HARVARD Kennedy School
MOSSAVAR-RAHMANI CENTER
for Business and Government

Challenges in assessing the impacts of regulation of Artificial Intelligence

Stephen Gibson
Harvard Kennedy School

Winston Tang
Harvard Kennedy School

July 2025

M-RCBG Associate Working Paper Series | No. 262

The views expressed in the M-RCBG Associate Working Paper Series are those of the author(s) and do not necessarily reflect those of the Mossavar-Rahmani Center for Business & Government or of Harvard University. The papers in this series have not undergone formal review and approval; they are presented to elicit feedback and to encourage debate on important public policy challenges. This paper is copyrighted by the author(s). It cannot be reproduced or reused without permission. Pursuant to M-RCBG's Open Access Policy, this paper is available to the public at hks.harvard.edu/centers/mrcbg free of charge. Papers may be downloaded for personal use only.

Mossavar-Rahmani Center for Business & Government
Weil Hall | Harvard Kennedy School | www.hks.harvard.edu/mrcbg

Challenges in assessing the impacts of regulation of Artificial Intelligence

Stephen Gibson and Winston Tang

Challenges in assessing the impacts of regulation of Artificial Intelligence

Contents

Abstract 3
1. Introduction 4
2. Why AI needs to be Regulated 4
2.1 Market Concentration 4
2.2 Potential for Bias 5
2.3 Privacy Concerns 6
2.4 Hallucinations, Disinformation and Deep Fakes 7
2.5 Intellectual Property Infringement 7
2.6 Black Box and Lack of Alignment 7
2.7 Existential or Near-existential Threat 8
3. Informing the design of AI regulation 9
3.1 Recognising the AI Lifecycle 9
3.2 Using AI as a Regulatory Tool 11
3.3 The Balance between Innovation and Regulation 11
3.4 Global Cooperation and the Risk of AI Arms Race 12
3.5 The boundaryless nature of AI risk 13
4. Principles for a Regulatory Framework 14
4.1 General Principles for AI regulation 14
4.2 The Precautionary Principle 16
5 AI Regulations Assessment Frameworks and Approaches 17
5.1 Potential AI Risk and Cost Benefit Analysis (CBA) 17
5.2 Quantitative Assessment using environmental precedent 19
5.3 Quantitative assessment using breakeven analysis 21
5.4 Qualitative Assessment using breakeven analysis 21
5.5 Qualitative assessment using the Precautionary Principle 23
5.6 Using Real Options in assessing AI Regulations 23
6. Conclusions 26

Challenges in assessing the impacts of regulation of Artificial Intelligence

Stephen Gibson[1] and Winston Tang[2]

Abstract

The recent surge in Generative Artificial Intelligence has introduced both opportunities and risks to society. This paper discusses the challenges in assessing the impacts of regulation of AI. It identifies a range of different concerns that might give rise to AI regulation and sets out approaches that may inform the design of AI regulation as well as principles for a robust AI regulatory framework.

The paper focuses on the methodologies and challenges involved in evaluating the impacts of AI regulation particularly where there is both significant uncertainty around the costs and benefits of the proposed regulation and the potential for near-existential risk, meaning that AI regulatory proposals are not easily susceptible to standard cost benefit analysis approaches. It outlines and considers the use of a range of quantitative and qualitative approaches to the assessment of AI regulatory proposals including breakeven analysis, using real options and applying the precautionary principle.

Given the potential for significant and near-existential harm from AI, it seems reasonable and appropriate that policymakers should err on the side of caution in designing AI regulation in line with the precautionary principle. However, there are important insights from both the approach adopted for assessing environmental regulations in terms of developing a standardized metric of regulatory risk, developing more robust qualitative reasoning, and also considering the regulatory framework as a real option, implying that policymakers should retain flexibility and monitor developments in designing regulations as new evidence becomes available. This effectively views AI regulation as an investment with embedded real options (to delay, expand, revise or abandon). It requires ongoing monitoring of the effectiveness (or otherwise) of regulation and implications of wider developments in the AI space, as well as a willingness to re-open regulatory decisions in the light of new information.

[1] M-RCBG Senior Fellow, Harvard Kennedy School; Senior Fellow, London School of Economics and Political Science; Chair of the UK Regulatory Policy Committee; Member of the Bank of England cost-benefit analysis panel; Director SLG Economics. This article is written in a personal capacity and the views expressed do not necessarily reflect the views of the Regulatory Policy Committee, the UK Government, the Bank of England or SLG Economics.
[2] Master in Public Policy candidate, Harvard Kennedy School.

1. Introduction

Artificial Intelligence (AI) is transforming society, generating both unparalleled opportunities and unprecedented risks. As AI applications permeate essential sectors like healthcare, finance and public safety, the potential costs to society associated with its development and deployment are high, spanning from social inequality to existential risks. This requires the introduction of comprehensive and well-balanced regulation, but at the same time avoiding overregulation that might deter AI innovations that could benefit society.

This paper considers how regulators and government should approach the design of AI regulation by understanding the risks of AI, providing a strategic basis for AI governance based on a rigorous understanding of the various stages of AI lifecycles, clear principles for a regulatory framework, and presenting both quantitative and qualitative assessments of the impacts of AI regulations. While many of the challenges associated with regulating AI are similar to those associated with other technologies or market failures across the economy, this paper highlights two aspects of AI that pose particular challenges to public policy assessment – the potential for near-existential risk and the speed and unpredictability of AI development.

These challenges warrant a proactive, flexible and precautionary approach, erring on the side of caution in designing AI regulations that seek to address potentially near-existential risks. This could involve using the precautionary principle and seeking a standardised proxy for AI risk (similar to the use of CO2 equivalent emissions in environmental regulation). Other potential methodological approaches include the use of qualitative and quantitative break-even analysis and applying real options methodology. Given that much of the scope and magnitude of risks associated with AI is unclear and constantly changing, AI regulations should be designed to anticipate and mitigate escalating risks while also being flexible enough to address unforeseen challenges and incorporate learning and feedback.

2. Why AI needs to be Regulated

This section identifies some of the risks and concerns around AI and the reasons why government may seek to impose a range of different regulatory remedies on the technology.

2.1 Market Concentration

The barriers to entry in AI development are high. To train generative AI models, an enormous volume of data is required. Meta's newest generation of Llama 4 AI models, for example, were pre-trained on over 30 trillion tokens (more than double the volume

used for its Llama 3 model)[3], such a level of data collection is hard to achieve. Companies like Google, Amazon and Meta have massive, proprietary datasets gathered from their users over the years. It is challenging for new companies to access similar datasets without significant financial or technological resources, posting hurdles for potential new players to enter the market and compete.

Additionally, to train such a huge amount of data, strong computing power is necessary. A set of complex and expensive infrastructure, including high-performance hardware, data storage, efficient networking infrastructure, as well as power and cooling solutions are needed. Small companies will usually have to rely on alternative cloud services, provided by companies like Amazon (AWS), Microsoft (Azure), and Google (Google Cloud). For example, Anthropic, a new player in the AI space, has an agreement with Amazon to use its cloud services and chips to build its models[4]. The UK Competition and Markets Authority has started investigating whether such arrangements might harm competition[5]. However, DeepSeek, a Chinese company developing large language models (LLMs) and an associated chatbot, trained its flagship model at a fraction of the cost of competitors using less powerful Nvidia H800 chips (due to US export controls on faster chips), suggesting that the barriers to entry may be significantly lower than previously thought.

Developing cutting-edge AI also requires highly specialised technical expertise which is expensive, in short supply and concentrated in large companies[6]. They can also invest in or acquire promising AI startups founded by such talents. Large existing AI companies' ability to hoard top talent and acquire emerging companies adds to the antitrust concerns in the AI space.

2.2 Potential for Bias

Another risk associated with AI is the potential for bias in its outputs. Both traditional and generative AI rely on the input data used to train the models and generate outputs.

---

[3] Meta AI. "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation." Meta AI Blog, 5 Apr. 2025, https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[4] Amazon Staff. "Amazon and Anthropic Deepen Their Shared Commitment to Advancing Generative AI." *Amazon*, 27 March 2024, https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment.

[5] Browne, Ryan. "Amazon's $4 billion investment in AI firm Anthropic faces UK merger investigation." *CNBC*, 8 August 2024, https://www.cnbc.com/2024/08/08/amazons-investment-in-ai-firm-anthropic-faces-uk-merger-investigation.html.

[6] "Superstar coders are raking it in: for a few AI whizzes, pay is going ballistic", The Economist, July 2025 https://www.economist.com/business/2025/07/01/superstar-coders-are-raking-it-in-others-not-so-much?utm_content=ed-picks-image-link-2&etear=nl_today_2&utm_campaign=r.the-economist-today&utm_medium=email.internal-newsletter.np&utm_source=salesforce-marketing-cloud&utm_term=7/1/2025&utm_id=2091192

Therefore, without careful consideration during the data collection and cleaning process, there is a risk that bias in historic training data transfers through into the AI output.

For example, a software used by many U.S. courts[7] to predict the likelihood of a defendant reoffending[8] mistakenly predicted Black defendants who did not reoffend had a 44.9% likelihood of committing another crime - almost double the 23.5% likelihood estimated for white defendants[9]. Although the training data did not include racial information, other aspects of the data may have been correlated to race, such as employment history, drug problems, and vocational / educational problems[10]. The 'black box' nature of AI models (see Section 2.6) can make it more difficult to identify and regulate examples of this 'unintentional bias'. Therefore, AI can reinforce historic bias and lead to unequal treatment and discrimination.

2.3 Privacy Concerns

Many AI models are trained on massive datasets from publicly available sources like government records, social media and blogs, many of which have personal information that may not have intended or agreed to be used for the purpose of training AI models. Techniques like web scraping can collect information from public platforms without individual consent - for example yt-dlp, a web scrapping tool, allows users to pull content from YouTube[11] and Meta has trained its AI models on text and photos published on Facebook and Instagram since 2007[12].

AI systems can also collect private and sensitive information from users' input via platforms like AI chatbots[13] or by including information from data leaks in their training data[14]. This allows potential misuses such as identity theft, digital profiling, bias and

---

[7] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
[8] Corbett-Davies, Sam, Pierson, Emma, Feller, Avi, and Goel, Sharad. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*, 17 October 2016, https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.
[9] Dressel, Julia, and Farid, Hany. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances*, 17 January 2018, https://www.science.org/doi/10.1126/sciadv.aao5580.
[10] Hamilton, Melissa. "Justice Served? Discrimination in Algorithmic Risk Assessment." *Research Outreach*, 19 September 2019, https://researchoutreach.org/articles/justice-served-discrimination-in-algorithmic-risk-assessment/.
[11] McGrady, Ryan, Zuckerman, Ethan, and Zheng Kevin "AI Companies Threaten Independent Social Media Research." *TechPolicy.Press*, 30 Jan. 2025, https://www.techpolicy.press/ai-companies-threaten-independent-social-media-research/.
[12] Weatherbed, Jess. "Meta Fed Its AI on Almost Everything You've Posted Publicly Since 2007." *The Verge*, 12 September 2024, https://www.theverge.com/2024/9/12/24242789/meta-training-ai-models-facebook-instagram-photo-post-data.
[13] Li, Jingquan. "Security Implications of AI Chatbots in Health Care." *Journal of Medical Internet Research*, vol. 25, 28 November 2023, https://www.jmir.org/2023/1/e47551.
[14] Trinckes, Jay. "AI Data Breach: Understanding Their Impact and Protecting Your Data." *Thoropass Blog*, Accessed 25 March 2025, https://thoropass.com/blog/compliance/ai-data-breach/.

discrimination, exclusion and social embarrassment[15]. Models can also disclose personal data they collect and train on to third parties without user consent[16].

2.4 Hallucinations, Disinformation and Deep Fakes

AI models, such as LLMs operate by predicting the next word or sequence based on pattern recognition without verifying the validity of the output they produce. Therefore, output can appears informative but may be inaccurate or simply untrue — a phenomenon called 'hallucinations'.

Additionally, deep fakes created by advanced generative models can be difficult or impossible to distinguish from genuine photos or videos. For example, during the 2024 U.S. presidential election, AI was used to generate deep fake video and voice-cloning content to spread disinformation on candidates, dissuade people from voting with false information, confuse poll workers and disrupt polling locations[17].

2.5 Intellectual Property Infringement

AI systems also raise concerns about intellectual property and copyright infringement. Training data used by AI models may include copyrighted data such as paintings, photos, and writing. The new content generated by systems like DALL-E, ChatGPT, and Stable Diffusion can incorporate elements or styles of the copyrighted materials without the creators' consent. The black-box nature of AI models and the mass collection of data, along with the nuances of intellectual property laws, have raised copyright infringement issues. In the United States, there is an ongoing debate on the application of the Fair Use Doctrine in AI-generated content; debated issues range from the purpose and amount of use of copyrighted materials in AI-generated content, to the nature and market impact of AI-generated work[18].

2.6 Black Box and Lack of Alignment

AI models have internal decision-making processes that are hard for humans to understand (a 'black box' problem). These models use complex machine learning and

---

[15] Li, Jingquan. "Security Implications of AI Chatbots in Health Care." *Journal of Medical Internet Research*, vol. 25, 28 November 2023, https://www.jmir.org/2023/1/e47551.
[16] Duffourc, Mindy Nunez, Gerke, Sara, and Kollnig, Konrad. "Privacy of Personal Data in the Generative AI Data Lifecycle." *NYU Journal of Intellectual Property & Entertainment Law*, 8 July 2024, https://jipel.law.nyu.edu/privacy-of-personal-data-in-the-generative-ai-data-lifecycle/.
[17] Dwyer, Devin, and Herndon, Sarah. "AI Deepfakes a Top Concern for Election Officials with Voting Underway." *ABC News*, 18 October 2024, https://abcnews.go.com/Politics/ai-deepfakes-top-concern-election-officials-voting-underway/story?id=114202574.
[18] Congressional Research Service. "Generative Artificial Intelligence and Copyright Law." *Congress.gov*, 29 September 2023, https://crsreports.congress.gov/product/pdf/LSB/LSB10922.

deep learning algorithms that lack transparency, making it difficult for users to grasp the reasoning and logical steps behind AI-generated predictions or decisions. This opacity raises significant issues for accountability, interpretability, potential ethical violations, and trustworthiness — factors critical for adoption of AI, especially in sensitive or high-stakes fields[19].

Relevant to the black box problem is the issue of lack of alignment. This happens when AI fails to execute tasks assigned by humans due to a false interpretation of the true objectives of the task or a failure to adjust to a changing environment. This could involve the AI executing the literal implications of a prompt rather than the full human intention or the AI continuing to pursue specified goals even when the initial environment or data distribution changes.

2.7 Existential or Near-existential Threat

One of the most mentioned and feared risks of AI (often appearing as a theme in science fiction novels and dystopian films) is the potential existential threat it poses to humanity.

Generally, there are four types of AI existential or near-existential risks[20].

- **Malicious use**, such as weaponising AI for bioterrorism autonomous weapons systems, Chaos-GPT[21], or slaughter bots[22]. These malicious uses of AIs can be autonomous, potentially causing large-scale devastation if humans lose control of the operation of AI or if law enforcement fails to obtain control.
- **AI race**, particularly the pressure to compete and dominate in AI technologies, in a corporate or military setting leading to unsafe and excessive development and deployment of AI systems without safety or ethical considerations (for example, keeping a 'human in the loop' – and whether even retaining human involvement is an effective safeguard[23]).

---

[19] Rawashdeh, Samir. "AI's Mysterious 'Black Box' Problem, Explained." *University of Michigan–Dearborn News*, 6 March 2023, https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained.
[20] Hendrycks, Dan, Mazeika, Mantas, and Woodside, Thomas. "An Overview of Catastrophic AI Risks." *Center for AI Safety*, https://www.safe.ai/ai-risk.
[21] Koebler, Jason. "Someone Asked an Autonomous AI to 'Destroy Humanity': This Is What Happened." *VICE*, 4 April 2023, https://www.vice.com/en/article/someone-asked-an-autonomous-ai-to-destroy-humanity-this-is-what-happened.
[22] Allen, Felix. "Terrifying Rise of AI 'Slaughterbots' Could Wipe Out Humanity." *News.com.au*, 22 December 2021, https://www.news.com.au/technology/innovation/military/terrifying-rise-of-ai-slaughterbots-could-wipe-out-humanity/news-story/a4c4886489bb544e69a222442514d6a8.
[23] Peter Rautenbach, "Keeping humans in the loop is not enough to make AI safe for nuclear weapons", *Bulletin of the Atomic Scientists*, February 2023 https://thebulletin.org/2023/02/keeping-humans-in-the-loop-is-not-enough-to-make-ai-safe-for-nuclear-weapons/#:~:text=interacting%20with%20it.-,Keeping%20humans%20in%20the%20loop%20is%20not%20enough%20to%20make,humans%20or%20in%20their%20place.

- **Operational failure or organisational risks,** causing a loss of control. Complex AI systems, some of which are used for security or military purposes, can suffer from accidental misuse and intentional leaking or stealing of source codes as well as intentional harm and hacking by bad actors.
- **Rogue AIs,** AI systems that 'outsmart' human beings, optimising their ability to deliver flawed (and potentially catastrophic) objectives. This leads to the risk of loss of control (particularly when accompanied by lack of alignment).

While there is no known incident of loss of control over an AI system and some dispute whether generative AI poses existential threats[24], AI technologies continue to evolve and become more layered and sophisticated with an increased ability to act independently. Closed proprietary systems (like Chat GPT) are typically accessed via an Application Programming Interface (API) with built-in safety filters and usage policies making it more difficult for users to produce harmful outputs. In contrast, open-source models (such as Llama 3.1 and Gemma) whose code and weights are publicly available, can be run and modified without restriction (including removing safeguards and filters), potentially enabling malicious actors to harness advanced AI capabilities with little oversight. The UK government's AI risk assessment[25] highlights this distinction noting the qualitatively different risks of malicious use posed by open-source models.

3. Informing the design of AI regulation

This section identifies factors that should inform the design of AI regulation: the AI lifecycle, AI's potential as a regulatory tool, a need for a balance between fostering innovation and enforcing oversight, and the importance of global cooperation.

3.1 Recognising the AI Lifecycle

To effectively regulate AI systems, it is important to consider the AI lifecycle and the potential for regulation at different stages[26] of this process. Designing regulations to address the various critical points of the AI lifecycle can help mitigate the associated risks and ensure that the regulation is as effective as possible in securing its objectives

---

[24] Lu, Sheng, Bigoulaeva, Irina, Sachdeva, Rachneet, Tayyar Madabushi, Harish, and Gurevych, Iryna. "Are Emergent Abilities in Large Language Models Just In-Context Learning?" *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, August 2024, pp. 5098–5139, https://aclanthology.org/2024.acl-long.279/.

[25] *Safety and Security Risks of Generative Artificial Intelligence to 2025,* Department for Science, Innovation and Technology, April 2025 https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/safety-and-security-risks-of-generative-artificial-intelligence-to-2025-annex-b

[26] U.S. Department of Commerce, National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST, July 2024, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

(although AI development is not linear and involves iterative feedback loops which need to be taken into account). Regulation also has to differentiate between regulating the application of the technology (for example its use in education or healthcare) and regulating the underlying development of the technology itself (for example requiring the developer to establish a risk management system throughout the AI system's lifecycle).

The first stage in AI development is design, training, and testing. This determines the system's scope and use, training with input data of the developer's choice, and testing it based on different evaluation thresholds for various standards such as accuracy and precision, generalisation and compliance with ethical standards [27]. This stage is embedded with the risks of AI model biases, ethical and privacy concerns, technical and safety vulnerabilities, and misalignment. During this stage, developers have full control over the design, training and testing processes, which regulation can potentially tightly control. However, even with well-designed regulations, AI systems may still harbour hidden biases or weaknesses that only emerge when deployed in real-world settings, often as a result of unforeseen interactions with diverse data and user behaviors that were not fully accounted for during development[28].

The second stage is initial deployment and usage. Here, potential risks can materialise through accidents or operational failures, intentional misuse and ethical violations, as well as malicious use and security risks. During this stage, developers and companies have discretion over the access, usage and application of AI models. For example, OpenAI limits API access for malicious actors though strict usage monitoring, rate limits and account verification processes[29]; it also refuses to answer questions that are harmful, offensive, discriminatory or could potentially incite violence[30]. However, due to the size of the user base and the complexities of user intent, it would be challenging for developers (and therefore regulators) to fully prevent AI models from being misused by malicious actors.

The final stage of the AI life cycle is longer-term diffusion[31]. Broadly speaking, there are two types of AI systems or products that are particularly relevant. The first is products like

[27] In some cases, **general-purpose** models such as a foundation models (FMs) are trained broadly and can be later repurposed for specific tasks. AWS. "What Is a Foundation Model?" Amazon Web Services, 2025, https://aws.amazon.com/what-is/foundation-models/.
[28] Ferrara, Emilio. "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies." *Sci*, vol. 6, no. 1, 2024, https://www.mdpi.com/2413-4155/6/1/3.
[29] OpenAI. "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors." *OpenAI*, 14 February 2024, https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/.
[30] Gewirtz, David. "6 Things ChatGPT Can't Do—and Another 20 It Refuses to Do." *ZDNet*, 16 February 2023, https://www.zdnet.com/article/6-things-chatgpt-cant-do-and-another-20-it-refuses-to-do/.
[31] Hansen, Hans Fredrik, Lillesund, Elise, Mikalef, Patrick, and Altwaijry, Najwa. "Understanding Artificial Intelligence Diffusion Through an AI Capability Maturity Model." *Information Systems Frontiers*, vol. 26, 2024, pp. 2147–2163, https://link.springer.com/article/10.1007/s10796-024-10528-4.

ChatGPT where the user base is large and usage is widely applicable in many aspects of life. As this kind of product rapidly spreads, its impact can soon be deeply integrated and diffused across society. The second category involves AI systems that might not have a large public audience but are an integral part of future AI models which build on each other. This layered approach is important to advance AI capabilities, improve performance, and address limitations of earlier models, but brings the risk of compounding errors or biases from earlier models, especially as they become integrated into high-impact domains. Compared to the other two stages, the final stage is more difficult to regulate as it concerns an accumulation of AI development. At this stage regulation can encompass not just the product concerned, but also other ways of minimizing harm – for example through user education about risks associated with usage and restrictions on who can use different products[32].

3.2 Using AI as a Regulatory Tool

The second consideration is the potential for leveraging AI to assist in guarding against AI risks. For instance, to address the 'black box' concerns of AI models, algorithms such as Explainable AI (XAI) attempt to describe an AI model's impacts and potential biases in ways that humans can easily comprehend[33]. Additional tools can help monitor AI-generated disinformation and evaluate AI models to identify and correct biases[34]. For instance, a Vision-Language Disinformation Detection Benchmark (VLDBench) tool supports both unimodal (text-only) and multimodal (text and image) disinformation detection[35].

3.3 The Balance between Innovation and Regulation

The tension between promoting innovation and precautionary regulation is prevalent in many regulatory fields, however, in the case of AI, it is amplified by the distinct features of AI technologies: its rapid speed of development, diffusion and integration, the essential need for data and the large scale of social impact. Therefore, while AI systems present a variety of risks, the government needs to consider how to balance its approach to AI

[32] School curriculums in the UK include lessons about the use and dangers of social media age-appropriate lessons on online safety, responsible use of technology, and how to identify and report harmful content https://www.gov.uk/government/publications/teaching-online-safety-in-schools/teaching-online-safety-in-schools

[33] IBM. "What Is Explainable AI?" *IBM Think*, 29 March 2023, https://www.ibm.com/think/topics/explainable-ai.

[34] Slapakova, Linda. "Towards an AI-Based Counter-Disinformation Framework." *RAND Corporation*, 29 March 2021, https://www.rand.org/pubs/commentary/2021/03/towards-an-ai-based-counter-disinformation-framework.html.

[35] Raza, Shaina, et al. "VLDBench: Vision Language Models Disinformation Detection Benchmark." arXiv, 17 Feb. 2025, https://arxiv.org/pdf/2502.11361.

regulation so that it avoids stifling desirable technological progress and supports international competitiveness while still avoiding consumer and societal harm.

There is an opportunity for the UK to establish itself as a pioneer in the global regulatory space to deliver more effective AI regulations that incentivise AI development, improve productivity and deliver economic growth while ensuring public safety and addressing the most fundamental risks of AI. The UK Department for Science, Innovation and Technology and Office for Artificial Intelligence has positioned the UK's approach as 'pro-innovation'[36], and has set out a 3 phase approach to providing guidance on AI, the first of which[37] sets out initial guidance for regulators to use in developing a principles-based regulatory framework within their remit. The principles are: safety, security & robustness; appropriate transparency and explainability; fairness; accountability and governance; and, contestability and redress.

3.4 Global Cooperation and the Risk of AI Arms Race

A final consideration for AI regulations is the need for global cooperation. Just like other technologies or commercial products, there is a competitive aspect to AI technologies between different companies and countries. However, equally important is the need for countries to work together to address some of the shared fundamental risks of AI and to leverage the technology for the common public good. As a global phenomenon, AI has applications and implications that cross national borders and regulating AI will require international cooperation to set standards for safety, transparency, and ethical use and to ensure interoperability. In 2023, the UK government organised the first AI Safety Summit[38], a global platform that brought together 28 countries to discuss the risk of AI and led to the signing of the Bletchley Declaration on AI safety[39]. This event set the precedent for a global platform for AI regulation discussion to foster global collaboration on devising effective AI governance frameworks. However, the huge profit and strategic opportunities for the firms and countries who are first to develop new AI models and applications may threaten any political consensus to regulate the risks involved.

---

[36] Department for Science, Innovation and Technology and Office for Artificial Intelligence. "AI Regulation: A Pro-Innovation Approach." *GOV.UK*, 29 March 2023, https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach.

[37] Implementing the UK's AI Regulatory Principles: Initial Guidance for Regulators, DSIT, February 2024, https://assets.publishing.service.gov.uk/media/65c0b6bd63a23d0013c821a0/implementing_the_uk_ai_regulatory_principles_guidance_for_regulators.pdf

[38] Foreign, Commonwealth & Development Office, Department for Science, Innovation and Technology, and AI Safety Institute "AI Safety Summit 2023." *GOV.UK*, 1 November 2023, https://www.gov.uk/government/topical-events/ai-safety-summit-2023.

[39] Ibid.

In an 'AI arms race', countries may disregard risks to gain a competitive advantage in AI development[40], causing a 'race to the bottom', mirroring aspects of the nuclear-arms race[41]. The key distinction between nuclear weapons and AI technologies is that the former is almost exclusively state-controlled and the risks are obvious and immediate, while AI is largely developed and driven by the private sector with more diffuse and subjective risks. This may make it easier for Governments (if they are willing) to agree to restrict nuclear proliferation; while even if Governments are able agree global AI regulation, this then needs to flow through to regulation of the businesses concerned.

Associated with the risk of an AI arms race is the strategic ambiguity of global AI cooperation due to the potential conflict between national interests and international considerations. In the area of climate change, countries are still debating how to balance domestic political concerns of possible negative electoral or short-term economic impacts brought by carbon or net-zero policies, and the compliance with international agreements like the Climate Agreement[42]. This problem of free riding, short-termism and lack of trust is also likely to inhibit and obstruct global cooperation in the AI space.

3.5 The boundaryless nature of AI risk

Another challenge that comes with AI risk is the relatively boundaryless nature of AI development and impacts. For example, even if strict UK and EU regulations prohibit AI models that can lead to widespread disinformation, these regulations will have a limited impact if disinformation is originated by AI companies in the US and then spreads to the UK or EU via social media. While AI models that cause significant harm may be banned in specific jurisdictions, they can still proliferate globally through cloud-based services, open-source releases and cross-border digital trade, limiting the effectiveness of UK regulations. In these cases, it will be challenging to assess the impacts of AI regulations in the UK, unless the UK can technologically isolate its market from external AI technologies to a significant extent.

[40] Meacham, Sam. "A Race to Extinction: How Great Power Competition Is Making Artificial Intelligence Existentially Dangerous." *Harvard International Review*, 8 September 2023, https://hir.harvard.edu/a-race-to-extinction-how-great-power-competition-is-making-artificial-intelligence-existentially-dangerous/.
[41] BlueDot Impact AI Safety Fundamentals Team. "Historical case studies of technology governance and international agreements (compilation - various authors)." *BlueDot Impact Blog*, 29 August 2023, https://bluedot.org/blog/historical-case-studies.
[42] Tosun, Jale, and Peters, B. Guy. "The Politics of Climate Change: Domestic and International Responses to a Global Challenge." *International Political Science Review*, vol. 42, no. 1, January 2021, pp. 3–15, https://journals.sagepub.com/doi/full/10.1177/0192512120975659.

4. Principles for a Regulatory Framework

Regulatory responses to potentially paradigm-changing technologies can (with the benefit of hindsight) fail to recognise either the threats posed by new technology and / or the benefits available from them. The Locomotives Act 1865[43] required a man with a red flag to walk at least 60 yards ahead of each vehicle to warn riders and the drivers of horses of the approach of the vehicle and set a speed limit of 4mph (2mph in towns) for road locomotives – hardly a measure designed to promote innovation[44]. On the other hand, the development of internet regulation was sporadic with "*no coherent strategy for regulation* … and *a failure to take early steps to structurally regulate the internet and instead focusing on individual harms*"[45].

4.1 General Principles for AI regulation

Given the complex nature of AI systems, it is imperative for the government to be proactive and flexible in its regulatory approach. Regulating AI is an adaptive governance process, requiring continuous recognition of and adjustments to new developments and challenges in the field. The government should therefore consider six key regulatory principles.

- **Consistency** - ensuring a uniform standard in evaluating and regulating AI, to avoid conflicting regulations from different government departments that confuse or complicate the compliance process for AI developers and deployers[46]. This also requires consistency across business sectors and industries to prevent regulatory arbitrage, where companies take advantage of gaps in regulations and operate in sectors or areas that have fewer or lighter restrictions[47].

- **Transparency** - building public trust and accountability by ensuring that stakeholders can understand the regulations and the decision-making or output-generating process of the AI models[48]. This principle is important for AI technologies that affect core sectors in society, such as civic rights, healthcare, energy, finance, and criminal justice. A relevant legislative precedent is the EU AI Act, which sets transparency

[43] An Act for further regulating the Use of Locomotives on Turnpike and other Roads for agricultural and other Purposes. 1865 https://www.legislation.gov.uk/ukpga/Vict/28-29/83/enacted
[44] It was repealed by the Locomotives on Highways Act 1896
[45] Regulating AI and Machine Learning: Setting the Regulatory Agenda, J Black and A Murray, December 2019 https://www.ejlt.org/index.php/ejlt/article/view/722
[46] Google. "Recommendations for Regulating AI." *Google*, Accessed 23 March 2025, https://ai.google/static/documents/recommendations-for-regulating-ai.pdf.
[47] Partnoy, Frank. "The Law of Two Prices: Regulatory Arbitrage, Revisited." *Georgetown Law Journal*, vol. 107, no. 4, April 2019, pp. 1017–1042, https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2019/04/5The-Law-of-Two-Prices_Regulatory-Arbitrage-Revisited_Partnoy.pdf.
[48] OECD. "Transparency and Explainability (OECD AI Principle)." *OECD.AI*, Accessed 23 March 2025, https://oecd.ai/en/dashboards/ai-principles/P7.

obligations for AI systems providers and deployers based on their risk level[49]. UNESCO's Recommendation on the Ethics of Artificial Intelligence also lists transparency and explainability as parts of its human-rights approach to AI[50].

- **Accountability** – ensuring that developers, providers and users of AI systems are held responsible for negative consequences arising from their actions. The EU AI Act, for example, requires providers of high-risk AI systems to set a quality management system, which contains "*an accountability framework setting out the responsibilities of the management and other staff*."[51] It also authorises market surveillance authorities to access the source code of the high-risk AI system "*upon a reasoned request*"[52]. Essentially, along with transparency, the accountability principle aims to ensure that any harm caused by AI can be traced back to responsible individuals or businesses.

- **Targeting** - government should avoid a 'one-size-fits-all' approach to regulating the AI industry. Instead, it should focus on companies and specific areas within AI systems that pose the highest potential impact or risk[53]. Given the rapid and far-reaching nature of AI development, the government should (as far as is possible) prioritise its regulatory capacity and resources by concentrating on entities and aspects of AI technology that require close monitoring due to the significant risks they pose to society.

- **Proportionality** - designing regulations that are proportional to the level of risk the AI system may pose[54]. This aims to avoid unnecessary burdens on AI enterprises with low risks, to avoid unnecessarily stifling innovation, and ensuring that government resources are focused on areas where they can best protect the public interest.

- **Fairness** – this is important for the legitimacy of the system and for addressing and mitigating the risks posed by potential biases in AI systems[55]. It aims to prevent the worsening of, or addition to existing social inequalities. AI systems that discriminate

[49] EU AI Act. "Key Issue 5: Transparency Obligations." *EU AI Act*, https://www.euaiact.com/key-issue/5.
[50] United Nations Educational, Scientific and Cultural Organization (UNESCO). *UNESCO's Recommendation on the Ethics of Artificial Intelligence*. UNESCO, 23 November 2021, https://unesdoc.unesco.org/ark:/48223/pf0000385082/PDF/385082eng.pdf.multi.
[51] EU AI Act. "Article 17: Quality Management System." *EU AI Act*, Accessed 23 March 2025, https://www.euaiact.com/article/17.
[52] Ibid.
[53] Sullivan, Scott. "Targeting in the Black Box: The Need to Reprioritize AI Explainability." *Lieber Institute – West Point*, 28 August 2024, https://lieber.westpoint.edu/targeting-black-box-need-reprioritize-ai-explainability/.

[54] Ibid.
[55] Digital Regulation Cooperation Forum (DRCF). "Fairness in AI: A View from the DRCF." *DRCF*, 15 April 2024, https://www.drcf.org.uk/publications/blogs/fairness-in-ai-a-view-from-the-drcf.

based on race, gender or sexual orientation, for example, are in violation of this principle.

4.2 The Precautionary Principle

The precautionary principle is used to support the decision-making process in areas that lack scientific certainty[56]. Such a process usually concerns risks that may not be precisely calculable in advance. Its purpose is to allow action or intervention even if the full extent or likelihood of harm cannot be confidently assessed. In an environmental context, this approach aims to prevent harm to human health, animal health, plant health, or the environment in situations where there is credible evidence of potential risks, but insufficient scientific clarity to assess those risks fully.

The Interdepartmental Liaison Group on Risk Assessment (ILGRA) guidance outlines specific criteria and steps for invoking the precautionary principle[57]. The guidance gives two prerequisites that must be met for the precautionary principle to be used:

- there is good reason to believe that harmful effects may occur to humans, animal or plant health, or to the environment; and
- the level of scientific uncertainty about the consequences or likelihoods is such that risk cannot be assessed with sufficient confidence to inform decision-making."[58]

Next, there are four steps under the precautionary principle that policymakers need to complete:

- policymakers should review the evidence to assess the existence and extent of harmful effects: empirical evidence of the actual harm, empirical evidence of an analogous harm or analogous activity / product / situation causing harm, and/or a strong theoretical argument that harm will result.
- policymakers should assess the possible impacts of inaction and whether such impact justifies proposed policies when the likelihood of harm and risks are unclear.
- policymakers need to determine that risks cannot be evaluated with adequate confidence to inform decision-making based on scientific uncertainty.

---

[56] U.K. Department for Environment, Food & Rural Affairs. "Environmental Principles Policy Statement." *GOV.UK*, 31 January 2023, https://www.gov.uk/government/publications/environmental-principles-policy-statement/environmental-principles-policy-statement.
[57] Regulatory Policy Committee. "Short Guidance Note: Precautionary Principle." *GOV.UK*, Accessed 23 March 2025, https://assets.publishing.service.gov.uk/media/5e21c408e5274a6c3be72203/short_guidance_note_-_precautionary_principle.pdf.
[58] Ibid.

- policymakers need to consider the need and necessity of invoking the precautionary principle if all criteria from previous steps are met.

As mentioned in Section 2, the rise of AI has brought potential risks to society, many of which are hard to assess. Chief among them is the existential threat of AI, which has both high uncertainty and high consequences, such that the application of the precautionary principle could be warranted.

5 AI Regulations Assessment Frameworks and Approaches

A key part of the UK government's approach to policymaking is providing robust evidence about the impacts (costs, benefits, risks and uncertainties) of the proposed regulation and potential policy alternatives (including 'Do Nothing') to inform the decision-making process. While it is not determinative – governments and regulators can have a wide-range of objectives including those that do not form part of an impact assessment (for example, wider social and political objectives), a rigorous and structured approach to assessing the effects of a proposed regulatory policy is an important part of the government decision-making process.

Given the unprecedented nature of AI, this section will discuss six policy assessment approaches to shed some light on potential approaches to assessing options for AI regulation.

5.1 Potential AI Risk and Cost Benefit Analysis (CBA)

The CBA approach as part of Regulatory Impact Assessments (RIAs) has been used in the UK for all new regulations that have an impact on business since 2010[59]. A key objective is to balance the costs and benefits of regulation to ensure it is neither too restrictive nor too lenient. However, there are two aspects of potential AI regulation that are not easily susceptible to a standard CBA approach – the potential for existential or near-existential threats and the difficulty and uncertainty around forecasting costs and benefits. Not all harms are equal in unpredictability, rapidity of development or impact and the appropriate regulatory response needs to be tailored to the particular area of concern.

Many potential AI-related regulations can be assessed in the same way as other regulations, collecting both quantitative and qualitative data on the impacts to see if the proposal is net-beneficial and disaggregating them to see if there are groups who are disproportionately affected (for example, vulnerable groups or small or micro businesses).

[59] Gibson, Stephen. "The Role of Regulatory Impact Assessments in the UK." In *Regulatory Governance: Learnings, Challenges and Way Forward*, edited by Abha Yadav, Routledge, 2025, https://www.routledge.com/Regulatory-Governance-Learnings-Challenges-and-Way-Forward/Yadav/p/book/9781032608952.

The techniques and approaches to doing this are well-understood[60,61] and have been applied widely in the UK[62] and the EU, including in the context of AI regulation[63]. Measures to address the risks in sections 2.1 to 2.5 of this paper would appear to be susceptible to this type of approach.

However, there are arguments that some aspects of frontier AI development has the potential to pose an existential or near-existential threat to humanity[64] (for example, through loss of agency and misalignment of objectives[65]), this leaves policy-makers with the challenge of assessing regulations that seek to reduce that risk. One could argue that given the nature of this threat, even a tiny reduction in an existential risk is worth any cost (for example, a global ban on all research or developments in AI), in the same way as some argue that gene editing in humans poses an existential risk[66] and all research in this area should be banned[67]. However, the development of AI also offers the prospects of huge benefits in terms of improved healthcare, autonomous vehicles, increased economic efficiency and growth etc – PwC estimate global GDP could be up to 14% higher in 2030 as a result of AI (equivalent to $15.7 trillion)[68]. In practice, Governments are clearly willing to trade-off some risk of extinction (or near-extinction) against its potential benefits, suggesting that either the risks are large (but not existential) and the

---

[60] U.K. HM Treasury. "*The Green Book: Appraisal and Evaluation in Central Government*." 14 May 2024, https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government/the-green-book-2020.

[61] Regulatory Policy Committee. "RPC Guidance for Departments and Regulators." *GOV.UK*, 12 April 2019, https://www.gov.uk/government/collections/rpc-guidance-for-departments-and-regulators.

[62] Regulatory Policy Committee opinions on many hundreds of Impact Assessments can be found at: Regulatory Policy Committee. "RPC Opinions." *GOV.UK*, 21 February 2019, https://www.gov.uk/government/collections/rpc-opinions.

[63] For example a system-level AI Impact Assessment (AIIA), developed by the Responsible Artificial Intelligence Institute (RAI Institute) https://www.gov.uk/ai-assurance-techniques/rai-institute-artificial-intelligence-impact-assessment-aiia#limitations-of-the-approach

[64] Bales, Adam, D'Alessandro, William, and Kirk-Giannini, Cameron Domenico. "Artificial Intelligence: Arguments for Catastrophic Risk." *Philosophy Compass*, vol. 19, no. 2, February 2024, https://doi.org/10.1111/phc3.12964.

[65] In August 2022, a survey of 738 AI experts found that 50% of them believed there is a 10% or greater chance that humans will go extinct due to our inability to control AI. In March 2023, the Future of Life Institute, specialized in human extinction risks, published an open letter signed by experts from around the world calling for a six-month pause on advanced AI models. Finally, in May 2023, the Center for AI Safety issued a statement signed by the executives of some of the leading AI companies, including OpenAI, DeepMind, Anthropic, and Turing Award winners. Their message was clear: "*Mitigating the risk of extinction from AI should be a global priority, alongside other societal-scale risks such as pandemics and nuclear war*".

[66] European Parliamentary Research Service. "*Genome Editing in Humans: A Survey of Law, Regulation and Governance Principles*." European Parliament, June 2022, https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729506/EPRS_STU(2022)729506_EN.pdf.

[67] What are the Ethical Concerns of Genome Editing?, National Human Genome Research Institute, August 2017 https://www.genome.gov/about-genomics/policy-issues/Genome-Editing/ethical-concerns

[68] *Sizing the prize: What's the real value of AI for your business and how can you capitalise?,* PwC, https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

probabilities very small, or that targeted regulation can reduce near-existential risks to levels that are acceptable trade-offs against the potential benefits to society.

A CBA would need to consider the benefits of a proposed regulation that aims to lower (by an unknown amount) the risk of a near-existential (but unquantifiable) threat from AI against the (extremely uncertain but large) costs of the AI benefits foregone. This challenge is applicable not just to this subject area, but also to other potential existential or near-existential risks that might be mitigated by regulation – for example the risk of: global warfare, a meltdown in the world's financial markets, a global climate disaster or a global pandemic.

5.2 Quantitative Assessment using environmental precedent

Typically, in assessing the benefits of regulations that reduce the probability of high impact outcomes, one would estimate the cost of the outcome (e.g. a financial crisis) and the reduced likelihood of a crisis occurring as a result of the proposed policy, and then multiply the two estimates together; one might then adjust this estimate in the CBA to take account of societal risk aversion or the precautionary principle.

Because of the inherent uncertainty of what factors (and the multitude of factors) that might lead to a future near-existential AI crisis, it is likely to be extremely difficult to estimate the change in probability of such a crisis as a result of a particular policy proposal – such policies are designing a precautionary framework rather than individually preventing or lowering the particular risk of a crisis. Therefore, probabilities are unlikely to be independent and there are likely to be 'cliff-edges', rather than linear changes in risks.

Estimating the cost of near-existential risks (such as those that AI may pose) appears inherently intractable, with no agreement between experts on the nature or extent of the outcome, therefore, while one could describe qualitatively the potential costs involved under various scenarios, it is unlikely that one could quantify those costs. Nevertheless, there are precedents one can rely on to shed light on a potential way forward, such as existing mechanisms for estimating the cost of carbon emissions with the goal of helping to avoid a global climate catastrophe. There are typically three approaches to estimating the cost of carbon:

- Social cost of carbon: An estimate of the economic damage caused by emitting one ton of carbon dioxide into the atmosphere[69]. This estimate is used to help policymakers decide if a proposed policy to curb climate change is justified.

[69] Rode, Ashwin, Carleton, Tamma, Delgado, Michael, Greenstone, Michael, Houser, Trevor, Hsiang, Solomon, Hultgren, Andrew, Jina, Amir, Kopp, Robert E., McCusker, Kelly E., Nath, Ishan, Rising, James,

- Marginal abatement cost: An estimate of the cost of lowering carbon emissions to help meet a national or international emission reduction target.
- Market prices of emissions allowances: An estimate based on the current and estimated future market values of carbon emissions allowances[70].

In the context of AI regulation, marginal abatement costs and the market price of emissions allowances parallel the benefits foregone by regulation. These could be approximated by assessing potential reductions in profit from the proposed regulation or estimating the amount firms might pay to avoid regulatory constraints, although even here the loss of innovation would be extremely difficult to assess. However, this does not capture the benefits of avoiding the near-existential risk that the regulation is designed to achieve.

All three mechanisms require a single consistent metric for assessing environmental harms: the equivalent impact of a unit of carbon dioxide (CO2e) emission. This is used as a standardised metric to target reduction strategies, for carbon offset and carbon taxes and is appropriate because carbon dioxide is a major driver of climate change.

It does not at first glance appear that any similar standard metric or assessment of harm that is applicable to AI risks, or that there is a quantifiable factor that directly or indirectly relates to the scale of the threat posed by AI. However the USA has attempted to restrict China's development of AI and AI chip design by banning the export to China of high-end chips (eg Nvidia A100 and H100) which are used to power the data centers needed to train AI models and setting limits on the advanced graphics processing units (GPUs) that are permitted to be exported[71]. This suggests that it may be possible to use the number of high-end GPUs or another similar measure as a proxy for AI threat potential and a viable control point for policy intervention. Sastry *et al* suggest that "*relative to other key inputs to AI (data and algorithms), AI-relevant compute is a particularly effective point of intervention: it is detectable, excludable, and quantifiable, and is produced via an extremely concentrated supply chain … policymakers could use compute to facilitate regulatory visibility of AI, allocate resources to promote beneficial outcomes, and enforce restrictions against irresponsible or malicious AI development and usage*".[72] Therefore by

---

and Yuan, Jiacan. "Estimating a Social Cost of Carbon for Global Energy Consumption." *Nature*, vol. 598, 13 October 2021, pp. 308–314, https://www.nature.com/articles/s41586-021-03883-8.

[70] World Bank Group. "What Is Carbon Pricing?" *Carbon Pricing Dashboard* World Bank Group, Accessed 23 March 2025, https://carbonpricingdashboard.worldbank.org/what-carbon-pricing.

[71] US tightens its grip on AI chip flows across the globe, Reuters, January 2025 https://www.reuters.com/technology/artificial-intelligence/us-tightens-its-grip-ai-chip-flows-across-globe-2025-01-13/#:~:text=The%20regulations%20cap%20a%20four,and%20global%20development%20of%20AI.

[72] Computing Power and the Governance of Artificial Intelligence, G Sastry et al, February 2024 https://cdn.governance.ai/Computing_Power_and_the_Governance_of_AI.pdf

monitoring and regulating access to a proxy for AI risk (such as high-end GPUs)[73], it may be possible for policymakers to gain early warning signs of emerging AI capabilities and intervene before deployment, using some of the precedents from environmental regulation.

5.3 Quantitative assessment using breakeven analysis

One technique used in impact assessments where the benefits are uncertain or difficult to quantify involves breakeven analysis. This attempts to determine the minimum benefit required to justify the costs of the regulation. If it can be demonstrated that the (probability adjusted) benefits are of a scale to exceed the expected costs of the measure, then the proposed regulation is net-beneficial even if those benefits cannot be formally quantified.

This approach to policy assessment works well if at least one side of the equation (costs or benefits) is reasonably applicable to quantification and the other side can be reasonably assessed to be higher (or lower) than that value (i.e. the range within which the values could potentially fall is not overlapping). However, in the case of regulation focussed on mitigating near-existential AI risk, the potential costs (the lost profits and growth opportunities from restricting AI) may be as difficult to define and quantify as the potential benefits (a small reduction in the likelihood of existential or near-existential risks). Therefore, it may be difficult to narrow the assessment of costs sufficiently to reasonably assert that the benefits (even in low benefit scenarios) will exceed that level.

5.4 Qualitative Assessment using breakeven analysis

For many proposed regulations it may be impossible to quantitively estimate the impacts, but it may be possible to qualitatively describe them – for example, the UK Artificial Intelligence Regulation Impact Assessment describes many benefits to businesses and consumers, but has a very limited quantification of the impacts[74]. However it may be possible to describe the size of the reduction in probability of a near-existential AI crisis using standard (deliberately unquantified) terminology to categorise the impacts of particular measures. A sample standard could have the following categories to rank AI policy proposals:

- No impact on reducing the risk of a near-existential AI crisis,
- A minimal impact on reducing the risk of a near-existential AI crisis,

[73] Any such proxy for AI risk would need to be updated over time to reflect technological change, compute-efficiency etc.

[74] UK Artificial Intelligence Regulation Impact Assessment, Department for Science, Innovation & Technology , March 2023 https://assets.publishing.service.gov.uk/media/6424208f3d885d000cdadddf/uk_ai_regulation_impact_assessment.pdf

- A small impact on reducing the risk of a near-existential AI crisis,
- A medium impact on reducing the risk of a near-existential AI crisis, or
- A significant impact on reducing the risk of a near-existential AI crisis.

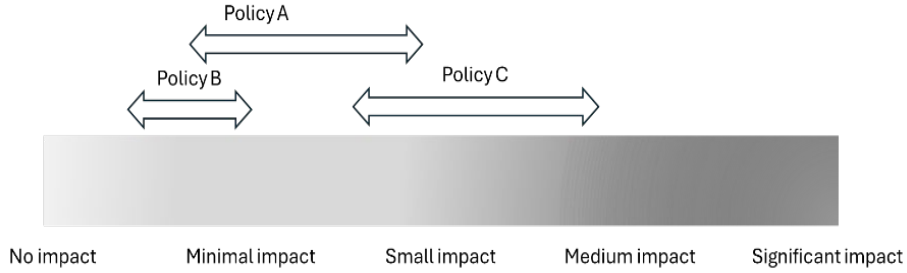
These standard terminologies would be qualitatively described and particular policies justified in the CBA as being in a range along a scale that is described by those standard terms. For example, and purely for exemplification purposes, three potential proposed policies might be categorised as follows:

Policy A: minimal or possibly small impact

Policy B: minimal impact

Policy C: small to medium impact

And this might be shown in the diagram below:



As the government or regulator builds up a portfolio of CBAs, it may be possible to infer logical consequences from the categorisation (a previous policy that had a small impact justified a certain level of costs, so that level should be at least justified for a policy with a medium impact), or apparent inconsistencies (a previous policy that had a medium impact was decided against because the level of costs was too high, but a policy with the same level of costs and a minimal impact was accepted). If one assumed consistency in policymaking, this would allow the qualitative assessment of risk for a new policy to better inform the costs that were justified in seeking to gain the benefits of that policy and therefore provide some basis for regulatory assessment of measures aimed at reducing the risk of a near-existential threats from AI.

However, such a design for qualitative assessment has a risk of “gaming”. Since the categories are subjective and intentionally lacks precise quantification, policymakers or interest groups might strategically use these labels to support or challenge regulations by inflating a weak regulation to be on the level of a mid or high impact category regulation to get more policy benefits or downgrading high impact regulation to the minimum impact category justified for non-implementation. Additionally, potentially artificial thresholds for

different categories might incentivise stakeholders to alter impact descriptions or assessment methodology so that they can artificially meet the policy threshold for a desirable category. It is therefore important to develop a set of tools that can mitigate the potential for gaming the system, such as standardised justification metrics, an independent entity to review and validate categorisations, and appropriate reliance on quantitative assessments whenever possible.

5.5 Qualitative assessment using the Precautionary Principle

As discussed in Section 4.2 above, the precautionary principle sets out two prerequisites – the reasonable expectation of harmful effects and scientific uncertainty about the consequences and likelihoods. As described above both prerequisites are potentially met for near-existential threats from AI – where there is the potential for substantial harm associated with significant uncertainty.

Employing the precautionary principle for AI regulations using the four steps in Section 4.2 allows significant policy flexibility. It provides a solid regulatory groundwork for the government to build up a toolkit of AI regulations, study the regulatory impacts with scalable pilot programs, and rooms for agile changes when more evidence suggests a change in the direction, magnitude or details of existing policies. To obtain this evidence and remain flexible, a robust monitoring and evaluation system specifically designed for AI regulations would need to be established. Such a system could provide continuous assessment mechanisms to evaluate the impact and effectiveness of AI regulations in risk reductions and harm mitigations. Similar existing systems include an AI Incident Database[75], which catalogues real-world instances where artificial intelligence systems have caused or nearly caused harm.

5.6 Using Real Options in assessing AI Regulations

The application of cost-benefit analysis is generally based on a static approach to analysis, often failing to handle the uncertainty inherent in AI developments (pace of innovation, future capabilities or broader societal impacts). Specifically, this approach forecasts the net present value (NPV) of the costs and benefits associated with the regulatory option compared to a 'Do Nothing' baseline. If the regulation's benefits outweigh the costs (appropriately discounted for society's value of time and risk preferences), then there is a prima facie case for the regulation. In this case, delays in introducing such a regulation will negatively impact society by delaying the associated benefits (discounted using the social discount rate). However, delaying the introduction of such a regulation could potentially allow regulators to learn more about its estimated

[75] Responsible AI Collaborative. "*Artificial Intelligence Incident Database*". Accessed May 12, 2025. https://incidentdatabase.ai/.

costs and benefits and those of alternative policy options and allow the regulator to amend the proposed regulation in the light of new information gained in the interim – for example whether the regulation is effective at preventing the potential harm, whether it has unintended consequences or whether it gives rise to perverse incentives or significant unforeseen costs. This new information may be important in informing the development of AI regulatory policy or suggesting further monitoring and waiting.

The expected value of a regulation may be considered as a 'real option' - an opportunity, though not an obligation, to make future regulatory decisions. Real options treat regulatory decisions as investments with embedded options (to delay, expand, revise or abandon), allowing policy-makers to preserve flexibility. The value of the option arises from the opportunity to make different decisions informed by new information in the future, noting that this information is not available when the option is first considered – this can be considered the 'option value' of postponing the regulatory decision. For policy areas like AI regulation, where the technology is developing at an extremely rapid pace and new evidence is being elicited continuously, the option value could be significant and highly relevant to the policy decision.

Real options valuation methods offer a structured way to assess the option value of regulatory actions, using a range of established methodologies[76,77] . In some cases, the process to decide on whether and when to implement new regulations resembles that of a real option that can be incorporated into the cost-benefit analysis[78]. Enabling legislation can provide the government or regulator with the potential, though not the obligation, to enact future regulations. This flexibility holds value, as delaying implementation allows regulators to revise, adjust, or forgo regulation based on lessons and developments observed in the interim[79] (see Box below).

Forecasting costs and benefits of AI regulation is likely to be very difficult, and the level of risk/uncertainty will be very high. Real options may provide a mechanism to manage (at least some of) that uncertainty by highlighting the benefits of acting now vs postponing decisions and waiting until more information is available. It also demonstrates the benefits of closely monitoring the impacts and effectiveness of regulatory measures and maintaining flexibility to allow regulatory design to change to take account of new information as it emerges. Specifically, it provides a grounded framework for regulators

[76] Vonortas, N.S. and Desai, C.A., 2007. 'Real options' framework to assess public research investments. Science and Public Policy, 34(10), pp.699-708

[77] Schwartz, E. S., & Trigeorgis, L. (Eds.). (2001). *Real options and investment under uncertainty: Classical readings and recent contributions*. MIT Press.

[78] Cave, Jonathan, and Stephen Gibson. "Assessing the impacts of primary and secondary legislation – a Real Options approach to rules for making rules." M-RCBG Associate Working Paper Series 2024.223, Harvard University, Cambridge, MA, January 2024.
https://dash.harvard.edu/handle/1/37377675

[79] Ibid.

to conduct scenario analysis based on a systematic understanding of different levels of AI risk. It will, however, require the regulators to stay informed of the ongoing development of AI and its impacts in real-time and adjust their scenario analysis with real options accordingly.

Estimating the costs and benefits of AI regulation is challenging, given the evolving nature of technology as well as its high levels of risk. Real options can offer a way to manage or partially contain this uncertainty by highlighting the value of acting now versus delaying decisions, as postponing until additional information is available can help optimise timing and adaptability in response to changing insights. For example, instead of introducing a full AI regulation immediately, a policy-maker may introduce a phased regulation with a review point after say 2 years – the ability to revise the policy is a real option with economic value (although to evaluate this option requires similar precision on the potential for different outcomes to arise as is required for the initial evaluation). Real options support adaptive regulation, where rules are updated based on new evidence or technological milestones. This helps to align regulation with actual risks and benefits over time rather than fixed and potentially outdated assumptions. Real options help to highlight the irreversibility of some regulatory costs (eg stifling innovation or deterring investment), incorporating this insight helps to balance short-term precaution with long-term innovation potential and promote more resilient and economically efficient regulatory frameworks that evolve with the AI itself.

A simple worked example of the use of real options in regulating a high-risk AI healthcare application*

A government is considering regulating AI systems that provide diagnostic recommendations in healthcare; the risks and benefits are uncertain and depend heavily on future developments. Two policy options are considered (alongside the 'Do Nothing' default option):

- Introduce strict regulation immediately – for example clinical trial evidence required before deployment. This has costs of £100m (compliance, enforcement, delays to innovation) and expected benefits of £120m (reduced misdiagnosis, increased safety)
- Phased regulatory approach – allow the use of regulatory sandbox environments with strict monitoring for 3 years and then decide whether to either introduce the strict regulation (if risks manifest) or relax (if it is safe and efficient). This has an initial cost of £30m (sandbox creation and monitoring); if it proves safe (50% likelihood) then it has £120m benefits but lower compliance costs of £30m, while if it proves risky (50% likelihood) then strict regulation is introduced with £120m benefits and compliance costs of £100m.

Value of immediate regulation: £120m - £100m = **£20m**

Value of flexible approach: 0.5 * (£120m - £30m) + 0.5 * (£120m - £100m) - £30m = **£25m**

In this case, the analysis suggests that it is better to delay the introduction of regulation while monitoring its effectiveness within a regulatory sandbox. It is of course possible that for other regulatory proposals, analysis of the costs, benefits and likelihood will point to immediate regulation (or the 'Do Nothing' option).

* Note, this simple example ignores adjustments for the time value of money for ease of exposition

6. Conclusions

This paper has identified different justifications for the regulation of AI – some of which are similar to those of other sectors / technologies and some (particularly the potential for uncertainty combined with near-existential risk) present novel and challenging problems to regulators and governments in assessing potential regulatory remedies.

It provides suggestions for the approach to designing and developing regulatory solutions and a set of principles for an AI regulatory framework: consistency, transparency,

accountability, targeting, proportionality and fairness, as well as the precautionary principle.

The paper identifies two aspects of potential AI regulation that are not easily susceptible to a standard cost benefit analysis approach to regulatory assessment - the potential for near-existential threat and the difficulty and uncertainty around forecasting costs and benefits in a rapidly changing technological environment.

It therefore considers a variety of approaches to evidence-based policy assessment and considers how far they might be useful in assessing potential regulatory proposals.

- The approach adopted for assessing environment regulations currently lacks applicability due to the lack of a similar metric to CO2 equivalent emissions to apply to the benefits of AI regulation, however an approach based on a proxy for AI risk (such as the number of high-end GPUs), may have the potential to be developed and adopted;
- Quantitative breakeven analysis is potentially applicable if at least the costs of the proposal are well understood or can be estimated within a reasonably narrow band, however if there is significant uncertainty around both the costs and benefits, then it may be difficult to evidence that the benefits clearly outweigh the costs;
- Qualitative breakeven analysis provides a potential approach as the government or regulator builds up a portfolio of regulatory decisions that new proposals can be compared to;
- Real options may offer a methodology for managing some of the uncertainty by assessing whether to act now or postpone a decision until better information is available;
- The precautionary principle provides a route to decision-making in the face of uncertainty, but may lead to over-cautious regulation that fails to maximise the potential opportunities available from AI.

In the face of deep uncertainty around both the costs, benefits and risks of potential AI regulation, it seems reasonable and appropriate that policy-makers should err on the side of caution in designing AI regulation. The use of quantitative breakeven analysis and more robust qualitative reasoning may offer a way forward if the evidence base supports the analysis. In addition, insights from environmental regulatory assessment suggest the benefits of attempting to develop a proxy or standardised metric of AI risk that can be employed in regulatory assessment. However, the real options methodology and approach should be considered in areas where evidence and information is changing rapidly – at a minimum it suggests that in designing regulations, policy-makers should retain flexibility to adjust and revise regulation as new evidence becomes available. This requires ongoing monitoring of the effectiveness (or otherwise) of the regulation and wider

developments in the AI space, as well as a willingness to re-open regulatory decisions in the light of new information.