[10] Microsoft and Amazon did not comment publicly, while Apple told reporters about its post hoc system of removing suppliers: 'If a foundry or refiner cannot or does not want to meet our strict standards, we will remove it from our supply chain and, since 2009, we have guided the removal of more than 150 smelters and refineries.'[11]

Google reiterated 'the rigor of [its] Supplier Code of Conduct… demanding the search for ores "only from certified and conflict-free companies"'. But it 'ruled out the adoption of protocols such as the audit by the Responsible Minerals Guarantee Process (RMAP),' which involves an independent third-party assessing a company's supply chain.[12] Organisations like the Responsible Minerals Initiative have developed best practice standards that provide guidance for the responsible sourcing of minerals in a supply chain, but companies will clearly need incentives to adopt them.

The Brazilian federal government, elected in early 2023, is planning legislation, and the Banco Central do Brasil and other government agencies 'have been studying the adoption of the electronic tax receipts for buying and selling gold in order to track whether it was illegally mined'.[13]

10 'Ouro Ilegal Da Amazônia é Ligado a Quatro Big Techs, Aponta PF e MPF' Gazeta Brasil (27 July 2022) https://gazetabrasil.com.br/ciencia-e-tecnologia/2022/07/27/ouro-ilegal-da-amazonia-e-ligado-a-quatro-big-techs-aponta-pf-e-mpf/ accessed 28 February 2023.

11 ibid.

12 ibid.

13 Anthony Boadle and Lisandra Paraguassu, 'Exclusive: Brazil Plans Legislation to Crack down on Laundering of Illegal Gold' Reuters (16 February 2023) https://www.reuters.com/world/americas/brazil-plans-legislation-crack-down-laundering-illegal-gold-2023-02-16/ accessed 28 February 2023.

This example could be considered analogous to requirements for detailed datasheets for AI models, which are documents that list details about a dataset such as: what data is included; how it was sourced; and how it should be used. It also highlights the importance of laws and regulations that establish the appropriate uses of data used to train an AI system, like data protection that covers the use of personal data and copyright law that covers the use of creative works.

Former US Federal Communications Commission Chair Tom Wheeler captured this concern, noting that machine learning 'is nothing more than algorithmic analysis of enormous amounts of data to find patterns from which to make a high percentage prediction. Control of those input assets, therefore, can lead to control of the AI future'.[14]

The EU is attempting to address some of these scarcity issues through its European Data Strategy,[15] including legislation such as the Data Governance Act and Data Act.

Reliance on third-party software components and libraries

Like almost all software, AI systems are likely to be developed making extensive use of software libraries and components from third parties, to 'benefit from the rich ecosystem of contributors and services built up around existing frameworks.'[16]

Researchers have noted: 'Much software is too complex, relying on too many components, for any one person to fully understand or account for its workings.'[17] One analysis of commonly used deep learning frameworks found: 'Caffe is based on more than 130 depending libraries … and TensorFlow and Torch depend on 97 Python modules and 48 Lua

14 Wheeler, (2019), History's message about regulating AI, Brookings Institution, https://www.brookings.edu/research/historys-message-about-regulating-ai/

15 European Commission, 'A European Strategy for Data' (2020) COM/2020/66 final https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX%3A52020DC0066 accessed 16 May 2023.

16 Information Commissioner's Office, 'Guidance on AI and Data Protection' (ICO 2023) https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/ accessed 19 January 2023.

17 Cobbe, Veale and Singh (n 6) 3.

modules respectively."[18]

These components can be used to introduce security vulnerabilities into end-systems. For example, researchers showed how a vulnerability reported in the 'numpy' Python library could be used to cause TensorFlow applications depending on it to crash. Other vulnerabilities could cause AI frameworks to misclassify inputs, and to enable an attacker to remotely compromise a system.[19]

To address these types of vulnerabilities, the US government is now limiting the procurement of 'critical' software unless it complies with standards issued by the US National Institute of Standards and Technology (NIST). This is to 'enhance the security of the software supply chain', including using automated tools to 'check for known and potential vulnerabilities and remediate them'.

The USA will also include standards regarding: 'maintaining accurate and up-to-date data, provenance (i.e., origin) of software code or components, and controls on internal and third-party software components, tools, and services present in software development processes, and performing audits and enforcement of these controls on a recurring basis.'[20]

This type of procedure may need to be considered by the UK and other governments in procuring AI systems for their own use, and in critical national infrastructure.

Data protection

Data protection issues arise wherever personal data is processed in a supply chain. For example, if a UK business purchases a database of marketing contacts from a UK or EU supplier, both parties must comply with data protection law (principally the General Data Protection Regulation (GDPR), which was transposed into UK law following Brexit).

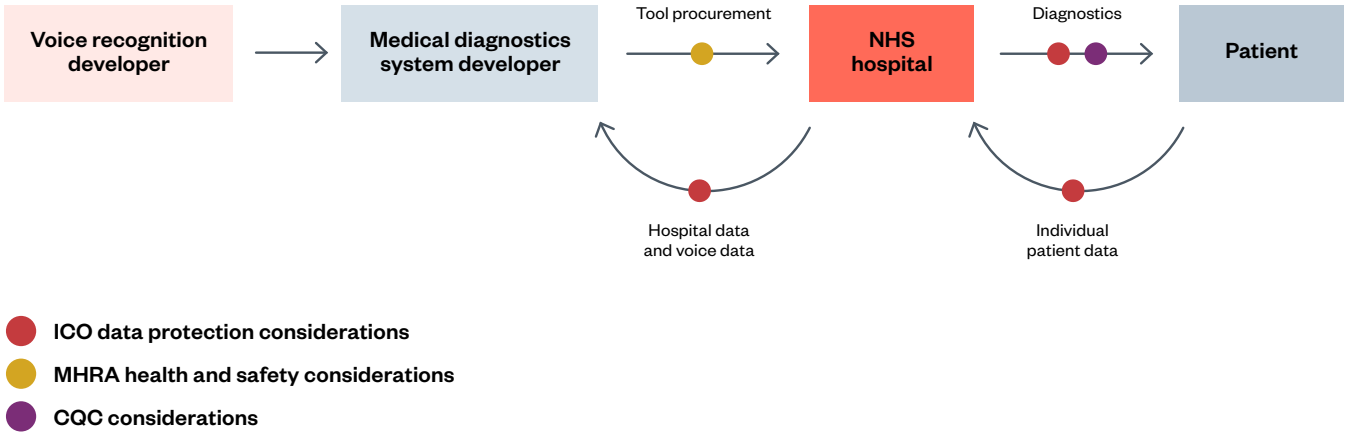
18 Qixue Xiao and others, 'Security Risks in Deep Learning Implementations', 2018 IEEE Security and Privacy Workshops (SPW) (2018) 124.

19 Xiao and others (n 18) 125–126.

20 Executive Order on Improving the Nation's Cybersecurity 2021 (Executive Order 14028).

Figure 2 shows an example supply chain for a medical diagnostics system, where the developer and its hospital customers both process patients' personal data to train and apply cancer detection models and voice recognition models.

Figure 2: AI supply chain for a medical diagnostics system



Description of Figure 2: A developer has created an AI-powered diagnostic tool which can assess an X-ray of a patient's lungs for signs of cancer. They have procured a voice recognition model to incorporate into this tool from another company, which can understand doctors' recorded voice comments on a patient to add to their diagnostics report. This diagnostics tool has been assessed for safety by the Medicines and Healthcare products Regulatory Agency (MHRA), while the models it contains have data sheets and model cards which can be produced to the ICO or EHRC if needed, as well as the Care Quality Commission when the CQC is assessing care provision.

Where personal data is used in either training an AI model or in its application, it raises a range of data protection issues, including data quality, fairness and legality of processing, and fairness of automated decision-making.[21] Where personal data is involved, the technological drive to train ever-larger models with ever-greater quantities of data is in significant tension with the notions of minimisation and purpose specification that are enshrined in current data protection legislation.

21 Information Commissioner's Office, (2023), *Guidance on AI and data protection*, ICO, https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/ (Accessed: 19 January 2023)

There may be trade-offs between principles, for example, data minimisation and statistical accuracy.[22]

This risk is beginning to be identified by jurisdictions around the world. In one of the first AI-related GDPR enforcement decisions, the Italian data protection authority temporarily blocked the US developer of AI chatbot Replika from processing the personal data of Italian users, due to risks to vulnerable users, especially children.[23]

Shortly afterward, the authority temporarily blocked OpenAI's chatbot product ChatGPT from processing the personal data of Italian users until it put in place several protections. This includes: the availability of privacy policies for users of the service and those whose data was used for model training; a clear statement about the legal basis for those uses; the provision of tools for individuals to exercise their privacy rights; and (again) better protection for children.[24]

The European Data Protection Board – made up of the national supervisory authorities – has created a task force to 'foster cooperation and to exchange information on possible enforcement actions' relating to ChatGPT.[25]

For models trained on vast quantities of uncurated data scraped from the web, there could be fundamental issues for GDPR compatibility. These could relate to consent for the processing of sensitive personal data and the question of whether rectification of errors and the 'right to erasure' extend to the (hugely expensive) retraining of models, rather than the suppression of a specific output, as noted in the outcome of a related case on search engines at the EU Court of Justice, *Google Spain*.[26]

22 ibid.

23 Garante per la Protezione dei Dati Personali, 'Provvedimento' https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9852214 accessed 11 March 2023.

24 Giangiacomo Olivi and Chiara Bocchi, 'Generative AI vs Privacy Compliance: Our Five-Point Checklist for the Way Forward' (Dentons Knowledge, 4 May 2023) https://www.dentons.com/en/insights/articles/2023/may/4/generative-ai-vs-privacy-compliance-our-fivepoint-checklist-for-the-way-forward accessed 15 May 2023.

25 European Data Protection Board, 'EDPB Resolves Dispute on Transfers by Meta and Creates Task Force on Chat GPT | European Data Protection Board' (13 April 2023) https://edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt_en accessed 15 May 2023.

26 Lilian Edwards, 'Can ChatGPT Be Compatible with the GDPR? Discuss.' (National Association of Data Protection and Freedom of Information Officers, April 2023) https://www.slideshare.net/lilianed/can-chatgpt-be-compatible-with-the-gdpr-discuss accessed 16 May 2023.

This may partly depend on the specific value of generative AI models to freedom of expression and other fundamental rights, and whether such benefits could be obtained through models trained on much more carefully curated datasets with explicit consent from data subjects.

An internal note allegedly leaked from Google suggests using small, carefully selected datasets can be an effective approach, compared to the use of 'the largest models on the planet' trained on a significant fraction of the entire world wide web.[27]

In the UK, the Information Commissioner's Office (ICO) has identified the high risk of data processing in the context of AI use, publishing guidance which states: 'In the vast majority of cases, the use of AI will involve a type of processing likely to result in a high risk to individuals' rights and freedoms, and will therefore trigger the legal requirement for [organisations] to undertake a data protection impact assessment (DPIA).' This must assess 'risks to the rights and freedoms of individuals, including the potential for any significant social or economic disadvantage'.[28]

Scholars suggest that the current methods for ensuring prevention or mitigation of potential harms are inadequate: for example, the 'right to an explanation' frequently discussed in an AI/GDPR context is not sufficient to deal with often-cited 'algorithmic harms' around fairness, discrimination and opacity.

The scholars suggest the GDPR's right to erasure and data portability, as well as its requirements for data protection by design, impact assessments and certifications/privacy seals, may be a better basis 'to make algorithms more responsible, explicable, and human-centred'.[29]

There are also open questions about the data protection responsibilities of companies in some supply chains under the GDPR, as personal 'data controllers'. Some major developers of AI technologies like Microsoft, Google and Amazon sell these products as a service to other companies

27 Dylan Patel and Afzal Ahmad, 'Google "We Have No Moat, And Neither Does OpenAI"' (semianalysis, 4 May 2023) https://www.semianalysis.com/p/google-we-have-no-moat-and-neither accessed 19 May 2023.

28 Information Commissioner's Office (n 16).

29 Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review https://papers.ssrn.com/abstract=2972855 accessed 4 March 2023.

(this is called 'AI as a Service', or AIaaS). When doing so, they commonly ask customers' permission to use their submitted data to improve their products.

When this includes personal data of end users, although the terms of service of these providers may specify that they are acting as data processors, in reality they may be acting as controllers or joint controllers and face significant GDPR requirements.[30] This remains the case for as long as the UK retains a data protection framework largely mirroring the GDPR, though it is worth noting that the UK Government has proposed a significant update in its Data Protection and Digital Information Bill.

Under these conditions, AIaaS providers will need a specific legal basis for processing to improve their models (and should check where they are joint controllers for normal processing). If data supplied by customers includes special category data (such as health data), explicit consent from data subjects is likely to be the only option available.[31]

There are some open questions to be addressed in relation to these issues.

- How can companies ask customers to provide explicit consent as data subjects 'where they do not directly interact with passive third parties (where customers are, for instance, directly or indirectly surveilling a physical space)'?[32]
- Are providers able to adequately inform data subjects and get their explicit consent?[33]
- And how will data protection authorities verify key details about the use of personal data, given the sparse documentation typically published by developers?

30 Jennifer Cobbe and Jatinder Singh, 'Artificial Intelligence as a Service: Legal Responsibilities, Liabilities, and Policy Challenges' (2021) 42 Computer Law & Security Review 105573, 22.

31 ibid 34.

32 ibid 37.

33 ibid 37.

Copyright

Businesses using copyrighted works anywhere in a supply chain must ensure they have permission from the copyright owners, or qualify under a limited number of 'fair dealing' exceptions under UK copyright law. For example, when an organisation buys photographs from a stock library, both parties are responsible for complying with any licence conditions set by the photograph's owner, or verifying that they qualify for an exception (such as educational use).

There are large quantities of public domain works – text, audio and video – that are available for training AI models and are not subject to copyright, as well as copyright material released under licences permitting use (although these rarely give explicit permission for use in training AI models). Where other copyrighted works are used, questions of fair dealing and licensing will arise – as well as issues relating to the outputs of generative models.[34]

In the USA, there are mixed academic views as to whether their broad 'fair use' copyright exception would allow the use of copyrighted works to train models without the copyright holder's consent.[35] International newspapers including The Economist have argued that companies such as Microsoft, which is now showing AI-produced summaries of articles in its search engine Bing, should have to license the use of such content.[36]

In the UK, a limited copyright exception allows text and data mining (TDM) for research for non-commercial purposes (section 29A of the Copyright, Designs and Patents Act 1988). The UK Intellectual Property Office (IPO) has proposed significantly widening this exemption,[37] and the UK Government has accepted the recommendations of the review by its Chief Scientific Adviser suggesting that it 'should work with the AI and creative industries to

34 Andrés Guadamuz, 'A Scanner Darkly: Copyright Infringement in Artificial Intelligence Inputs and Outputs' (26 February 2023) https://papers.ssrn.com/abstract=4371204 accessed 26 February 2023.

35 ibid 16–21.

36 'Artificial Intelligence Is Reaching behind Newspaper Paywalls' [2023] The Economist https://www.economist.com/business/2023/03/02/artificial-intelligence-is-reaching-behind-newspaper-paywalls accessed 4 March 2023.

37 Intellectual Property Office, 'Artificial Intelligence and Intellectual Property: Copyright and Patents: Government Response to Consultation' (2022) https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation accessed 8 March 2023.

develop ways to enable TDM for any purpose'.[38]

The EU has already introduced such a reform, also allowing commercial TDM on an opt-out basis, in Article 4 of the Copyright in the Digital Single Market Directive (2019/790/EU).

The spawning.ai website has gathered opt-outs for 78 million works, mainly from organisations such as Shutterstock. The developers of one of the most popular text-to-image AI systems, Stable Diffusion, announced they will honour these opt-outs in training future models.[39] However, European artists' associations have argued this is insufficient protection, and have lobbied for the EU's AI Act to require explicit, informed consent from the authors of works.[40]

Current copyright regimes are unlikely to find images produced by AI systems such as Stable Diffusion 'in the style of' a specific artist to be an infringement of copyright. But they will give protection to AI-generated images which contain copyrighted elements, such as cartoon superheroes.[41]

The US Copyright Office has issued policy guidance that under US law, copyright will apply to the human-authored elements of AI-generated works where an author has 'select[ed] or arrange[d] AI-generated material in a sufficiently creative way', or 'modif[ied] material originally generated by AI technology to such a degree that the modifications meet the standard for copyright protection'.[42]

Human rights

Businesses' responsibility to respect human rights throughout their supply chains is enshrined under the United Nations' Guiding Principles

38 Patrick Vallance, 'Pro-Innovation Regulation of Technologies Review' (2023) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1142883/Pro-innovation_Regulation_of_Technologies_Review_-_Digital_Technologies_report.pdf accessed 17 March 2023.

39 Spawning.AI, 'We Are Thrilled to Announce That Our Campaign to Gather Artist Opt Outs Has Resulted in 78 Million Artworks Being Opted out of AI Training. This Establishes a Significant Precedent towards Realizing Our Vision of Consenting AI, and We Are Just Getting Started!' https://twitter.com/spawning_/status/1633196665417920512 accessed 8 March 2023.

40 Molly Killeen, 'Generative AI Keeps Creative Industries on Their Toes' EURACTIV (27 January 2023) https://www.euractiv.com/section/artificial-intelligence/news/generative-ai-keeps-creative-industries-on-their-toes/

41 Guadamuz (n 33).

42 US Copyright Office, 'Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence' https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence accessed 29 March 2023.

on Business and Human Rights, which says they should 'avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved'.[43]

In the UK and EU, data protection regulation further states that 'data protection aims to protect individuals' rights and freedoms with regard to the processing of their personal data, not just their information rights' – including, for example, the right to non-discrimination.[44]

This means that, particularly where governments are using AI systems for decision-making, they will need to consider the full range of human rights issues raised. As a recent Chatham House report concluded: 'While human rights do not hold all the answers, they ought to be the baseline for AI governance. International human rights law is a crystallization of ethical principles into norms, their meanings and implications well-developed over the last 70 years.'[45]

For example, state responses to AI-generated disinformation must be carefully guided by the impact on freedom of expression, as, for example, over-moderating online content can inadvertently remove valuable political speech.[46] Similarly, the use of AI-based tools such as facial recognition and gunshot detection by law enforcement and national security agencies can have serious impacts on privacy and equality, particularly for some marginalised groups.

Associated Press found that in the USA, the widely used ShotSpotter system 'is usually placed at the request of local officials in neighborhoods deemed to be the highest risk for gun violence, which are often disproportionately Black and Latino communities'.[47]

---

43 John Ruggie, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework (United Nations High Commissioner for Human Rights 2012) 13 https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights accessed 22 March 2023.

44 Information Commissioner's Office (n 14).

45 Kate Jones, 'AI Governance and Human Rights' (2023) https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights accessed 4 March 2023.

46 Chris Marsden, Trisha Meyer and Ian Brown, 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?' (2020) 36 Computer Law & Security Review 105373.media pluralism and the exercise of democracy, from the wider lens of tackling illegal content online and concerns to request proactive (automated

47 Garance Burke and Michael Tarm, 'Confidential Document Reveals Key Human Role in Gunshot Tech' Associated Press News (20 January 2023) https://apnews.com/article/shotspotter-artificial-intelligence-investigation-9cb47bbfb565dc3ef110f92ac7f83862 accessed 21 March 2023.

The private sector must 'respect' human rights,[48] for example avoiding discrimination by using AI-based employment tools, which is also required by specific equality and anti-discrimination laws in many countries, such as the UK's Equality Act 2010.[49]

This may be a more likely outcome where sectoral regulators are enforcing these duties – for example, in the UK finance sector compared to the human resources sector – even where cross-sectoral regulators are in place, such as the UK Information Commissioner's Office (ICO) and the Equality and Human Rights Commission (EHRC).[50]

Some of these issues are starting to be specifically addressed at the international level in the Council of Europe's draft AI Convention,[51] building on regional human rights instruments such as the European Convention on Human Rights.

In the EU, civil society groups have called for the extension of the AI Act's outright prohibition on AI systems that pose an unacceptable risk to fundamental rights, including 'social scoring systems; remote biometric identification in publicly accessible spaces (by all actors); emotion recognition [and] systems to profile and risk-assess in a migration context'.[52]

Distinctive features of AI supply

Distinctive features of AI supply chains that are potentially of interest to regulators include items such as components' complexity, speed of change or opacity, and their diffuse impact on people and society.

48 Jones (n 48).
49 Equality Act 2020 (c.15).
50 Centre for Data Ethics and Innovation, 'Industry Temperature Check: Barriers and Enablers to AI Assurance' (2022) https://www.gov.uk/government/publications/industry-temperature-check-barriers-and-enablers-to-ai-assurance/industry-temperature-check-barriers-and-enablers-to-ai-assurance accessed 21 March 2023.
51 Victoria Hendrickx and Peggy Valcke, 'The Council of Europe's Road towards an AI Convention: Taking Stock' (25 January 2023) https://www.law.kuleuven.be/ai-summer-school/blogpost/Blogposts/AI-Council-of-Europe-draft-convention accessed 21 March 2023.
52 European Digital Rights, 'Civil Society Calls on the EU to Put Fundamental Rights First in the AI Act' (30 November 2021) https://edri.org/our-work/civil-society-calls-on-the-eu-to-put-fundamental-rights-first-in-the-ai-act/ accessed 7 March 2023.

Features of models and systems

> The novelty, complexity and speed of ongoing change and adaptation of AI models will make it difficult to standardise or even precisely specify their features – hence the importance of transparency tools like transparency registers, model cards, datasheets and other methods for sharing information about a system.[53]

AI systems can be highly opaque, leading to a lack of standardisation and societal understanding. Even before the proliferation of AI systems, researchers noted 'the lack of means, legal or technical, for uncovering the nature of the supply chains on which online services rely.'[54]

Countries including France and the Netherlands are partly addressing this by setting up algorithm registers and regulators (in these two cases, within their data protection authorities).[55] Spain is setting up an independent agency to monitor the public and private sector's compliance with the EU's forthcoming AI Act, which will require 'high-risk' systems to be registered in a public database.[56]

Seven European city administrations (Barcelona, Bologna, Brussels Capital Region, Eindhoven, Mannheim, Rotterdam and Sofia) have been developing an Algorithmic Transparency Standard, which will 'help people understand how the algorithms used in local administrations work, and what their purpose is.'[57]

---

53 For a summary of these tools, see: Ada Lovelace Institute, (2023), Mechanisms for assessing and mitigating risks that AI systems pose for people and society. Forthcoming.

54 Open Government Partnership. Building Public Algorithm Registers: Lessons Learned from the French Approach (2021). Available at: https://www.opengovpartnership.org/stories/building-public-algorithm-registers-lessons-learned-from-the-french-approach/; Government of Netherlands. Het Algoritmeregister van de Nederlandse overheid (2022). Available at: https://algoritmes.overheid.nl/;

55 Open Government Partnership. Building Public Algorithm Registers: Lessons Learned from the French Approach (2021). Available at: https://www.opengovpartnership.org/stories/building-public-algorithm-registers-lessons-learned-from-the-french-approach/; Government of Netherlands. Het Algoritmeregister van de Nederlandse overheid (2022). Available at: https://algoritmes.overheid.nl/;

56 Pablo Jiménez Arandia, 'What to Expect from Europe's First AI Oversight Agency' (AlgorithmWatch, 1 February 2023). https://algorithmwatch.org/en/what-to-expect-from-europes-first-ai-oversight-agency/ accessed 17 March 2023.

57 Alex Godson, 'Nine Cities Set Standards for the Transparent Use of Artificial Intelligence' (Eurocities, 19 January 2023) https://eurocities.eu/latest/nine-cities-set-standards-for-the-transparent-use-of-artificial-intelligence/ accessed 21 March 2023.

AI systems can impose significant, diffuse external costs on a range of downstream actors

The UK has also introduced a public-sector Algorithmic Transparency Standard for use by public sector organisations.[58] Notably, public sector organisations will not be legally required to use this standard, under the latest language of the Data Protection and Digital Information Bill (DPDIB).

Opacity is an issue, because it can be reinforced by companies to protect commercial secrets. An Associated Press review of potential serious miscarriages of justice resulting from the use of one company's gunshot detection system in the USA found that it 'guards how its closed system works as a trade secret, a black box largely inscrutable to the public, jurors and police oversight boards'.

Even in court cases, it 'has shielded internal data and records revealing the system's inner workings, leaving defence attorneys no way of interrogating the technology to understand the specifics of how it works'.[59] The EU's proposed AI Liability Directive will enable courts to order disclosure of evidence from a provider or user of a high-risk system subject to the AI Act, if the system is 'suspected of having caused damage' (Article 3(1)).[60]

Impact on companies, people and society

More broadly, AI systems can impose significant, diffuse external costs on a range of downstream actors, including intermediary companies and members of the public (such as the impact of a proliferation of disinformation on the quality of democratic debate). These impacts can extend beyond the direct users of such systems.

58 Central Digital and Data Office and Centre for Data Ethics and Innovation, 'Algorithmic Transparency Recording Standard Hub' (GOV.UK, 5 January 2023) https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub accessed 22 March 2023.

59 Burke and Tarm (n 47).

60 European Commission, Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) 2022 [2022/0303 (COD)].

Policymakers and regulators need feedback loops along AI supply chains to identify harms, errors or issues

> It can be difficult to attribute responsibility and legal liability for harms resulting from complex supply chains, given the potentially significant effects in some systems of even small changes by one node in the chain (the 'many hands' problem).[61]

Some emerging policy proposals have sought to address this issue.

European policymakers have debated following 'established practices in other sectors, such as pharmaceutical and chemical products, where producers' liability is normally exempted in cases of misuse, but producers are increasingly prompted to think about "reasonably foreseeable misuses".'[62]

In the broader case of platform regulation, researchers have suggested the idea of 'cooperative responsibility', with a division of responsibility between actors depending on 'the capacities, resources, and knowledge of both platforms and users, but also on economic and social gains, incentives and arguments of efficiency, which vary from sector to sector and case to case.'[63]

In the closely related area of software liability, the US's Biden administration has adopted a cybersecurity strategy which will 'ask more of the most capable and best-positioned actors'.[64] The USA will move towards 'preventing manufacturers and service providers from disclaiming liability by contract, establishing a standard of care, and providing a safe harbour to shield from liability those companies that do take reasonable measurable measures to secure their products and services'.[65]

61 Cobbe, Veale and Singh (n 8).
62 Engler and Renda (n 1) 24.
63 Natali Helberger, Jo Pierson and Thomas Poell, 'Governing Online Platforms: From Contested to Cooperative Responsibility' (2018) 34 The Information Society 1.
64 'National Cybersecurity Strategy' (White House 2023) https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf accessed 8 March 2023.
65 Jim Dempsey, 'Cybersecurity's Third Rail: Software Liability' (Lawfare, 3 March 2023) https://www.lawfareblog.com/cybersecuritys-third-rail-software-liability accessed 3 March 2023.

Under EU law, scholars have noted that AI system providers are 'not protected by the E-Commerce Directive from liability for their customers' activities, since they are not mere conduits, caches or hosts, and anyway cannot be said to be operating without knowledge of their customers' activities.'[66]

Other scholars warn that US and EU regulatory initiatives do not cover the risk of 'black swan' events, which are potentially catastrophic but low probability. These include 'high-impact accident risks from general purpose AI systems; the uncontrolled proliferation and malicious use of AI systems; and applications of AI that could cause long-term systemic harm to social and political institutions'.[67]

To address these risks, policymakers and regulators need accelerated feedback loops along AI supply chains. These should identify harms, errors or issues, and should include a mechanism for feedback up the supply chain, for action by the relevant actor.

A regulator could put in place ex ante requirements for the design and testing of an AI system and conduct ex post evaluations of a system's actual performance. For example, the Financial Conduct Authority (FCA) and/or Equalities and Human Rights Commission (EHRC) could place requirements in relation to fairness testing of mortgage assessment processes, and take ex post enforcement action if the system were found to still be biased. A regulated bank may need to work with its suppliers, potentially all the way up a supply chain, to ensure it can meet those requirements.

Research by the UK's Centre for Data Ethics and Innovation found industry participants were keen to understand how the use of assurance tools such as impact assessments and certifications would help meet regulatory obligations, which would be a 'key motivator for industry engagement' in the continued use of these tools.[68]

66 Cobbe and Singh (n 30).

67 Noam Kolt, 'Algorithmic Black Swans' (2023) 101 Washington University Law Review 31 https://papers.ssrn.com/abstract=4370566 accessed 10 March 2023.

68 Centre for Data Ethics and Innovation (n 50).

Examples of different kinds of AI supply chains

In a key analysis, a research team at the Centre for European Policy Studies identified the following seven basic configurations making up AI supply chains, containing one, two or three developers/suppliers, in increasing order of complexity.[69] These supply chains vary in terms of who has the skills, power and access needed to assess and mitigate risks. In more complex supply chains with multiple suppliers or developers, these responsibilities can be less clear.

Throughout this chapter, we make a distinction between **deployers** of an AI system (those who actually use it) and **developers** of an AI system (those responsible for training and creating the system), although elements of these roles can merge. **End users** refer to people or businesses who are ultimately affected by a deployed system.

Below, we distinguish between three broad categories of AI supply chains:

1. systems built in-house
2. systems relying on an API
3. systems built for a customer (or fine-tuned for one).

69 Engler and Renda (n 1) 6–14.

Systems built in-house

1. [One actor] A company develops AI systems in-house, using their own staff, software and data. This provides the company with the maximum level of control over the system, and clear responsibility to assess and mitigate risks. This configuration is most likely for specialised use cases with limited economies of scale in production and operation. This scenario implies that the company will have the staff needed to monitor and mitigate resulting risks, *if* those staff are retained beyond initial deployment (and this monitoring and evaluation is not outsourced to a third party).

Systems relying on an API

2. [Three or more actors] A developer company buys AI systems and components from several other companies, integrating them into a complete system of its own before supplying it to companies and end users directly or via an API. The developer/deployer company may have a high-level understanding of these systems, but may not have specialist AI expertise to monitor and mitigate resulting risks.

3. [Two actors] A company develops and trains an AI system which a second company accesses by sending queries via a limited API. This gives the developer a high level of control over how its system is used, including the potential to include technical as well as legal restrictions on prohibited uses. The customer/deployer company may have a high-level understanding of the system, but may not have specialist AI expertise to monitor and mitigate resulting risks, or to correct errors in the underlying model.

Systems built for a customer (or fine-tuned for one)

4. [Two actors] A company deploys a system custom-developed by another company under contract. This gives the deploying company a high level of control but presents some challenges in its ability to assess and monitor for risks. The contracting/developer company may specify permitted and prohibited uses of the system in the contract – although may have limited resources to monitor and enforce how the deploying company uses it. The customer/deployer company will be likely to have a high-level understanding of the system but may not have specialist AI expertise to monitor and mitigate resulting risks.

5. [Two actors] A developer company sells a complete AI system to a second company, which inputs its own data to enable the system to undertake additional training, and deploys it. This provides the first developer company with *some* higher level of control over the use of the resulting system. This scenario implies that the deploying company will have the staff needed to monitor and mitigate resulting risks, *if* they are retained beyond initial deployment.

6. A developer company sells a complete AI system, including direct access to the underlying model(s), which the second company can access and train using its own data. The developer has a lower level of knowledge and control over the use of the resulting model. This scenario implies that the deploying company will have the staff needed to monitor and mitigate resulting risks, *if* they are retained beyond initial deployment.

7. An AI system developer sells code to a deploying company, which uses it along with its own data to train and deploy a specific type of model. The system developer has a lower level of knowledge and control over the use of the resulting model by the deploying company. This scenario implies that the deploying company will have the staff needed to monitor and mitigate resulting risks, *if* they are retained by the deploying company beyond initial deployment.

Open-source components

In all these cases, companies may incorporate resources (including data, software and models) released under open-source licences. These licences grant companies and researchers the freedom to use, examine and modify these resources as they wish (discussed further in from page 56).

Where the use of such resources is business-critical, companies may choose to pay for external support from specialist providers of open-source tools such as Red Hat.[70] Where companies have the expertise, they will be able to modify open-source components directly to fix faults – and (if they choose, or are required to by the licence) contribute those fixes back to the community using and maintaining the component. This can create some ambiguity about which parties are providers/developers of an AI system and which are the deployers.

70 Red Hat, 'Siemens Improves Uptime and Security with Red Hat OpenShift' (15 July 2022) https://www.redhat.com/en/resources/siemens-amberg-case-study accessed 22 March 2023.

Assurance intermediaries

A final consideration for assigning responsibility for assessing and mitigating risks in an AI supply chain is whether a third-party organisation (such as a law firm, auditing agency or consultancy) has taken on a contractual obligation to manage some risks.

Third-party organisations can conduct independent certifications, audits and other processes to provide additional information about AI components, which would give assurances to the public, regulators and companies making use of them.[71]

Companies that perform these kinds of third-party evaluations of an AI system are a major part of the UK's strategy for the development of trustworthy AI systems, with the Government's Centre for Data Ethics and Innovation producing a roadmap towards an effective ecosystem conducting AI assurance.[72]

71 For further information about risk assessment methodologies, see: Ada Lovelace Institute, (2023), Mechanisms for assessing and mitigating risks that AI systems post for people and society, [Internal briefing for DCMS]

72 Centre for Data Ethics and Innovation, 'From Roadmap to Reality: Insights from Industry on Advancing AI Assurance' (7 December 2022) <https://cdei.blog.gov.uk/2022/12/07/from-roadmap-to-reality-insights-from-industry-on-advancing-ai-assurance/> accessed 21 March 2023.

A conceptual framework for regulators to apply to AI supply chains in their sector

Policymakers and regulators must make difficult choices when determining where to assign distinct responsibilities for addressing the risks that can arise throughout an AI system's supply chain. Below, we provide an initial conceptual framework that regulators can build from to determine where responsibilities might apply, which relies on four principles:

1. **Transparency:** what information can each actor in a supply chain provide to enable risks to be identified and addressed.
2. **Incentivisation:** who is best incentivised to address these risks, and how can regulators create those incentives while minimising the overall costs of fixing problems.[73]
3. **Efficacy:** who is best positioned to most effectively address the risks that can emerge from an AI system (potentially multiple parties working together).
4. **Accountability:** how can the use of contracts assign responsibilities, and what are the limitations of this method.

Transparency

> To ensure effective regulation, regulators and policymakers will need to incentivise transparency and information flow across the supply chain.

This will allow information about, and evaluation of, systems and potential risks to travel up and down chains, supporting remediation of identified problems.

73 Timnit Gebru and others, 'Datasheets for Datasets' (2021) 64 Communications of the ACM 86.

Mechanisms needed to ensure this flow of information, including via contractual terms and regulatory requirements on all actors in a supply chain, include:

- Transparency and accountability processes, including mechanisms such as model cards and datasheets which provide information on an AI model's architecture and the data it was trained on. These 'have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks'.[74]

- Certifications, audits, impact assessments, technical standards and similar mechanisms, which give organisations reliable evidence on the trustworthiness of AI systems.[75] These mechanisms seek to establish standardised processes for organisations to evaluate and monitor the behaviour of their AI systems, and other aspects that are important for fulfilling their regulatory duties, and are important to their end users. For example, the Centre for Data Ethics and Innovation found respondents from the connected and automated vehicle sector were keen to see the development of certification and kite-marking mechanisms 'to demonstrate their compliance to customers'.[76]

- Requirements for sector-specific information sharing, like the UK's Cyber Security Information Sharing Partnership. Similar efforts for AI could potentially be facilitated by regulators. Fora like these could also develop voluntary sectoral codes of conduct, building on those envisaged in the GDPR's Articles 40 and 41, and developing standards for certifications.[77]

- Requirements to share data with insurers and regulators, that are modelled on other domains like cybersecurity. A US review found 'a lack of data, a lack of expertise, and an inability to scale rigorous security audits have rendered cyber insurers unable to play a significant deterrent role in reducing cybersecurity incidents or

74 ibid.
75 Centre for Data Ethics and Innovation (n 50).
76 ibid 37.
77 Edwards and Veale (n 29) 70–80.

exposure to cyber risks.' The review highlights the approach of the Singaporean government in improving this issue.[78]

- Mechanisms for reporting and remedying faults. Researchers from Stanford's Human-Centered AI project suggested: 'If downstream users have feedback, such as specific failure cases or systematic biases, they should be able to publicly report these to the developer, akin to filing software bug reports. Conversely, if a model developer updates or deprecates a model, they should notify all downstream users' including deployers or end users whose products and services rely on that model.[79] This would require a mechanism to keep track of all such users, which may not be straightforward where models or components may be downloaded and used without specific notification to their developer. However, the GDPR's Article 17(2) attempts to deal with this issue (in this case, in terms of secondary use of personal data), by mandating that the main party takes reasonable steps to inform other parties.

More broadly, it may be most efficient for a government body to play a cross-sectoral role for information-sharing and learning.[80] In the Netherlands, for example, an algorithm regulator, situated within the Data Protection Authority, 'will identify cross-sector risks related to algorithms and AI and will share knowledge about them with the other regulators. It will also, in cooperation with already existing regulators, publish and share guidance related to algorithms and AI with market parties, clients and governments.'[81]

78 Shauhun Talesh, 'Cyber Insurance and Cybersecurity Policy: An Interconnected History' (Lawfare, 4 November 2022) https://www.lawfareblog.com/cyber-insurance-and-cybersecurity-policy-interconnected-history> accessed 23 March 2023.

79 Percy Liang and others, 'The Time Is Now to Develop Community Norms for the Release of Foundation Models' (Stanford University Human-Centered Artificial Intelligence, 17 May 2022) https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models

80 For a greater discussion on AI monitoring, see: Ada Lovelace Institute (2023), Approaches to government monitoring of the AI landscape. [Forthcoming]

81 Martijn Schoonewille and others, 'Introduction New Algorithm Regulator and Implications for Financial Sector' Lexology (5 January 2023) https://www.lexology.com/library/detail.aspx?g=3e71f01b-2cb7-4294-b8f2-68ea2ab67261 accessed 20 January 2023.

Transparency in a supply chain: Supply Chain 1

Scenario: *A developer company sells a complete AI system to a second company, which inputs its own data to enable the system to undertake additional training. It then deploys the system. This provides the first developer company with some higher level of control over the use of the resulting system. This scenario implies that the deploying company will have the staff needed to monitor and mitigate resulting risks, if they are retained by the company beyond initial deployment.*

Using this example, the deployer has modified the system in ways that might explain some of its behaviour. However, to fully assess these risks, it is likely that the deployer will still require access to information about how the developer's model was trained, what data it used, and how it was tested. Without this information, it will be challenging for either the deployer or the developer to assess this system holistically for potential risks and mitigate them. If the developer has used transparency mechanisms like model cards and datasheets to share critical information about how the underlying model was trained, that *may* be enough for the deployer to take on more responsibility for addressing or mitigating potential risks. Conversely, if the deployer spots an issue in the developer's model, transparency mechanisms could enable that information to pass back up the supply chain and for the developer to address the risk.

Regulators can play an important role by incentivising information transfer and transparency practices in AI supply chains, while noting its limitations: making information transparent does not necessarily mean it will be acted upon.

These bodies can collaborate internationally in venues such as the Organisation for Economic Co-operation and Development and Council of Europe. Centre for Data Ethics and Innovation research found participants were keen for regulators to coordinate internationally, to encourage consistency and 'to ensure the alignment of national and international standards objectives'.[82]

The EU's AI Act will significantly rely on the production of technical standards for AI systems by bodies such as CEN and CENELEC.

82 Centre for Data Ethics and Innovation (n 50).

However, there are problems with regulatory regimes relying too heavily on technical standards with a significant impact on fundamental rights, which the private organisations producing have little experience and less legitimacy in managing[83] (attempts to improve this have faced significant obstacles[84]).

So-called 'explainable' AI (XAI) systems may help with allocation of responsibility, in that developing systems that 'can explain their "thinking" will let lawyers, policymakers and ethicists create standards that allow us to hold flawed or biased AI accountable under the law.'[85] However, some researchers have noted the limitations of current XAI approaches, which demonstrate how explanations can be brittle and their meaning can change over time.[86]

Relatedly, research suggests that the most important accountability mechanism for AI systems will be preserving a snapshot of the state an AI system at the time when a harm occurs. AI systems can change with new inputs or tweaks to their architecture. This means saving time-stamped versions of systems so that the cause of harms can be examined later, as happens already with self-driving vehicles.[87]

Finally, regulators and policymakers must acknowledge the limits of transparency. Simply making information about AI systems, data or risks available does not mean that information will be acted on by relevant parties. Regulation must create proportionate incentives and penalties for them to do so.

83 Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22 Computer Law Review International 97, 105.

84 C Cath, 'Changing minds and machines: a case study of human rights advocacy in the Internet Engineering Task Force (IETF)' (http://purl.org/dc/dcmitype/Text, University of Oxford 2021) https://ora.ox.ac.uk/objects/uuid:9b844ffb-d5bb-4388-bb2f-305ddedb8939 accessed 22 May 2023.

85 Mason Kortz and Finale Doshi-Velez, 'Accountability of AI Under the Law: The Role of Explanation' (Berkman Klein Center 2017) https://cyber.harvard.edu/publications/2017/11/AIExplanation

86 de Bruijn, H., Warnier, M. and Janssen, M. (2022) 'The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making', Government Information Quarterly, 39(2), p. 101666. https://doi.org/10.1016/j.giq.2021.101666;

87 Joanna J Bryson, 'The Past Decade and Future of AI's Impact on Society', Towards a New Enlightenment? A Transcendent Decade (Turner Libros 2019) https://www.bbvaopenmind.com/en/articles/the-past-decade-and-future-of-ais-impact-on-society/ accessed 4 March 2023.

Incentives, penalties and value chains

Regulators can also incentivise those who are best placed to address emerging risks in an AI supply chain. This approach reduces the risk of a 'diffusion of responsibility' across a supply chain, which can potentially lead to an insufficient consideration of risks by any actor.[88]

Current corporate practices often do not align with incentives to produce systems that prioritise societal benefit. In interviews with 27 AI practitioners, scholars found a 'deeply dislocated sense of accountability, where acknowledgement of harms was consistent but nevertheless another person's job to address, almost always at another location in the broader system of production, outside one's immediate team... Current responsible AI interventions, like checklists, model cards, or datasheets ask practitioners to map the technology to their end use. They attempt to put "out of scope" harms back in scope, but here we show how contemporary software production practice works against that attempt, leaving developers in a bind between countervailing cultural forces.'[89]

Moving away from thinking about a supply chain (the resources and actors that move a product from supplier to customer), towards a value chain model (which considers how value can be added along this supply chain for different actors) could be beneficial for regulators. This shift in thinking would encourage regulators to consider how to set particular incentives and penalties for addressing harms within a supply chain.

One suggested approach for strengthening collaboration to address harms along a supply chain is to use a communication tool, such as a model card or a datasheet, that can help create this shift towards a value chain. In this scenario, a model card 'is not a one-and-done affair, but a place where partiality comes together and relations occur, making the relationship less like a supply chain and more like a value chain where collaborators are working together to collectively address the problem'.[90]

88 John M Darley and Bibb Latane, 'Bystander Intervention in Emergencies: Diffusion of Responsibility' (1968) 8 Journal of Personality and Social Psychology 377.

89 Widder and Nafus (n4).

90 ibid.

Incentives in a value chain: Supply Chain 2

Scenario: *An AI system developer sells components of an AI system (such as code) to a deploying company, which uses it along with its own data to train and deploy a specific type of model. The developer has a lower level of knowledge and control over the use of the resulting model by the deploying company. This scenario implies that the deploying company will have the staff needed to monitor and mitigate resulting risks, if they are retained by the deploying company beyond initial deployment.*

In this example, regulators could focus on which actors – the developer and the deploying company – can be incentivised to proactively assess risks and take action to mitigate them. While the deployer in this case has more information about the final AI system's architecture and data, some risks may be a result of an error or issue in the upstream developer's code.

Regulators may want to incentivise both parties to undertake risk assessments and model evaluations, and engage in transparency mechanisms like datasheets or model cards. To do this, with statutory authority, regulators could require the deployer to ensure that both they, and any upstream developers, have undertaken these risk assessments.

Conversely, if they had legal authority to do so, regulators could place the onus on the upstream code developer. The upstream code developer could require the downstream deployer to only use this code if they agree to undertake specific risk assessments and evaluations of the final AI system.

In a UK scenario, it is likely that deployers will be UK-based and developers could be outside the jurisdiction of UK regulators, raising questions of enforcement against the latter.

Interfaces along a supply chain could be strengthened through the use of contracts that specify clear responsibilities and increase communication between non-developers: 'Those playing customer roles in the supply chain might routinize asking suppliers for model cards, if the data it was trained on was properly consented, if crowd workers labelling the data were paid an appropriate wage, etc., which is commonplace in supply chains for physical goods'.[91]

91 ibid.

Companies and regulators should pay attention to what information and/or responsibility could potentially be lost between the nodes in a supply chain, particularly as a supply chain becomes more complex. As some scholars have noted, 'Thinking interstitially means moving away from the binary question of "do I have full control, or not?" and reasoning in probabilities and frictions. What does the technology make easier or harder, faster or slower, and what do contractual obligations or marketing messages make easier or harder?'.[92]

Efficacy up and down the AI supply chain

> Regulators and policymakers must also consider who in a value chain can most easily identify risks, and who is best-placed to take action to mitigate them.[93]

In an open letter describing how the proposed EU AI Act's ex ante legal requirements for assessing the risk and quality of an AI system should apply to complex AI systems, a group of European civil society organisations argue that shifting the obligations entirely to downstream users in a supply chain 'would make these systems less safe.'[94]

This is because downstream companies are likely to lack the capacity, skills and access to make any changes. However, the letter also argues that downstream companies deploying the system are best placed to comply with other requirements of the act like 'human oversight, but also any use case specific quality management process, technical documentation and logging, as well as any additional robustness and accuracy testing.'[95] This is because downstream deployers are closer in proximity to the final context in which the system is operating.

92 ibid.
93 Engler and Renda (n 1) 24.
94 Access Now et al., 'Call for Better Protections of People Affected at the Source of the AI Value Chain' (25 October 2022) https://futureoflife.org/wp-content/uploads/2022/10/Civil-society-letter-GPAIS-October-2022.pdf accessed 21 March 2023.
95 ibid.

Considering efficacy in the supply chain: Supply Chain 3

Scenario: *A company develops and trains an AI system which a second company accesses by sending queries via a limited API. This gives the developer a high level of control over how its system is used, including the potential to include technical as well as legal restrictions on prohibited uses. The customer/deployer company may have a high-level understanding of the system, but may not have specialist AI expertise to monitor and mitigate resulting risks.*

In this example, a regulator could consider which actor – the company developing the system or the deployer who uses it – can most easily identify and take actions to address risks. In this case, the developer is the only party with control over the system's data and model architecture, meaning that any tweaks or changes to the system will have to be made by them.

The use of an API also gives the developer greater control to prohibit certain uses and even monitor actual uses by the deployer. By considering the principle of efficacy, a regulator could assign responsibility for identifying and addressing risks primarily to the developer – but this would require a regulator to have the necessary powers to do so.

Sometimes it may not be possible for the parties closest to the person affected by an AI system to deal with risks in a manageable way. This is not a new problem: in a case that later became influential across the USA, the New York Court of Appeals found in 1916 that Buick Motor Company 'was not at liberty to put [a car] on the market without subjecting the component parts to ordinary and simple tests' (MacPherson vs Buick, 217 NY 382) since 'neither the consumer nor the local dealership' they acquired it from 'had meaningful insight into or control over the manufacturing process or material supply chain'.[96]

Some researchers have made a more recent assessment of the decision's relevance to regulation of 5G networks, which also has echoes for AI regulation, although these products are at a much earlier stage of development: 'The decision firmly placed the risk assessment and mitigation responsibility with the corporation in the best position to know details regarding assembled sub-systems and to control the

96 Dempsey (n 65).

processes that would address risk factors.'[97]

The US legal approach to AI is likely to evolve via 'multiple specific (and often simultaneous) theories of liability that can be asserted in a products liability claim, including negligence, design defects, manufacturing defects, failure to warn, misrepresentation, and breach of warranty'.[98]

In US state law, 'risk-utility tests have long been employed in products' liability lawsuits to evaluate whether an alleged design defect could have been mitigated "through the use of an alternative solution that would not have impaired the utility of the product or unnecessarily increased its cost".'[99]

Other researchers have expanded on these tests to cover more complex networks of liability, concluding that a 'strict liability regime is the best-suited regime to apply when AI causes harm and will provide indicators to identify the entity who should be held strictly liable.'[100]

Another scholar suggests: 'This same [risk utility] test can be applied in relation to AI as well; however, the mechanics of applying it will need to consider not only the human-designed portions of an algorithm, but also the post-sale design decisions' and aspects of a system that automatically update as new data is fed into it.[101] These tests may offer a way for regulators to make clearer allocations of responsibility.

Finally, when considering efficacy, regulators may need to pay special attention to the jurisdiction where a company or supplier is operating. In some supply chains, it may be easier for regulators to create incentives for suppliers that sit within their own jurisdiction.

97 Tom Wheeler and David SImpson, 'Why 5G Requires New Approaches to Cybersecurity' (Brookings Institution 2019) https://www.brookings.edu/research/why-5g-requires-new-approaches-to-cybersecurity/ accessed 21 March 2023.

98 John Villasenor, 'Products Liability Law as a Way to Address AI Harms' (Brookings Institution 2019).

99 ibid.

100 Anat Lior, 'The AI Accident Network: Artificial Intelligence Liability Meets Network Theory' (2021) 95 Tulane Law Review 1103.

101 Villasenor (n 102).

Accountability through contracts

Companies offering products and services to the market that contain or are based on AI components will generally bear the legal liability of doing so. Where courts or regulators fine or order compensation payments against such companies, they will in turn need to examine whether their suppliers should be responsible for some (or all) of these remedies.

As researchers have observed: 'Apportioning blame within the supply chain will involve not only technical analysis regarding the sources of various aspects of the AI algorithm, but also the legal agreements among the companies involved, including any associated indemnification agreements.'[102]

Accountability through contracts: Supply Chain 4

Scenario: *[Two actors] A company deploys a system developed by another company under contract. This gives the deploying company a high level of control but presents some challenges in their ability to assess and monitor for risks. The contracting/developer company may specify permitted and prohibited uses of the system in the contract, although they may have limited resources to monitor and enforce how the deploying company uses these. The customer/deployer company will be likely to have a high-level understanding of the system, but may not have specialist AI expertise to monitor and mitigate resulting risks.*

Using this example, the use of a contract can allow both parties to agree who is responsible for assessing, mitigating and being held accountable for certain risks. This contract can also create a structure for the flow of information (for example, documentation about the model or datasets used) between the two companies. In this case, the deployer may be best placed to monitor for potential errors or issues, but the developer could be contractually obligated to address those issues once reported back to them. However, it may be the case the developer is larger and more powerful than the deployer; and the deployer may have limited ability to negotiate custom contracts. Regulators must be attentive to these power imbalances.

102 ibid.

At a minimum, those companies will need to use contract law to ensure they have all the data they need about the models and systems they make use of to do so effectively.[103] Japan's government is encouraging this by issuing interpretive guidance on AI contracts.[104] In turn, companies' suppliers will need to ensure they can do the same with all of the components making up the systems they are offering. Similarly, those contracts will need to provide mechanisms by which companies using AI can notify suppliers and request remediation of problems, all the way up the supply chain.

Debate in EU institutions has also highlighted 'the belief that original AI developers will often be larger entities such as tech giants. These larger entities can be assumed to possess more resources and greater knowledge compared to the (arguably smaller) companies that will eventually become the providers, as they will place the high-risk AI systems on the market.'[105]

Upstream suppliers will often be larger and more powerful, and downstream deployers may have limited ability to negotiate custom contracts – as already seen with cloud services. This may leave small and medium-sized enterprises (SMEs) in a weak position to determine important aspects of contracts, which could, for example, determine joint controllership under data protection law (and thus create more liability).[106] Regulators must carefully consider these issues, and may find it beneficial to issue guidance on the use of contracts in AI supply chains.

103 Engler and Renda (n 1) 15.

104 MEIT expert group, 'Governance Guidelines for Implementation of AI Principles Ver. 1.1' (Japan Ministry of Economy, Trade and Industry 2021) 35 https://www.meti.go.jp/english/press/2022/0128_003.html

105 Engler and Renda (n 1) 23.

106 Cobbe and Singh (n 30) 43.

Foundation models

Foundation models are worth considering as a separate element of an AI supply chain, as they can make it harder for regulators to assign responsibilities, and more challenging for sectoral regulators to identify the boundaries of their remit.

> Foundation models, sometimes called 'general purpose AI/GPAI systems', are characterised by their training on especially large datasets to perform many tasks, making them particularly well suited for adaptation to more specific tasks through transfer learning. These models – especially those used for natural language processing, computer vision, speech recognition, simulation, and robotics – have become more foundational in many commercial and academic AI applications.'[107]

OpenAI's chief scientist Ilya Sutskever has commented: 'These models are [...] becoming more and more potent. At some point it will be quite easy, if one wanted, to cause a great deal of harm with those models.'[108]

The supply chains of foundation models are similar to Supply Chain 3 described earlier, but differ in a crucial way – a single model can be adapted (or 'fine-tuned') for a wide variety of applications, which means:

1. It becomes harder for upstream providers of a foundation model to understand how it will be used and to mitigate its risks.
2. A much wider number of sectoral regulators will have to evaluate its use.
3. A single point of failure by the developer (for example, an error in the training data) could create a cascading effect that causes errors for all subsequent downstream users. As European civil society groups

107 Engler and Renda (n 1).

108 James Vincent, 'OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: "We Were Wrong"' The Verge (15 March 2023) https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview accessed 24 March 2023.

> have noted: 'A single GPAI system can be used as the foundation for several hundred applied models (e.g. chatbots, ad generation, decision assistants, spambots, translation, etc.) and any failure present in the foundation will be present in the downstream uses.'[109]

In this section, we discuss some of the active, relevant debates in EU and US policy circles around how to regulate foundation models, and how the regulation of these systems is further complicated by the dynamics of 'open source' models.

Supply chains and market dynamics for foundation models

Some foundation models have been released on a cloud computing platform and made accessible to other developers via an API but (unlike Supply Chain 3 above) with the capability to fine-tune models using their own data. Many end users will also likely experience products built using foundation models, which may be built into existing products and services such as operating systems, web browsers, voice assistants and workplace software (such as Microsoft Office and Google Workspace).

Figure 3 (next page) shows the current market structure of cloud computing, where Amazon and Microsoft (and to a lesser extent Google's parent company, Alphabet) already have large market shares,[110] with substantial investments into machine learning research and development, and global computing and communications infrastructure.

It therefore seems likely that these three companies will also become highly successful in offering foundation models on their platforms. These companies already offer a range of AI/machine learning services to clients, such as Google's AI Infrastructure and Microsoft's Azure AI Platform. They are already able to 'offer their services at lower cost, broader scale, greater technical sophistication, and with potentially easier access for customers than many competitors.' [111]

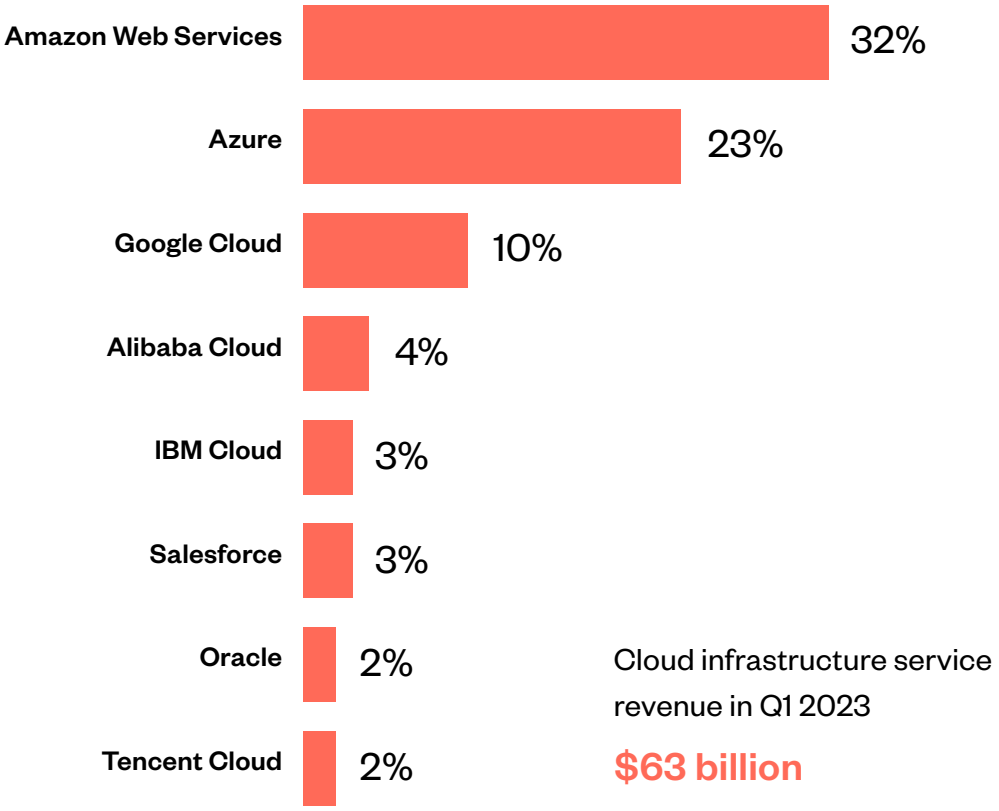
109 Access Now et al. (n 94).

110 Felix Richter, 'Amazon, Microsoft & Google Dominate Cloud Market' (Statista Infographics, 23 December 2022) https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers accessed 21 March 2023.

111 Cobbe, Veale and Singh (n 8) 9.

Figure 3: Amazon, Microsoft and Google's dominance of the global cloud market[112]

Worldwide market share of leading cloud infrastructure service providers in Q1 2023*



* includes platform as a service (PaaS) and infrastructure as a service (IaaS) as well as hosted cloud services.

Source: Synergy Research Research Group/Statista

112 Source: Richter, (2022), Amazon, Microsoft & Google Dominate Cloud Market, https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers Accessed: 21 March 2023. Adapted with permission of author.

A product briefing leaked from OpenAI has described the platform Foundry, with tiers of pricing based on the computational load and model sophistication, starting from hundreds to millions of dollars per year.[113] Specialised microchip manufacturer NVIDIA has announced a similar platform.[114]

However, scholars have noted that 'the fact that AIaaS operates at scale as an infrastructure service does offer potential points of legal and regulatory intervention. Given AI services will likely be widely used in future, then regulating at this infrastructural level could potentially be an effective way to address some of the potential problems with the growing use of AI'.[115] This would mean focusing regulatory attention on the large providers of these foundational models.

A contrary view has been provided in an internal note allegedly leaked from Google, which assesses the gap between proprietary and open source AI models (discussed further below) is 'closing astonishingly quickly'.

The note describes how Meta released the model weights for its large language model LLaMa in March 2023, which caused open source developers to quickly recreate the model and build novel applications from it. The note concludes: 'The barrier to entry for training and experimentation has dropped from the total output of a major research organization to one person, an evening, and a beefy laptop.'[116]

How EU regulators are assigning responsibility to foundation models

Other jurisdictions (notably the various EU institutions developing the AI Act) are planning to go further than the UK and place (non-contractual) regulatory requirements on suppliers higher up the AI supply chain,

113 Erik Torenberg and Nathan Labenz, 'OpenAI's Foundry Leaked Pricing Says a Lot – If You Know How to Read It' (The Cognitive Revolution, 27 February 2023) https://cognitiverevolution.substack.com/p/78a8bc84-59ab-47d2-bcdc-dfda00131549 accessed 28 February 2023.

114 Andrew Tarantola, 'NVIDIA Unveils AI Foundations, Its Customizable Gen-AI Cloud Service' Engadget (21 March 2023) https://www.engadget.com/nvidia-ai-foundations-customizable-genewrative-ai-cloud-service-161505625.html accessed 22 March 2023.

115 Cobbe and Singh (n 30) 52.

116 Patel and Ahmad (n 27).

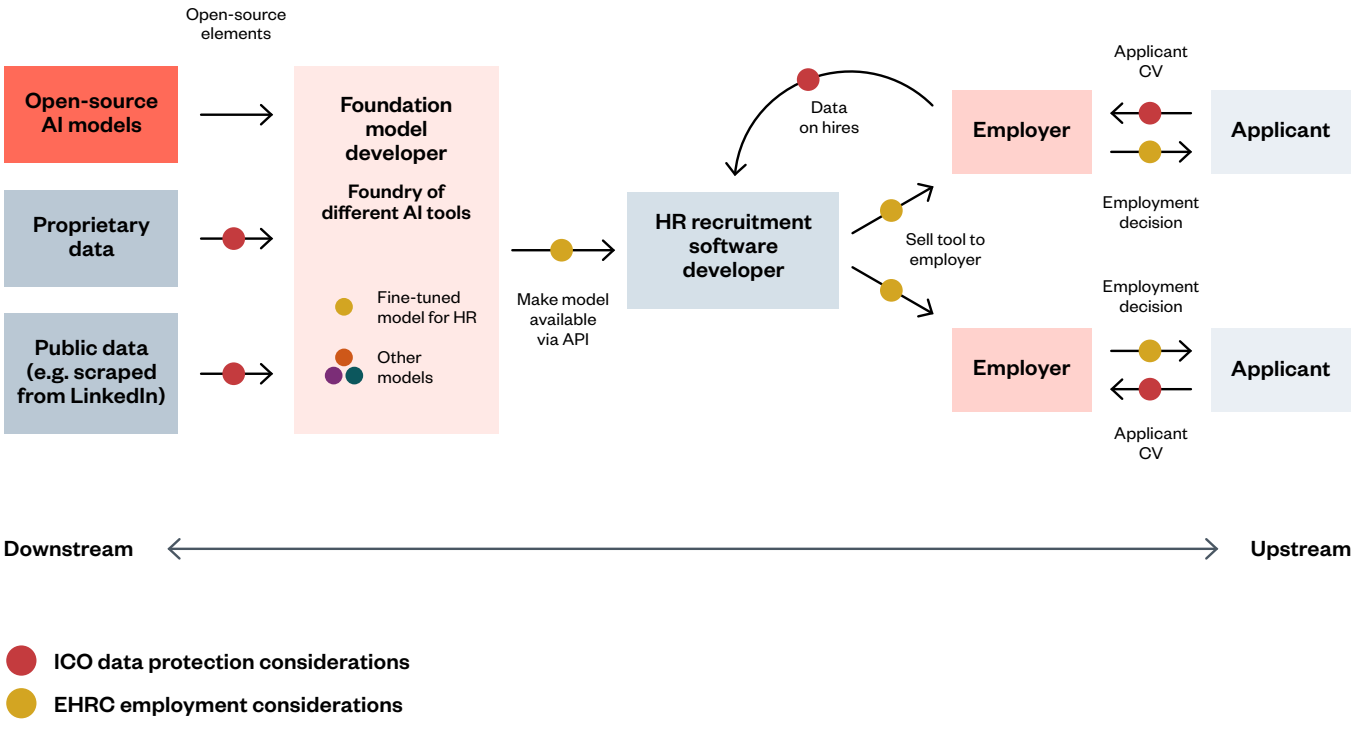
including for foundation models and systems (especially where those are assessed as high risk).

These EU regulatory requirements could include transparency mechanisms around the data and model architecture of the model. This would enable academics, civil society groups and the media to more effectively scrutinise those systems for public-interest concerns such as fairness and non-discrimination. Regulators could also set baseline requirements for the information downstream developers building on foundation models must acquire from upstream developers of the system.

Considerations for assigning responsibility for foundation models

Below, we include an example supply chain of a foundation model that has been fine-tuned to provide an HR recruitment service. Figure 4 shows how different sectoral regulators may need to intervene at different points.

Figure 4: AI supply chain for HR recruitment built from a foundation model



Description of Figure 4: A foundation model system developer has built a 'foundry' of AI tools using proprietary data, public and scraped data, and open-source AI models. An HR recruitment software developer has fine-tuned an HR-specific version of the foundation model via an API, using their own proprietary data from their clients, and created an HR recruitment online service. That service is then procured by downstream employers and recruitment agencies, which use it to contribute to decisions about potential job applicants. Employers and agencies must have regard to the EHRC's Statutory Code of Practice on Employment.

The foundation model provider updates the fine-tuned model's data sheet and model card regularly, which the HR tool developer and its customers can retrieve on demand to show the Equality and Human Rights Commission (EHRC) they are not discriminating between candidates based on protected characteristics or proxies for them. These organisations can also share the data sheet and model card with the Information Commissioner's Office (ICO) if there are questions about the fairness of processing, or other data protection law compliance issues.

Drawing on our framework and the principles of efficacy and transparency, it may be more efficient to deal with risks such as bias in suppliers that are higher upstream in supply chains, if their models/systems are being used by large numbers of downstream deployers and developers. Otherwise, 'excluding [GPAI] models could potentially distort market incentives, leading companies to build and sell GPAI models that minimise their exposure to regulatory obligations, leaving these responsibilities to downstream applications'.[117]

In the HR supply chain example (Figure 4) above, this would mean placing requirements to evaluate for issues of bias and performance on the foundation model provider, as only they would have the access and proximity to assess for bias in that model.

That information could be made available for downstream developers via a model card. Similar obligations could be placed on the downstream HR recruitment service tool developer as they further refine the tool for use by specific employers.

117 Engler and Renda (n 1) 23.

There are concerns that SMEs building systems on top of foundation models will not have the resources to address many risks. This will present problems because 'shifting responsibility to these lower-resourced organizations [...] simultaneously exculpates the actors best placed to mitigate the risks of general purpose systems, and burdens smaller organizations with important duties they lack the resources to fulfil'.[118]

Locating responsibility with foundation model developers higher up the supply chain would enable them to 'control several levers that might partially prevent malicious use of their AI models. This includes interventions with the input data, the model architecture, review of model outputs, monitoring users during deployment, and post-hoc detection of generated content.'

But it will not create a perfect system, rather: 'the efficacy of these efforts should be considered more like content moderation, where even the best systems only prevent some proportion of banned content.'[119] Scholars suggest mechanisms that are already familiar from the EU Digital Services Act and UK Online Safety Bill: 'notice and action mechanisms, trusted flaggers, and, for very large [generative AI model] developers, comprehensive risk management systems and audits concerning content regulation'[120]

The US Federal Trade Commission has announced a potentially far-reaching approach under its consumer protection authority, warning businesses creating generative AI systems they should 'consider at the design stage and thereafter the reasonably foreseeable – and often obvious – ways it could be misused for fraud or cause other harm. Then ask yourself whether such risks are high enough that you shouldn't offer the product at all.'[121]

118 Kolt (n 67) 33.

119 Alex Engler, 'Early Thoughts on Regulating Generative AI like ChatGPT' (16 February 2023) https://www.brookings.edu/blog/techtank/2023/02/16/early-thoughts-on-regulating-generative-ai-like-chatgpt/ accessed 21 February 2023.

120 Philipp Hacker, Andreas Engel and Marco Mauer, 'Regulating ChatGPT and Other Large Generative AI Models', Proceedings of Fairness, Accountability and Transparency '23 (ACM 2023) 22 http://arxiv.org/abs/2302.02337 accessed 16 May 2023.

121 Michael Atleson, 'Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale' (Federal Trade Commission Business Blog, 20 March 2023) https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale accessed 22 March 2023.

It goes on to say that companies should take 'all reasonable precautions before [such a system] hits the market', and adds: 'Merely warning your customers about misuse or telling them to make disclosures is hardly sufficient to deter bad actors. Your deterrence measures should be durable, built-in features and not bug corrections or optional features that third parties can undermine via modification or removal. If your tool is intended to help people, also ask yourself whether it really needs to emulate humans or can be just as effective looking, talking, speaking, or acting like a bot.'

However, as AI software and models become more generalisable and have potentially more users, it becomes harder for their developers to consider customer-specific contexts and potential harms.

As scholars have pointed out, 'AI practitioners encounter difficulty in engaging with downstream marginalized groups in large scale deployments. Even where a company is working directly with a client to develop a system for them, it may be unable to know what the customer later did with that system after the initial prototype phase, as follow up work does not scale'.[122] Some responsibilities for foundation model supply chains must be placed on deployers who are using the system in a specific context.

Other scholars suggest that systems such as ChatGPT are so general-purpose and usable in so many contexts they should be regulated as a specific category. This would place a duty on developers to actively monitor and reduce risks, in a similar manner to the obligations on platforms of the EU Digital Services Act (Article 34) and the UK Online Safety Bill.[123]

Scholars also suggest regulators should monitor the 'fairness, quality and adequacy of contractual terms and instructions' between providers and end-users, as is also considered for platforms under the Online Safety Bill.[124] Researchers suggest a specific category of regulation, which

122 Widder and Nafus (n4).

123 Michelle Donelan and Lord Parkinson of Whitley Bay, Online Safety Bill 2023. European Parliament and Council of the European Union, 'Digital Services Act' art 34. The EU's AI Act is moving in this direction as it is negotiated.

124 Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review https://policyreview.info/essay/chatgpt-and-ai-act accessed 22 February 2023."plainCitation":"Natali Helberger and Nicholas Diakopoulos, 'ChatGPT and the AI Act' (2023

imposes limited transparency obligations on generative AI developers, but imposes the duty to implement a risk-management system on companies using such a system in high-risk applications.[125]

In the EU approach, so-called 'providers' (or developers) of high-risk AI systems would face 'provisions on a risk management system, data governance, technical documentation, record keeping (when possible), transparency requirements, accuracy and robustness [...] and go through the formal legal registration process, including performing an ex ante conformity assessment procedure, registering the high-risk AI system in the EU-wide database, establishing an authorised representative as a point of contact for regulators, and demonstrating conformity upon the request of regulatory agencies'.[126]

In 2022, the Council of the EU proposed that GPAI models are addressed as a stand-alone category of AI system. They propose the exact obligations to be placed on GPAI developers are decided via an 'implementing act' (a piece of secondary legislation that allows the European Commission to take 18 months to address this question.)

In May 2023, the European Parliament followed suit by also proposing tailored requirements for GPAI[127], 'foundation models'[128] and 'generative AI'.[129] They conceptualise foundation models and generative AI as sub-categories of GPAI, and set different rules for each:

- GPAI providers will be required to share information downstream in order to support downstream providers (e.g. fine-tuners) to comply, if deploying the GPAI in a high-risk area.

- Foundation model providers will have to obligations at the design and development phase, and throughout the lifecycle. The requirements focus on risk and quality management, data governance measures, and testing the model for predictability, interpretability, corrigibility,

125 Philipp Hacker, Andreas Engel and Theresa List, 'Understanding and Regulating ChatGPT, and Other Large Generative AI Models: With input from ChatGPT' (Verfassungsblog, 20 January 2023) https://verfassungsblog.de/chatgpt/ accessed 20 January 2023.

126 Engler and Renda (n 1) 4–5.

127 "an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed"

128 "an AI model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks"

129 defined as "foundation models specifically intended to be used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video"

safety and cybersecurity. These rules are aimed to be 'broadly applicable', i.e. independent of distribution channels, modality, or development method.

- Finally, generative AI providers will be compelled to follow transparency obligations to make clear to end users that they are interacting with an AI model, and will also have to document and make publicly available a summary of the use of training data protected under copyright law.

The Council of the EU and the European Parliament will therefore regulate GPAI – including foundation models and generative AI – in some form, but the exact requirements will be dependent on the inter-institutional 'trilogue' negotiations, which will conclude by the end of 2023 or early 2024.

Large technology companies have argued strenuously against regulation of foundation models. Google has told the European Commission such 'systems [...] are not themselves high-risk', and Microsoft has argued that regulation would have a negative impact on startups and SMEs (although the European Digital SME Alliance disagrees.[130]

However, when asked whether 'it's too early for policymakers and regulators to get involved,' OpenAI's Chief Technology Officer, Mira Murati, responded: 'It's not too early. It's very important for everyone to start getting involved, given the impact these technologies are going to have.'[131]

OpenAI's CEO Sam Altman told a US Senate hearing that 'the U.S. government should consider a combination of licensing or registration requirements for development and release of AI models above a crucial threshold of capabilities, alongside incentives for full compliance with these requirements.'[132]

130 Natasha Lomas, 'Report Details How Big Tech Is Leaning on EU Not to Regulate General Purpose AIs' TechCrunch (23 February 2023) https://techcrunch.com/2023/02/23/eu-ai-act-lobbying-report/ accessed 1 March 2023.

131 Steve Mollman, 'ChatGPT Must Be Regulated and A.I. "Can Be Used by Bad Actors," Warns OpenAI's CTO' Fortune (5 February 2023) https://fortune.com/2023/02/05/artificial-intelligence-must-be-regulated-chatgpt-openai-cto-mira-murati/ accessed 21 March 2023.

132 'Oversight of A.I.: Rules for Artificial Intelligence' https://www.judiciary.senate.gov/download/2023-05-16-testimony-altman accessed 21 May 2023.

Altman and OpenAI have not specified what that threshold would be, or whether any of the commercial products they are building would be underneath that threshold. In a paper co-authored by Altman, OpenAI has also referred to models above this threshold as 'frontier models' and 'superintelligence,' which is distinct from 'artificial general intelligence.'[133]

These terms are undefined and create ambiguity around what the object of regulation should be. Altman has also noted that 'regulatory capture is bad, and we shouldn't mess with models below the threshold. Open source models and small startups are obviously important'. [134]

To conclude, foundation models raise additional challenges around the assignment of responsibilities in an AI supply chain. Transparency will be necessary but not sufficient – access to information about the model's data, architecture and weights (which determine how complex models interpret data) will be essential for downstream providers to remedy any issues they spot and meet regulatory obligations.

Transparency measures are a key instrument for reducing these challenges, but the general capabilities of these technologies will also require regulatory approaches that focus on all stages of their supply chain and approach these with the principle of efficacy in mind. For example, it may be that upstream providers are best placed to address data quality and evaluation issues, given the serious risk of cascading errors these systems could cause.

These matters are further complicated by open-source models, which have community-centred benefits but come with a trade-off concerning access and auditability of a system and a loss of constraints on its uses (see section on 'The challenges of open-source' below).

AI system release strategies

One of the biggest factors affecting an AI component's supply chain

133 Sam Altman, Greg Brockman and Ilya Sutskever, 'Governance of Superintelligence' (OpenAI, 22 May 2023) https://openai.com/blog/governance-of-superintelligence

134 Sam Altman https://twitter.com/sama/status/1659341540580261888?s=20.

and how subsequent responsibilities are assigned is how it is released. In some cases, AI components will be released in ways that make downstream developers or deployers incapable of accessing or understanding critical details of how they are trained. In the case of foundation models, how a model is released will have significant impacts on how responsibilities for addressing misuse should be applied.

Researchers have summarised various trade-offs for the degree of openness with which developers of 'generative' AI models (those that create new content) make them available to third-parties (shown below in Figure 5). More openness can bring benefits, as it increases the ability of a wider range of organisations and experts to audit models, increases the transparency of how models work and brings a broader range of perspectives to bear.

At the most open end of the spectrum (on the right side of Figure 5), models released under open-source licences (alongside resources such as training datasets and software) can be developed by communities of developers. This 'fully open' release allows the full details of the model to be made available, which maximises transparency and the opportunity for third-party assessment and development.[135]

Figure 5: Considerations for different kinds of AI system access[136]

Considerations	Internal research only High-risk control Low auditability Limited perspectives		Gated to public			Community research Low-risk control High auditability Broader perspectives
Level of access	**Fully closed**	**Gradual / staged release**	**Hosted access**	**Cloud-based / API access**	**Downloadable**	**Fully open**
System (Developer)	PaLM (Google) Gopher (DeepMind) Imagen (Google) Make-A-Video (Meta)	GPT-2 (Open AI) Stable Diffusion (Stability AI)	DALL-E 2 (Open AI) Midjourney (Midjourney) ChatGPT	GPT-3 (OpenAI) GPT-4 (OpenAI)	OPT (Meta) Craiyon (Craiyon)	BLOOM (BigScience) GPT-J (EleutherAI)

135 Irene Solaiman, 'The Gradient of Generative AI Release: Methods and Considerations', Proceedings of Fairness, Accountability and Transparency '23 (ACM 2023) http://arxiv.org/abs/2302.04844 accessed 25 February 2023.

136 Source: Irene Solaiman (2023) 'The Gradient of Generative AI Release: Methods and Considerations', FAccT '23, doi: 10.48550/arXiv.2302.04844

However, this openness comes with a significant trade-off: reducing the technical ability of developers to constrain their systems' use or misuse. Developers can still implement legal constraints via licences like Responsible AI Licenses (RAIL) that contractually prohibit the use of the model in a certain way, but it remains unclear how viable this method is as a remedy for preventing misuse.[137]

Fully open-source software does not generally impose such limits on deployers, and researchers have noted: 'Open source licensing invokes ideological frames that reject the idea that developers should exercise any control at all over harmful use: "the whole point is you can't control that – can't control what people do."'[138]

At the left end of the scale in Figure 5, models are kept entirely in-house. This gives the developer the highest level of control over usage, but provides limited ability for third parties to audit the model or provide broader perspectives on its use. A slightly broader version of this is to provide external researchers with access to some or all the model details and data needed to assess it. Some researchers propose doing this under the auspices of a foundation model review board.[139]

The next stage on the scale is to provide progressive releases of more sophisticated versions of a model, ideally limiting the opportunity for misuse while enabling broader assessment. Further stages give increasingly flexible access to the public, either hosted online (with the capability to constrain the output of the model's responses, as ChatGPT does) or enabling downloading of software containing the model (which gives a limited ability to constrain its use).

This is not necessarily a fail-safe: researchers have shown the potential to bypass controls even on model access provided via a constrained API, and as an example did so to use ChatGPT to create personalised scam e-mail.[140]

137 Danish Contractor and others, 'Behavioral Use Licensing for Responsible AI', 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM 2022) https://dl.acm.org/doi/10.1145/3531146.3533143 accessed 24 March 2023.

138 Widder and Nafus (n 4).

139 Liang and others (n 79).

140 Daniel Kang and others, 'Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks' (11 February 2023) http://arxiv.org/abs/2302.05733 accessed 7 March 2023.

Gradual or staged releases and hosted access releases can enable developers to implement more controls over misuse. Cohere, OpenAI and AI21 Lab have voluntarily adopted a preliminary set of best practices for deploying large language models (LLMs). These include: 'prohibit misuse' through guidelines and terms of use, and technical usage restrictions (such as rate limits); 'mitigate unintentional harm' via model evaluation and documenting known vulnerabilities; and 'stakeholder collaboration', including diverse teams and consultation, public disclosure of learning and respectful treatment of all workers in supply chains.[141]

However, researchers warn that 'best practices and other deployment-focused frameworks primarily target the immediate risks from current AI systems. Far less attention is directed toward longer-term and larger-scale societal risks.'[142]

Applying our framework above, the principles of efficacy and transparency are critical. If a model is released in a more closed manner, it makes it harder for deployers or downstream users in the supply chain to identify these risks.

The further to the left on this spectrum, the more control a developer has on how a model is designed and used, and therefore the greater the responsibility they should have. The principle of transparency is also critical here, as developers will have far more information than a deployer about the model's architecture. Without transparency mechanisms in place, it will be hard for downstream deployers to identify or mitigate risks.

141 Cohere, OpenAI and AI21 Labs, 'Best Practices for Deploying Language Models' (Context by Cohere, 2 June 2022) https://txt.cohere.ai/best-practices-for-deploying-language-models/ accessed 21 March 2023.

142 Kolt (n 67) 20.

Examples of risks from generative models

Generative models like MidJourney or ChatGPT enable the generation of complex text or images through simple text prompts. They are one example of an AI system that can pose particular risks if made widely accessible, since their text/audio/video outputs are becoming increasingly difficult to distinguish from authentic human writing, speech and (in the foreseeable future) video footage of real events. There are significant concerns around their use in fraud, disinformation, 'deep fakes' (such as false representations of individuals in sexual imagery and politicians making false statements) and the generation of hate speech.[143] In one reported case, a chatbot trained to be 'more emotional, fun and engaging' appears to have contributed to the suicide of a Belgian man.[144]

While there has been some research into 'watermarking' an output of a generative model so that it can later be automatically detected,[145] the history of digital watermarking as a mechanism to limit copyright infringement suggests it will be difficult to make this robust (especially where it is possible to generate multiple versions of the same output, and the tools are publicly available to identify such watermarks).[146]

A preliminary analysis suggests tools to detect watermarks or other 'signatures' of generative model output 'are not reliable in practical scenarios'.[147] The US Federal Trade Commission has warned businesses: 'Researchers continue to improve on detection methods for AI-generated videos, images, and audio. Recognizing AI-generated text is more difficult. But these researchers are in an arms race with companies developing the generative AI tools, and the fraudsters using these tools will often have moved on by the time someone detects their fake content. The burden shouldn't be on consumers, anyway, to figure out if a generative AI tool is being used to scam them.'[148]

143 Laura Weidinger and others, 'Taxonomy of Risks Posed by Language Models', 2022 ACM Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery 2022) https://dl.acm.org/doi/10.1145/3531146.3533088 accessed 21 May 2023.

144 Chloe Xiang, '"He Would Still Be Here": Man Dies by Suicide After Talking with AI Chatbot, Widow Says' (Vice, 30 March 2023) https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says accessed 21 May 2023.

145 John Kirchenbauer and others, 'A Watermark for Large Language Models' (arXiv, 27 January 2023) http://arxiv.org/abs/2301.10226 accessed 21 March 2023.

146 Ed Felten, 'How Watermarks Fail' (Freedom to Tinker, 24 February 2006) https://freedom-to-tinker.com/2006/02/24/how-watermarks-fail/ accessed 7 March 2023."plainCitation":"Ed Felten, 'How Watermarks Fail' (Freedom to Tinker, 24 February 2006

147 Vinu Sankar Sadasivan and others, 'Can AI-Generated Text Be Reliably Detected?' (arXiv, 17 March 2023) http://arxiv.org/abs/2303.11156 accessed 24 March 2023.

148 Atleson (n 125).

An extreme example of a potential risk of generative AI is the use of drug-design models to generate novel chemical weapons. A pharmaceutical research company found it was possible to generate the structure of 40,000 highly toxic molecules in under six hours, using a model trained on a public database and software based on open-source tools.[149]

The researchers decided it would be unethical to carry out any further analysis on the lethality of the molecules. Other scientists have commented: 'The development of actual weapons in past weapons programs have shown, time and again, that what seems possible theoretically may not be possible in practice.'[150] Given the potential risks of these models, developers may be inclined to use more restricted forms of model releases.

The challenges of open-source

To the far right side of the scale in Figure 5, models are released via open-source licences and made fully available for other researchers and companies to use. Because so much more information is made available when models are released in this way, it provides researchers with a 'better means to validate provided information and test the capabilities of foundation models.'[151]

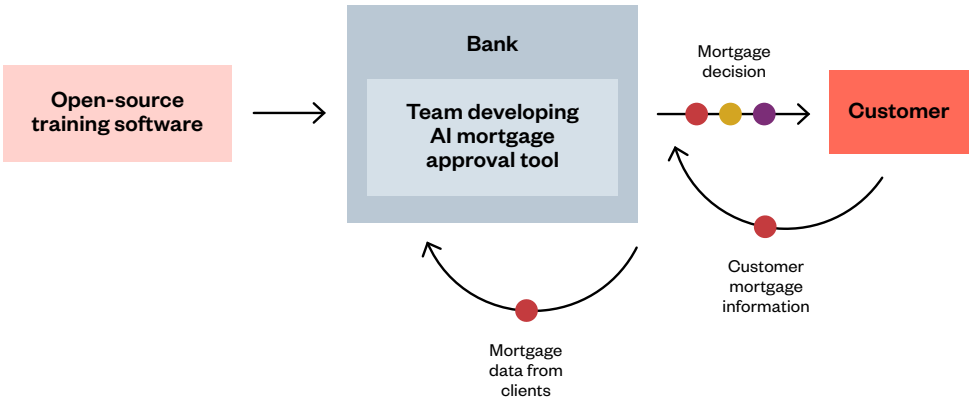
When developers use open-source training software and their own data to create a model, they are similarly in a much better position to test and update it. An example supply chain containing open-source software is shown below in Figure 6.

149 Fabio Urbina and others, 'Dual Use of Artificial-Intelligence-Powered Drug Discovery' [2022] Nature Machine Intelligence 189.

150 Rebecca Sohn, 'AI Drug Discovery Systems Might Be Repurposed to Make Chemical Weapons, Researchers Warn' Scientific American (21 April 2022) https://www.scientificamerican.com/article/ai-drug-discovery-systems-might-be-repurposed-to-make-chemical-weapons-researchers-warn/ accessed 7 March 2023.

151 'Artificial Intelligence Act: How the EU Can Take on the Challenge Posed by General-Purpose AI Systems' (Mozilla 2023) https://openfuture.eu/wp-content/uploads/2023/01/AI-Act_Mozilla-GPAI-Brief_Kx1ktuk.pdf accessed 23 February 2023.

Figure 6: AI supply chain for bank mortgage approval



Description of Figure 6: A bank has a development team which has used open source training software and internal customer data to produce a mortgage approval model. This is used to assess customer applications, with application data as well as mortgage performance data used to fine-tune the model. The Financial Conduct Authority (FCA) and Bank of England oversee the bank's use of this system, including whether it is meeting the bank's equality obligations. The ICO will also have oversight of uses of personal data by the bank.

For this reason, open-source developers such as Mozilla have argued that 'GPAI released open source and not as a commercial service should therefore be excluded from [regulation] if the information necessary for compliance is made available to downstream actors.'[152]

Other researchers suggest that this should include model cards and datasheets that contain essential information about the model's architecture and the data used to train it, access to training data, and 'licensing conditions that guarantee the right of licensees and third parties to audit and explain the behaviours of models'.[153]

152 ibid.

153 Paul Keller, 'How Will the AI Act Deal with Open Source AI Systems?' (13 December 2022) https://openfuture.eu/blog/how-will-the-ai-act-deal-with-open-source-ai-systems accessed 23 February 2023.

Open-source foundation model projects play two key roles:

1. 'They disseminate power over the direction of AI away from well-resourced technology companies to a more diverse group of stakeholders.

2. 'They enable critical research, and thus public knowledge, on the function and limitations of GPAI models."[154]

Support for fully open-source AI systems (including training data and software as well as models) is therefore one way to deal with the tendencies towards market concentration identified earlier in this paper.

While it may seem in the financial interest of companies investing heavily in the development of proprietary models to control their availability, even the largest technology companies are also contributing to open-source systems. For example, Microsoft has contributed to research leading to improvements in the Stable Diffusion image generation system.[155]

However, it is likely such contributions will be in the interests of the companies concerned.[156] An expert review for the European Commission found that platforms generally shape innovation within their own ecosystems to bolster their business models.[157]

It is not yet clear whether the very high resource requirements of creating the highest-capability models (such as OpenAI's GPT-4 and Google's LaMDA) will mean regulating their safety and availability via those companies will be feasible (as called for by OpenAI's CEO[158] and others).

154 Alex Engler, 'The EU's Attempt to Regulate Open-Source AI Is Counterproductive' (Brookings Institution TechTank, 24 August 2022) https://www.brookings.edu/blog/techtank/2022/08/24/the-eus-attempt-to-regulate-open-source-ai-is-counterproductive/

155 Yuheng Li and others, 'GLIGEN: Open-Set Grounded Text-to-Image Generation' (17 April 2023) http://arxiv.org/abs/2301.07093 accessed 6 March 2023.

156 Meredith Whittaker, 'The Steep Cost of Capture' (2021) 28 Interactions 50.

157 Ariel Ezrachi and Maurice E Stucke, 'Digitalisation and Its Impact on Innovation' (European Commission DG Research and Innovation 2020) 978-92-76-17462–2, KI-BD-20-003-EN-N https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/digitalisation-and-its-impact-innovation_en accessed 21 March 2023.

158 'Oversight of A.I.: Rules for Artificial Intelligence' (n 132).

While open source generative language models have been advancing at a rapid pace, so far they have been significantly based on models from companies such as Meta, whose model LLaMA was leaked in March 2023.[159] The AI Now Institute suggests: 'Even if costs are lower or come down as these systems are deployed at scale (and this is a hotly contested claim), Big Tech is likely to retain a first mover advantage.'[160]

Researchers note that despite the public availability for some time of capable generative AI systems (such as GPT-J), we are yet to see documented cases of resulting malicious use, suggesting that other obstacles to such use are still present. They argue providers should 'release audits of how the tools have been used and abused', and that 'Social media platforms should study and report the prevalence of [Large Language Model]-generated misinformation.'[161] Platforms could be encouraged to do so as part of the risk-management approach mandated by the UK Online Safety Bill.

While it would be possible for legislation to go further in applying obligations to online distribution of open-source AI components, its likely efficacy would be severely open to question, given the following observations:

- Without comprehensive international agreement (which is difficult to imagine in the current geopolitical climate), unrestricted development and sharing would be likely to continue in other jurisdictions (including the USA, whose constitution includes strict restrictions on government limits on publication).[162]

- The underlying techniques and data used for training models are likely to continue circulating freely (open-source software and a public molecule database were used to train the model used to identify potential chemical weapons described on pages 55/56.).

159 Patel and Ahmad (n 27).

160 Amba Kak and Sarah Myers West, 'AI Now 2023 Landscape: Confronting Tech Power' (AI Now Institute 2023) 17 https://ainowinstitute.org/2023-landscape accessed 21 May 2023.

161 Arvind Narayanan and Sayash Kapoor, 'The LLaMA Is out of the Bag. Should We Expect a Tidal Wave of Disinformation?' (Algorithmic Amplification and Society, 6 March 2023) http://knightcolumbia.org/blog/the-llama-is-out-of-the-bag-should-we-expect-a-tidal-wave-of-disinformation accessed 21 March 2023.

162 Andrea Matwyshyn, 'Hacking Speech: Informational Speech and the First Amendment' (2013) 107 Northwestern University Law Review 795.

- Such restrictions would be likely to significantly impede the pace of research and development relating to AI tools and techniques, including those to identify and remedy potential harms, particularly outside of the large companies which already and increasingly dominate AI research.[163]

While not a precise analogy (because large AI models are much more complex and resource-intensive to create than encryption software), attempts by the USA and its allies to control the global spread of encryption technology throughout the 1980s and 1990s ultimately failed for similar reasons.[164]

163 Ahmed, Wahed and Thompson (n 4).

164 Whit Diffie and Susan Landau, Privacy on the Line (Updated and Expanded Edition, Random House 2010) https://www.penguinrandomhouse.com/books/654750/privacy-on-the-line-updated-and-expanded-edition-by-whitfield-diffie-and-susan-landau/ accessed 12 March 2023.

Conclusion

The UK's approach to AI regulation focuses on companies offering services incorporating AI functionality to customers. These companies are in the best position to assess and mitigate the context-specific risks and potential harms of systems they offer to end-users, although the UK's principles 'will ultimately apply to any actor in the AI lifecycle whose activities create risk that the regulators consider should be managed through the context-based operationalisation of each of the principles'.[165]

Any approach to AI regulation will need to grapple with different supply chains behind those services and with assigning responsibilities to actors in those supply chains. Broadly speaking, policymakers and regulators will need to understand 'in terms of who is doing what for whom, who is performing what key functions for others, who is core to certain supply chains, and who is systemically important.'[166]

Transparency mechanisms such as model cards and datasheets are an essential component of supply chain accountability, but can come into tension with other incentives, such as trade secrecy. OpenAI's recent release of GPT-4 and Google's recent release of Bard saw both companies refuse to provide details on the models' architecture and data sources, citing reasons of competition and safety.[167]

The refusal by companies to make these details accessible should alarm regulators and policymakers, as it removes the ability of downstream users and third-party auditors to assess the safety, performance and ethical considerations of these models. These transparency mechanisms should be standardised by governments and regulators, ideally via international standards and requirements, and made a legal requirement from companies putting AI models and services on the UK market.

165 Department for Digital, Culture, Media and Sport, 'Establishing a Pro-Innovation Approach to Regulating AI' https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement accessed 19 January 2023.

166 Cobbe, Veale and Singh (n 8) 12.

167 Vincent (n 108).

Where AI components are used by many downstream companies in a supply chain, it will be more efficient for some issues to be fixed by the component developer. Allocation of responsibility must also account for the power imbalances between different actors and how AI systems are released.

Those developing an AI system may be in a greater position of power over their suppliers or users to contractually offload responsibilities. Depending on how an AI system is released, upstream providers may need to bear more responsibility to evaluate and address the potential issues within their system.

Foundation models complicate supply chain considerations. Determining what kinds of responsibilities should apply will require both ex ante assessments of risk and assignments of responsibility by regulators and policymakers, along with ex post regulation of the actual uses of these systems.

As with other digital markets such as search, social networking services and especially cloud computing, competition concerns are likely to arise in the provision of AI services, due to high returns to scale and the importance of access to specific data resources. In forthcoming legislative reforms, the UK should ensure its Competition and Markets Authority Digital Markets Unit has powers to set ex ante rules where needed to deal with such concerns.

Open-source technologies further complicate supply chain considerations. Regulation must address how AI technologies (and powerful components of those AI technologies, like underlying models, datasets or model weights) are released.

But there are strong practical benefits for innovation, public accountability and competition from the availability of open-source tools. Limits on publication of components to manage risks face significant constraints, not least the small probability of the international agreement which would be needed to make them remotely effective, and the freedom of expression implications of trying to limit access to the underlying knowledge.

Further questions

- How should specific kinds of AI accountability practices be mapped to the supply chains for foundation models?
- What kinds of business models (and supply chains) are emerging for foundation models? How should specific regulatory responsibilities be applied?
- What are the externalities of different AI system release methods?
- Will the capability gap between proprietary and open-source general-purpose AI systems narrow or widen?
- What lessons can regulators and policymakers learn from other kinds of supply chains? What governance measures work best in different contexts?
- What kinds of access requirements will regulators and auditors of AI systems need with different actors in an AI supply chain?

Methodology

This report draws on a review of relevant literature (including preprints, reflecting how fast the field is moving) relating to AI supply chains, risk monitoring and regulation of supply chains in other sectors. Relevant literature was identified through keyword searching of online databases of academic literature and through snowball sampling via conversations with experts in AI supply chains and risk management.

Partner information and acknowledgements

This report was authored by Ian Brown, with substantive contributions from the Ada Lovelace Institute's Elliot Jones and Andrew Strait.

This work was originally undertaken with support via UKRI, the BRAID programme at the University of Edinburgh, and the Department for Digital, Culture, Media & Sport (DCMS) Science and Analysis R&D Programme. It was developed and produced according to UKRI's initial hypotheses and output requests. Any primary research, subsequent findings or recommendations do not represent DCMS views or policy and are produced according to academic ethics, quality assurance and independence.

The author would like to thank Luca Belli and Centre for Technology and Society colleagues for their input, and Reuben Binns, Connor Dunlop, Hamed Haddadi, Natali Helberger, Jat Singh, and Chris Marsden for their helpful comments on drafts.

About the Author

Dr Ian Brown is an internet regulation consultant, a visiting professor at the Centre for Technology and Society at Fundação Getulio Vargas Law School in Rio de Janeiro, and an ACM Distinguished Scientist. He was previously Principal Scientific Officer at the UK Government's Department for Digital, Culture, Media and Sport; Professor of Information Security and Privacy at the University of Oxford's Internet Institute; and a Knowledge Exchange Fellow with the Commonwealth Secretariat and UK National Crime Agency. His books include *Cybersecurity for Elections* (2020 with Marsden, Lee & Veale), *Regulating Code* (2013), and *Research Handbook on Governance of the Internet* (2013).

About the Ada Lovelace Institute

The Ada Lovelace Institute was established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

The mission of the Ada Lovelace Institute is to ensure that data and AI work for people and society. We believe that a world where data and AI work for people and society is a world in which the opportunities, benefits and privileges generated by data and AI are justly and equitably distributed and experienced.

We recognise the power asymmetries that exist in ethical and legal debates around the development of data-driven technologies, and will represent people in those conversations. We focus not on the types of technologies we want to build, but on the types of societies we want to build.

Through research, policy and practice, we aim to ensure that the transformative power of data and AI is used and harnessed in ways that maximise social wellbeing and put technology at the service of humanity.

We are funded by the Nuffield Foundation, an independent charitable trust with a mission to advance social wellbeing. The Foundation funds research that informs social policy, primarily in education, welfare and justice. It also provides opportunities for young people to develop skills and confidence in STEM and research. In addition to the Ada Lovelace Institute, the Foundation is also the founder and co-funder of the Nuffield Council on Bioethics and the Nuffield Family Justice Observatory.

Find out more:

Website: Adalovelaceinstitute.org
Twitter: @AdaLovelaceInst
Email: hello@adalovelaceinstitute.org

Preferred citation: Ian Brown. *Allocating accountability in AI supply chains: a UK-centred regulatory perspective* (Ada Lovelace Institute 2023)
Available at:
https://www.adalovelaceinstitute.org/resource/ai-supplychains/

ISBN: 978-1-7392615-2-8