Towards Responsibly Governing AI Proliferation

Edward Kembery

Supervisor: Dr John Burden

July 2024

This dissertation was originally submitted for the degree of Master of Philosophy at the University of Cambridge, in July of 2024, and has been lightly edited for upload on arXiv. Readers should consult more recent publications for updates on AI developments.

Abstract

This paper argues that existing governance mechanisms for mitigating risks from AI systems are based on the 'Big Compute' paradigm—a set of assumptions about the relationship between AI capabilities and infrastructure—that may not hold in the future.

To address this, the paper introduces the 'Proliferation' paradigm, which anticipates the rise of smaller, decentralized, open-sourced AI models which are easier to augment, and easier to train without being detected. It posits that these developments are both probable and likely to introduce both benefits and novel risks that are difficult to mitigate through existing governance mechanisms.

The final section explores governance strategies to address these risks, focusing on access governance, decentralized compute oversight, and information security. Whilst these strategies offer potential solutions, the paper acknowledges their limitations and cautions developers to weigh benefits against developments that could lead to a 'vulnerable world'.

Contents

1 Introduction

> *All the committees, the politicking and the plans would have come to naught if a few unpredictable nuclear cross sections had been different from what they are by a factor of two.*
>
> - Emilio Segrè, on the making of the atomic bomb [1, pg. 1]

New science enables new technologies; these technologies bear new affordances; those affordances create risks; and those risks inform the mechanisms of mitigating them. But as Segrè's comment suggests, not all risks are equally straightforward to mitigate. As recent philosophers have noted [2], if nuclear weapons were easier to develop, or more widely available, it would be far harder to mitigate the risks posed effectively in a manner in keeping with current ethical values. In a rapidly changing scientific field, decision-makers that think they are dealing with one type of 'bomb' may find themselves quickly dealing with another. Strategies based on the old paradigm may fail to mitigate the risks of the new.

The central contention of this paper is that contemporary AI technologies, specifically those based on the well-celebrated generative pre-trained transformer (GPT) architecture (henceforth, 'AI'), are rapidly diverging from and challenging traditional assumptions about the development of AI. The first contribution of this paper is to set out this forthcoming paradigm shift; the second, to suggest strategies for mitigating the emerging risks both ethically and effectively.

What sort of shift is proposed? The established AI paradigm emphasises the correlation between AI capabilities, the attendant risks, and the immense number of computational operations required to train and run the system, or 'compute' [3, 4]. In practise, this means focusing on risks from powerful frontier models which are closed-weight, high profile, and trained and run on large, trackable compute clusters (e.g. [5, 6]). Following Giovanni Dosi's theory of technological "paradigms", [7], this paper calls this the 'Big Compute' paradigm. It permits clear mechanisms for mitigating risks from AI systems responsibly [4] which inform governance decisions from company policy (e.g [8, 9, 10]) to international legislation (e.g. [11]).

Today, however, scientific developments are diverging from and challenging these assumptions. Five interoperating technology pathways—Small models, Hidden models, Augmented models, Decentralized processes and Open-Weight models ('SHADOW')–increasingly threaten, like Segrè's cross-sections analogy, to put governance mechanisms out by a factor of two. This paper calls this the 'Proliferation' paradigm of AI, or AI proliferation for short. It describes a developing network of technologies and capabilities that is less visible to regulators, harder to control, and presents new risks for irreversible harms. By coining the term, and defining the contours of the Proliferation paradigm, this paper hopes to make a contribution to future research towards modelling and mitigating the risks of models beyond the compute-intensive frontier.

What sort of governance strategies will work in this new paradigm? Instead of focusing just on compute, governance should focus on the other aspects

of what has been termed the 'AI triad'—algorithms and informational inputs like data and capability keys—as well as securing non-traditional decentralized compute. Building ethical policy here will require serious and critical attention to the ethical trade-offs of regulation and strategies to mitigate them, as well as improved mechanisms for securing these promises. Only a clear view of the technological paradigm, an evolving picture of stakeholder values, and the mechanisms to secure policy promises will be sufficient to govern the proliferation of AI.

This paper builds its contribution across three main sections. The first sets out key concepts: the argument for governing AI risks; how AI operates in paradigms; and how scientific and technological assumptions have configured Big Compute. The second sets out the Proliferation paradigm by assessing the viability, risks and benefits of each of the SHADOW technologies, drawing attention to how they challenge (compute) governance strategies. The final section explores promising strategies for governing algorithms, decentralized compute and capability keys, paying critical attention to ethical trade-offs of governance and future research priorities. A conclusion positions these policy recommendations in a broader context of AI development, philosophy of science and existential risk.

2 Key Concepts: Risk & Governance, Paradigms & 'Big Compute'

2.1 The Risk-based Case for AI Governance

> *AI governance refers (1) descriptively to the policies, norms, laws, and institutions that shape how AI is built and deployed, and (2) normatively to the aspiration that these promote good decisions (effective, safe, inclusive, legitimate, adaptive).*
>
> - Allan Dafoe, 'AI Governance: A Research Agenda' [12]

In its broadest sense, 'AI governance' simply refers to set of operations designed to affect and improve decisions about how AI models are built and deployed across society [12]. It might refer to decisions in private companies [13], national governments [14], or international organisations [15]. It might also refer to decision frameworks designed to achieve very different goals. Governance frameworks built to maximise the capabilities of AI systems overall, for example, might be very different to those built to minimise the likelihood of an adversarial nation acquiring those capabilities. The analysis in this paper focuses on two priorities necessary for responsible governance: mitigating harms from AI systems effectively, and doing so ethically.

The possibility of severe harms from powerful AI is a well-established prospect [5, 16, 17]. For instance, highly capable models could assist in the creation and deployment of biological weapons [18, 19] or cyberattacks against organisations or critical infrastructure [20, 21]. They could also experience technical failure modes like cybersecurity vulnerabilities or misalignment [22], either at the level of individual models, or those in larger groups [23]. Even if models are used with good intentions, some applications, like automation, may create social harms in which disproportionately disempower some portions of the labour market more than others [24]. Even if they were willing, it is unlikely that corporations alone would be able to mitigate such risks [25]. Effective countermeasures are likely to require coordination from both private and public actors (e.g. [26]).

At the same time, it is crucial that both AI technologies and governance interventions function *ethically*: that is, that they strive to reflect the normative aspirations of the parties which make and experience the results of the intervention [12]. That might include implementing governance measures which prioritise certain definitions of fairness [27], certain varieties of privacy [28, 29], clear lines of accountability [30], or support certain definitions of democracy [31], especially inter-generational [32] and international perspectives [33, 34]. Finding a 'one size fits all solution' is not the goal here. Rather, the aim should be to encourage constructive debate around the values *necessarily* embedded in AI governance regimes so as to work progressively towards generally held principles.

A pervading concern is where to set the *limits* of such governance. As Zeng (2020) and others have noted, over-regulation could enable and enact the centralization and predominance of some values or actors over others, compromis-

ing broadly liberal democratic values [35, 36]. Although empirical studies might go some way towards illuminating these trade offs, AI systems present significant unknowns, and policy must necessarily make values-driven interpretations bearing inductive risks [37]. Developing an ethical framework for making these values-based decisions (for instance, involving democratic processes) is consequently also a concern. Great as the risks from AI systems are, over-regulation resulting from unwarranted interpretations based on scarce information might also bear unethical consequences. These issues will be discussed in more detail in section 4.

2.2 The 'Big Compute' Paradigm

> *Computing power, or "compute", is crucial for the development and deployment of artificial intelligence (AI) capabilities. As a result, governments and companies have started to leverage compute as a means to govern AI.*
> - Sastry et al., 'Computing Power and the Governance of Artificial Intelligence' (2024) [4]

Understanding how AI proliferation challenges established assumptions about how to mitigate the risks from AI technologies requires first understanding what those established assumptions are. A useful lens here is the 'technological paradigm', an embodied intellectual construct Giovanni Dosi defines, following Kuhn, as "a 'model' and a 'pattern' of solution of *selected* technological problems, based on *selected* principles derived from natural sciences and on *selected* material technologies" (*sic*) [7, p. 150], including the "industrial structures" (labs, firms, government institutions) associated [p. 157]. For Dosi, such paradigms were described by 'technological trajectories', intellectual and industrial progress towards the solution of paradigmatic problems "whose outer boundaries [define and] are defined by the nature of the paradigm itself" (p. 154). Well-established across various fields (e.g. [38, 39, 40]), this lens can help decision makers to appreciate technological paradigms like webs of interconnecting assumptions: grounded in evolving scientific research, and embodied not only in experimental and commodity technologies, but the corollary industries, corporations, and community practises of those researching and developing them.

Dosi's 'technological paradigm' provides a useful lens to understand AI today. Of the three facets necessary to develop powerful AI capabilities—informational inputs like data, algorithms, and compute (sometimes referred to as the 'AI triad' [41])—compute has been assigned the dominant position. This is for good reasons. The hypothesis that performing more operations in the training and inference stages of a model's operation will improve that model's performance is one of the most well-supported findings in machine learning [42, 3, 43]. This means that models involving a larger number of parameters, or those trained for longer, typically perform better than smaller models [42]. Companies capable of gathering more compute can therefore build more capable models.

This training process is critically expensive: Sastry et al. (2024) point out that "development of frontier AI systems has become increasingly synonymous with large compute budgets, access to large computing clusters, and the proficiency to leverage them effectively" [4, 44, p.9].

This bears a number of corollaries for how the AI industry has developed. Since most leading AI companies rely on colossal personal datacentres or those of high-visibility compute providers or 'hyperscalers' [45], capable models are difficult to hide.[1] Each order-of-magnitude jump in investment here is extremely expensive. Training a model on one order of magnitude more compute than GPT4 might cost billions of dollars [44]. Releasing information about such a model's architecture, weights and training data online for free without usage constraints would be financial suicide (or at the very least a huge advantage to competing companies). Beginning with the compute hypothesis, one arrives at a paradigm of powerful models which are difficult to replicate, high profile, and trained and run on large, expensive, trackable compute clusters. The spiritual successor to 'Big Data"s emphasis on scale [46], this paper refers to this paradigm as 'Big Compute'.

The 'Big Compute' paradigm in turn configures what Wirtz, Langer and Weyerer (2024) have called the regulatory "ecosystem" [47]. It bears three advantageous features for responsible risk-mitigation governance. First, it creates *moats*: huge datacenters are expensive to run and maintain, so there are relatively few actors creating frontier models, lowering the burden on regulators and ensuring against over-regulation of non-threatening actors [48]. Second, it is *anticipatory*: the amount of compute used for training runs serves as a useful proxy for capabilities, allowing policymakers to better assess which models are and are not dangerous and decide when to trigger further evaluations *before* models are released or fully developed. One such governance mechanism, "compute thresholds" [6], forms a critical part of both the US AI Executive Order [49] and the EU AI Act [11], and some AI companies deploy them as part of "responsible scaling policies" (e.g. [8]). Third, compute is *physical*: more specifically, as Sastry et al. argue in their seminal paper on compute governance, it is "detectable, excludable, and quantifiable, and produced via an extremely concentrated supply chain" [4, p.1]. This helps to ensure transparency, enable reallocation of resources, and "enforce restrictions against irresponsible or malicious AI development and usage" (p.1).

Taken as a whole, the 'Big Compute' paradigm represents a particular set of scientific hypotheses, a set of technological trajectories, a narrow set of key actors, and an advantageous paradigm for governance. It is empirically well-founded, and likely to be an important tool for structuring effective risk-mitigation governance mechanisms from large and frontier models in the future. Nonetheless, as Sastry et al. themselves note, "governance of compute is not the whole story of AI governance" [4, p. 2]. That AI will remain bound to these assumptions as the technology matures is far from guaranteed. The next section

[1] For context, a two-order-of-magnitude increase in compute, comparable to the jump between GPT2 and GPT4, would require the same power footprint as the Hoover Dam [44].

will explore some of the scientific research and technological developments that offer an supplementary paradigm to the contemporary model of 'Big Compute', and their associated risks and governance challenges.

3 The Proliferation Paradigm: 'SHADOW' Pathways

> *The success of a paradigm ... is at the start largely a promise of success discoverable in selected and still incomplete examples.*
> - Thomas Kuhn, *The Structure of Scientific Revolutions* [50, p.23]

The central contention of this paper is that recent scientific and technological developments increasingly interoperate to provide and embody alternative assumptions to those of 'Big Compute'. Taken together, one could call this the 'Proliferation' paradigm of AI, or AI proliferation for short. It proposes serious new challenges for risk-mitigation which will be explored in Section 4.

A robust technological description of this paradigm has yet to be made. To this end, this paper contributes the 'SHADOW' framework, a map of five emerging pathways—Small models, Hidden models, Augmented models, Decentralized processes and Open-Weight models ('SHADOW')—which help define the contours of the Proliferation paradigm. Each proceeding subsection sets out the technical viability of the pathway, assesses its risks and opportunities, and describes how it challenges a governance paradigm based on compute, thus setting out the landscape for risk-mitigation governance that will be explored in Section 4.

3.1 Small Models

The more compute required for models to demonstrate dangerous capabilities, the more effective governance via compute becomes. Yet such a framework might overlook small models: specifically, those which demonstrate dangerous capabilities, yet have far fewer parameters or training token requirements, requiring little compute to train, run and fine-tune. This paper defines 'small' in a relative sense:

- Fewer than large frontier labs, such that compute governance thresholds designed to catch frontier systems do not apply.
- Few enough to make (the cost of) creating, running and fine-tuning them accessible to a dramatically larger pool of people.

Technical Pathway(s)

The algorithms behind AI systems may become far more efficient over time.[2] As Brown and Hernandez demonstrated in their 2020 paper popularising the idea, "the number of floating-point operations required to train a classifier to AlexNet-level performance on ImageNet decreased by a factor of 44x between 2012 and 2019", corresponding to a doubling in efficiency "every 16 months over a period of 7 years" [52]. State of the art models will most likely shrink faster: Ho et al. conclude that language model efficiency is doubling on average every 8.4 months (albeit with a broad 95 percent confidence interval of 5.3 to 13 months) [53]. Technical strategies like quantization [54], model pruning [55], knowledge distillation [56], and parameter-efficient fine-tuning [57] have all been cited as promising avenues for diminishing the amount of compute required to train a model, or the parameters needed, for a model to achieve advanced capabilities.

Two technical pathways to smaller models stand out. One, best demonstrated by Microsoft's Phi series, trained on a data corpus of real and synthetic (GPT 3.5 generated) textbook questions, matches or outperforms models like Gemma 7B and Llama 8B on benchmarks despite possessing about half the number of parameters [58].[3] Apple's OpenELM 3B model, uses such a specialised training regime to score a 92.7 on SciQ, an undergraduate level science benchmark [60]. Secondly, improvements to algorithmic architectures have also seen models become more capability dense [61]. Models like Databricks' DBRX 132B and Artic, a 40B parameter model with 17B parameters active, use only a fraction of its parameters on any given query, improving inference efficiency [62, 63]. Such systems work by allowing models to 'search' through individual subnetworks; improving the ability of models to allocate compute strategically might allow developers to make them even smaller [64]. Deeper architectural changes, like a shift from MLPs to Kolgomorov-Arnold Networks, might further reduce the compute requirements [65]. If frontier AI models become capable of automating AI research and development and improving algorithmic efficiency (as suggested in e.g. [66]), this might create a 'slipstream effect', by which growing scale of frontier models paradoxically rapidly *decreases* the compute required to achieve dangerous capabilities.

Skepticism is necessary here. Timelines are unclear: whilst it could be that models get smaller and better quickly (see section 3.3), unknown ceilings may cause progress to plateau. Whether improving frontier models make building small models easier, or less financially attractive, and how this will affect the pathway development, is also currently unclear. Nonetheless, it seems likely that some model shrinking will occur in the short term, increasing the number of actors capable of building models with sub-frontier capabilities, a phenomenon Pilz and Heim refer to as 'diffusion' [67].

Benefits, Risks, and Governance Challenges

[2]The efficiency of an algorithm can be understood as the number of floating-point operations required to demonstrate a particular capability, where more efficient algorithms use less compute [51, p. 9].

[3]'3B' refers to 'three billion parameters', a proxy for the compute used to train the model. OpenAI's GPT3, released in 2020, possessed 175 billion parameters [59].

Small models hold real promise. As a recent paper notes, the current size of larger models makes them unsuitable for scenarios requiring "on-device processing, energy efficiency, low memory footprint, and response efficiency" [68]. Smaller models are also suitable for running on-device, protecting them from privacy issues associated with cloud compute providers, and potentially making them easier to audit [60]. They are also, perhaps most importantly, cheaper to run, suggesting that (as Nathan Lambert argues) they are well positioned to "unlock the most economic value by unlocking applications reliant on edge computing or low-costs" [69].

These benefits come alongside new risks. A downside of what Pilz, Heim & Brown call the "access affect" [70] is the increased likelihood of malicious actors gaining access to dangerous capabilities. As a Center of New American Security paper points out, within five years, at current trends, "the cost to train a model at any given level of capability decreases roughly by a factor of 1,000, or to around 0.1 percent of the original cost" [51, p. 30]. By this trajectory, training a model in 2029 to today's state of the art lies in the price range of a family car (based on cost estimates, p. 13). Assuming similar training costs per parameter, training a model like Microsoft's Phi 3 the same year would cost around $83, about the cost of a family restaurant meal. Other technological trajectories discussed in this paper, like the falling price of compute access (section 2.3), capabilities augmentation (2.4) and open-weight ready-mades (2.5), might make them even cheaper.

Dangerously powerful small models present obvious risk-mitigation challenges, especially to a governance paradigm centered on compute. A key loss is the viability of compute thresholds. Sparing intervention by "scientific panel" (51.1.b), models developed in the EU below the threshold need only be deemed safe to release by the company or individuals developing them, a decision for which there is as yet no commonly accepted standard [23]. Although proponents of compute governance suggest that these thresholds should adjust over time (e.g. [4, p. 70]), the evidence and procedures that would be necessary to do so are unclear. Even if model evaluation mechanisms improved substantially to address this threat, another challenge is the increased opportunity for regulatory evasion. Smaller models would also be easier to hide from government oversight based on compute, especially if trained on decentralized architectures (2.3). This is discussed in more detail in the next section.

3.2 Hidden Models

The more information regulators, governments and the public have about which actors are developing and deploying what sort of AI models, the better equipped they are to make responsible decisions about how best to govern AI systems. In contrast, 'hidden' models – models with dangerous capabilities which go unreported, unregistered, or unlicensed in the jurisdiction in which they are used, or which are otherwise overlooked by government building AI regulation – present a variety of risks. Whilst not a technological pathway *per se*, hidden mod-

els constitute an important and as-yet undocumented aspect of the "industrial structure" of the Proliferation paradigm [7, p. 157].

Hidden Model Pathway(s)

Models might be 'hidden' in a variety of ways: call them *below, above*, and *outside*. *Below* would concern the models that are *legitimately* not covered by ('below') compute-threshold based reporting requirements (e.g. any model trained on sub-10^{26} FLOPs in the US) discussed in Section 2.1. This is likely to become more viable as eliciting dangerous capabilities requires less compute (section 2.3 and 2.5). However, it should be noted that the more widely used the model is, the more likely it will come to the attention of regulators.

Above would concern models with dangerous capabilities which are covered by ('above') compute threshold-based regulation, but go inadequately unreported. One culprit here might be major labs. While it is unlikely a leading lab would release an untested model outright, labs can and do release models 'into the wild' covertly for testing purposes. The origins of 'gpt2', briefly the most powerful LLM on the LMSYS leaderboard, were unknown before it was revealed to be GPT4o [71]. If a powerful model was trained and deployed covertly, copied (see 'slipstream' effect, section 3.3), or leaked, a massively dangerous model might proliferate in the wild. Even if models do not go entirely unreported, nascent or low-quality reporting might still mean that developers building frontier models fail to prepare [72]. In the longer term, another might be malicious actors using decentralized compute (see section 3.4) to build powerful models that regulators cannot see.

Outside would concern models located outside a jurisdiction, thus 'hidden' from the regulation within it, but accessed via mechanisms like VPNs. The present likelihood that this would happen is unclear. On the one hand, countries like China, Saudi Arabia and the UAE have all proposed alternative governance regimes to the EU and US, open-weighting their state-led models like Qi [73], Allam [74] and Falcon [75] respectively. On the other, threatening models may be used within national borders as much as outside them, limiting the likelihood that unstable countries might wish to invest in them. At least two high risk pathways look likely here. One is that a national government uses some form of hidden AI system to perpetrate harms in a warfare context [17]. The second is that a national government with less robust evaluation procedures mistakenly ratifies the safety of a model that then causes international harm [5].

Benefits, Risks, and Governance Challenges

Government oversight imposes burdens. Reporting costs time and attention; evaluations are expensive to run and difficult to check. If capabilities can be attained faster without members of small teams having to spend their time reporting to government bureaucrats, with no risk of releasing a model that might cause harms, it stands to reason that they might protest this, especially if their reporting requirements were similar to those of an AI company many

times their size [4, p. 67]. This has been the argument of many 'pro-innovation' or 'anti-regulation' arguments for governing AI [76].

At the same time, however, a laissez-faire approach to unreported models could have serious risks (see Section 2). For instance, a model hidden *below* reporting requirements could be exposed to less strict evaluation requirements, leading to the deployment of a model with flaws, like security vulnerabilities, in critical infrastructure. A model hidden *above* requirements governments are taken by surprise by the powerful capabilities of a less-reported model, leading to societal unrest [23]. And a model hidden *outside* requirements could enable malicious actors to perpetrate serious harms either domestically or in nations with higher-stringency governance regimes [17]. In each of these cases, having the information necessary to dissolve the threat before manifests would seem to be an urgent priority.

Governing hidden models presents two main challenges. One is how to calibrate the risk sensitivity when designing evaluation regimes that might affect small actors more than large ones (see section 4). Another is how to protect against regulatory flight. In a 'Big Compute' paradigm, AI-capabilities drivers (among other things) are high profile, possess big teams, and require huge data centers. AI proliferation might reduce these requirements, making regulatory flight substantially easier. If governance standards are not global, actors with higher risk thresholds may release models that will undermine the governance regimes established in nations working to establish lower-risk thresholds (see section 4).

3.3 Augmented Models

Where Section 3.1 focused on the diminishing compute requirements of training dangerously capable models, this section focuses on the fleet of strategies for eliciting new capabilities ('capability keys') from AI systems which bypass or invalidate safeguards ('jail-break' models), without retraining. Although a related problem is mentioned by Anderljung et al. (2023) [5] as the 'unexpected capabilities problem' (Sastry et al. (2024) [4, p. 34] also note this possibility in a footnote), this section explores more up-to-date literature on these risks in more detail.

Technical Pathway(s)

Two technical pathways stand out as promising for augmenting the capabilities of existing models: advanced prompting, and precise fine-tuning. Well-known prompt-based strategies, like chain of thought reasoning [77] and prompt scaffolding [78], have been shown to have impressive effects on the quality of model outputs. Davidson et al. (2023) report "gains equivalent to training with 5 to 20x more compute at less than 1% of the cost" [79]. Working in the context window (sometimes referred to in technical circles as 'adaptive compute' [80]) looks to become more popular in future. Google's recent Gemini model can learn new languages with efficiency using only context inputs [81]. These tech-

niques, compatible with both open- and closed-source models, might proliferate online, to be used at will. In the style of some video games, one could think of these pieces of information as 'capability keys'.

A complementary possibility is fine-tuning, which has been shown to enable both greater capabilities from small models [82] and remove safeguards on existing models with ease [83]. Model merging, a related technique, allows developers to systematically combine models to create more powerful capabilities, releasing "new state-of-the-art models back to the open-source community" [84]. It is even possible to fine-tune smaller models on the output data of larger ones, leading to a bootstrapping effect which small models become more capable. As in section 3.1, this might create a slipstream effect, by which the improvement in frontier models radically raises the quality of sub-frontier-compute-level models. These model weights could thus also be thought of, in a way, as 'capability keys': though on a different scale to prompt-based attacks, and only affecting models with open fine-tuning interfaces like open-weight models (section 3.5). Both strategies—prompt-based and fine-tuning—might also be used together.

A catalysing factor here is the emergence of powerful supporting infrastructure for sharing information about augmenting AI model capabilities. Open-source AI communities HuggingFace, LAION, or Eleuther AI proudly support the proliferation of state-of-the-art machine learning techniques. Semi-automated systems, like Cohere's Command R+ [85] and Github's in-platform Copilot [86] aim to accelerate coding and bring capabilities within the grasp of less experienced researchers. Microsoft's AutoGen already designs multiagent systems automatically [87, 88], as does Evolutionary Model Merge for model merging [84]. If models were to become substantially better at AI R&D (as highlighted in DeepMind's recent frontier safety framework [89]), these capabilities might proliferate more widely still.

Benefits, Risks, and Governance Challenges

Eliciting more powerful capabilities from existing models has obvious benefits. Prompting strategies can make outputs more robust, allowing smaller models to be scaffolded in ways that make them more effective agents [90]. Fine-tuning allow models to be cheaply and efficiently adapted for specific use cases, and has been shown to protect against certain failure modes [91]. Infrastructure that helps these techniques proliferate in society more broadly has advantages for decentralizing power over model capabilities [92]; it might also disproportionately empower smaller industry actors [93], and help to create broader bases for AI talent [76].

However, model augmentation brings hard-to-forsee risks that Anderljung et al. refer to as the 'unexpected capabilities problem' [5, p. 12]. If models quickly became extremely powerful due to a slipstream effect, the best case scenario might be societal disruption. Alternatively, if prompt scaffolding compounded capabilities, but created cybersecurity vulnerabilities, it could compromise critical infrastructure. More worryingly, present adversarial prompting [94] or fine-tuning [83] can bypass safeguards with ease, uplifting malicious actor

capabilities. A paradigm in which many older unsecured models could be upgraded to possess dangerous capabilities would be an immense and challenging regulatory target, especially if many were open-weighted (section 3.5). How to secure capability keys—which might proliferate online over platforms like HuggingFace or EleutherAI (see Section 3.5)—should be a governance priority. If augmentable models are distributed open-source for actors to download them (section 3.5), or difficult to track (3.2), it would be especially difficult for auditing parties to spot these issues and address them, and for governments to evaluate their spread and impact (see Section 4).

3.4 Decentralized Processes

'Big Compute' generally involves a few high-profile 'hyperscalers', working with a few high-profile companies. This section focuses on alternative pathways by which AI models might access the necessary compute, either by accessing decentralized compute providers, or by operating as a decentralized service across many devices in real time. Although decentralization has been widely discussed in a governance context around blockchain, applauded in AI circles [95, 96], and briefly featured in a 2020 AI review [97], no up-to-date governance-facing papers on the technology existed the time of writing.

Technical Pathway(s)

There are two ways decentralized processes might transform AI. The first involves training a centralised model on a decentralized compute platform; the second, training a decentralized model on lots of individual sites. To understand the first, consider Akash, a "decentralized compute marketplace" or "supercloud" that allows users to buy compute from providers at low costs [98].[4] The system already runs small in-context learning models (Llama 7B), and the team have trained at least one large language model ("Thumper") entirely on the distributed architecture [98]. Other similar examples include Golem [100], Brain Chain [101] and iExec [102]. Other frameworks like Gensyn go further still, implementing an automated payment system over blockchain that allows "direct and immediate rewards for those contributing their computational resources for machine learning tasks" [103]. If companies could unlock the 'latent compute' already in existence and make it public cheaply, Aksh Garg (a decentralized compute researcher) points out, "even tapping into a tiny fraction of this widely distributed network of computation would be game-changing" [104].

A second stream might come from distributed models themselves. Douillard et al. (2024) argue that models trained on modular tasks from discrete sites can be more efficient: training a model on the C4 dataset, with paths "of size 150 million parameters", the team was able to match "the performance in terms of validation perplexity of a 1.3 billion model, but with 45% less wall clock training time", which they hope will enable more energy and compute efficient scaling

[4] For a technical review of how decentralized platforms work, see [99].

(see Section 2.1) [105]. Meanwhile, papers like LinguaLinked promise models that will be able to be run (perform in-context learning) using the compute from several devices simultaneously [106]. As models become smaller (Section 2.1) or more efficient (2.3), this might be an increasingly viable strategy for eliciting greater capabilities from models without relying on traditional cloud compute providers.

At present, decentralized processes do not play a significant role in AI. Decentralized compute networks are very small (Akash offers access to 85 A100 GPUs; a mid-tier AI company like Stability AI might own 5,400). They are also hamstrung by technical limitation associated with running a model on a virtual machine rather than a data center co-located in the same place [104]. Research into decentralized processes like DiPaCo is still in the theoretical stages. Nonetheless, the amount of investment in the industry is substantial. DeepMind's investment in distributed path decomposition is illustrative of their confidence that they will be able to create an AI paradigm in which individual "paths, trained on any available hardware type, communicate infrequently across the world, exchanging useful information and enabling new forms of composition" [105]. If progress on this frontier continues, perhaps accelerated by the other SHADOW technologies, the risks should be taken seriously.

Benefits, Risks, and Governance Challenges

Proponents of decentralized compute networks point to benefits like "lower cost and greater choice", "access standardisation", "community driven" governance regimes and income streams for smaller actors [98]. They argue that centralized compute provisions push for longer contracts, enforce unsuitable UX, and drive up prices [103]. Distributed models might work better across decentralized compute: they also promise to bring down computing costs, and leverage existing compute more efficiently, helping to make capabilities more widely available [104].

Two potential risks of a paradigm built around decentralized computing stand out. The first is that this reduces the transparency of model development, leading to sub-optimal policy or harms from unsafe systems (see section 2.2). Yet building pro-transparency policy is challenging, too. Companies like Akash, vocal in their criticism of the "now-largely-discredited... cautious approach to technological progress", note that "a demand exists for access to open-source models in a permissionless environment" where "anyone can run these models without unnecessary restrictions" [98]. It is unlikely that they would enact pro-transparency Know-Your-Customer policies like traditional cloud computing providers, even if it was technically feasible [48]. Even then, their privacy-minded client base (both buyers and sellers) might just as easily find another marketplace elsewhere.

The second, more concerning risk, is that decentralized compute might constitute a resource for malicious actors to access compute to fine-tune or build small models for offensive purposes (see sections 2.1 and 2.3) without raising the suspicion of authorities (see section 2.2). If decentralized compute networks

were powerful enough, they might prove a way for malicious actors to get around on-chip governance measures such as location tracking, in effect skirting anti-chip-smuggling policy regimes. Even if actors are not malicious, they might make it more difficult to enforce policy regimes than traditional physical compute, or large corporations (the decentralized computing provider might not be based in the same jurisdiction as the model is trained, for instance). Distributed systems make enforcement even more difficult: enforcing a model trained on a system like like DiPaCo would compel action against 256 geographically distributed entities in the case of reluctant or malicious actors. Like the capability keys of Section 3.3, once the information necessary to build such systems is publicized, it may be impossible to contain.

3.5 Open-Weight Models

Copying, downloading or using (components of) models that are freely available online uses far less compute than training a model from scratch. The more of these components that are freely available (which might include pre-training data, fine-tuning data, alignment data, evaluation frameworks, model architecture, weights, and their respective implementation code [76]), the more a model can be described as 'open-source'. Since some facets of the training process (eg. the inference code) cost far less compute to generate than others, this paper follows Seger et al. in focusing most on open-weight models "for which at least model architecture and trained weights are publicly available" [107, p. 2].

Technical Pathway(s)
When Seger et al. (2023) wrote their seminal piece on the risks and benefits of open-sourcing highly capable models, there was yet to be an open-source model with capabilities close to the frontier. Today, open-weight models are a significant and competitive part of the AI ecosystem. They routinely beat frontier models like OpenAI's GPT4 on the LMSYS arena [108], both in narrow domains (e.g. Cohere's Command R +, [85]) and more generally (e.g. Alibaba's Qwen 1.5 72B [73]). Popular models like Meta's Llama 3 70B (April 2024) outperform or tie across multiple benchmarks with closed-source models produced by larger companies only a few months before (e.g. Google's Gemini 1.5, February 2024 or Claude 3 Sonnet, March 2024) [109]. Looking towards Facebook's Llama 405B, it is possible that around the time of this paper's publication with see a frontier-leading model be open-weighted for the first time [110].

Open-weighting is an increasingly popular and well-consolidated practise. Huggingface, a platform for downloading and sharing open-source models, supports over 350k models today [111]. Other platforms include Open-LAION, Red Pyjama, and Eleuther AI. In terms of big players, Meta has committed to open-weighting their frontier models (with some constraints) for the foreseeable future [110], and Open-AI has shown a pattern of open-sourcing their models roughly a year after they are first deployed. At the smaller end, Apple's OpenELM series is one of the most extensively open-sourced models currently available, including training datasets with their model reports [60]. In lieu of responsible

scaling policies with stipulations around open sourcing, there is nothing to stop companies from open-sourcing increasingly large and powerful models.

Benefits, Risks, and Governance Challenges

Providing the weights of capable models to the public for free has obvious and significant advantages. One concerns the progress of AI systems themselves. As Eiras et al. note in a recent paper (supported by Meta), open-weight probably increases the number of people capable of working in frontier research and contributing to AI development more generally, potentially reduces the need for retraining models at personal expense (thereby being comparatively cheap), and potentially empowers developers to work on architectures that they might not otherwise be able to [76]. This might also lead to the recognition of bugs, improving safety. A correlative argument is that open-weighting "democratizes AI", giving "more people influence over how AI is developed and used, and promoting the representation of more diverse interests and needs in the direction of the field" [107, p. 10].

At the same time, open-weight models introduce and exacerbate a number of risks [107, Section 3]. First, while closed source models themselves are not currently as well secured as would be desirable, open-weight models have been shown to be extremely vulnerable to adversarial detuning [83], allowing malicious actors to both disable safeguards and elicit dangerous capabilities [112]. Even if models were closed thereafter, making understanding of model internals more broadly available might give malicious actors a greater chance of finding exploits that work around safeguards. Second, while it some bugs may be relatively easy to fix, others might be neither easy to spot nor fix, leading to vulnerabilities proliferation. Even if patches are built, it is difficult to ensure that these will actually be applied. It can also be difficult to assign liability to such harms, making it harder to incentivise actors against committing them [112].

Such risks operate in tandem with open-source infrastructure. A malicious actor could exploit the resources on HuggingFace, for instance, by downloading an open-weight model and fine-tuning it on an offensive dataset by following the instructions online (see section 3.3). This did in fact happen in 2023 with GPT4chan, a model trained on racist hate speech which was downloaded nearly 1,500 times before it was taken down [113]. Such platforms can make it harder to defend powerful models against jail-breaking or misuse. Although platforms like HuggingFace have already made a number of statements on ethics and policy [114], it is relatively easy to find posts on getting around safeguards on datasets and models (see Section 4.3). Similar sites might remain zones for building dangerous models, and sharing means to elicit dangerous capabilities. Securing against this possibility is difficult: were HuggingFace to tighten their reporting requirements, users might always move elsewhere.

3.6 Governing the Proliferation Paradigm: Principles

The preceding subsections have sought to set out the contours of the Proliferation paradigm. Although the word 'proliferation' aims to emphasise the dissemination of capabilities across actors and models that characterises most of the increase in risk, other words might also describe it. 'Sub-frontier', 'underbelly', 'distributed' or simply 'emerging' all capture important features. A set of overlapping hypotheses about the future, the paradigm is necessarily uncertain, new technological pathways may emerge, and lines dividing the aforementioned technologies may alter over time. 'SHADOW' is not the final word on the AI proliferation, but a tool for beginning a conversation.

All SHADOW technologies weaken the assumptions of the 'Big Compute' paradigm, including governance strategies like compute governance. This creates three main disadvantages for governance mechanisms aiming to mitigate risks.

First, it substantially reduces capability *visibility*. Though it might seem like open-weighting models would increase visibility, doing so publicly means that it is very hard to keep track of which other actors might have access to those capabilities. Similarly, the lower the requirements for accessing sufficient compute, or capability keys necessary to elicit dangerous capabilities, the easier it will be to perpetrate harms. Paradoxically, the openness of benevolent or neutral actors enables secrecy to thrive.

Second, the massively increased target site of the Proliferation paradigm would make policy far more difficult to *enforce*. One axis here is model relevance. In an augmented-model paradigm, old models can be augmented to create powerful capabilities, just as compute sold to hidden actors years ago or latent in smart phones can be used to train dangerous systems. For models, the analogy of a 'frontier' serves poorly; better to think of a 'high water mark', with models below the surface remaining potentially (and dangerously) in play. The second axis is jurisdictional. Hidden models may operate across jurisdictions; information about capabilities augmentation or model weights is inherently international. Whilst *compute* might have a centralised supply chain [4], *capabilities* might not. Recognising and coordinating these is necessarily a global project, that will need to take into account global perspectives and various jurisdictions.

Third, in a Proliferation paradigm, actions may be *non-reversible*. Whereas a company can release access to a closed-source model and then restrict it, in open-weighting, as Seger et al. (2023) point out, "there are no take-backs" [107]. The same goes for publishing capability keys online, or sharing highly efficient architectures. While threats of 'regime lock-in' around AI governance should be taken seriously [115], the term might apply equally to the technological paradigm these new developments create, for better or for worse.

Such technological affordances pose daunting challenges for risk-mitigation. The next section proposes three sets of strategies for meeting them effectively, foregrounding an ethical perspective, and making clear where more research will be required.

4 The Proliferation Paradigm: Towards Responsible Governance

The reader should now have an understanding of the Proliferation paradigm and the challenges it presents to the assumptions of 'Big Compute'. This raises the question: if not compute, what aspects of the AI ecosystem should be governed to mitigate risks from AI risks effectively? How should these governance mechanisms be calibrated to ensure that this respects the values and principles of a broadly liberal, democratic society?

To answer this question, this paper returns to the concept of the 'AI triad' mentioned in Section 2.2 [41]. If not 'Big Compute', it argues, there are three main components of AI that governance can usefully target: *algorithms* (the weights and architectures required to build dangerously capable systems), *decentralized compute* (the facilities for actors to access compute anonymously or bypass regulation) and *dangerous inputs* (the information required to elicit dangerous capabilities without additional compute, or bypass safeguards in powerful models). In each domain, a subsection sets out the key ethical trade-offs risk-mitigation strategies have to face, sets out promising strategies for mitigating them, and clearly stipulating the directions required for further research.

4.1 Governing Algorithms: Towards Responsible Access Policies

A prevalent concern about both AI technologies and AI governance is that it might limit the number of actors capable of building AI models, keeping power centralised in the hands of a few [36, 31, 76]. This has attracted criticism from contemporary ethico-political philosophers. In 'Toward a Theory of Justice for Artificial Intelligence', for example, Iason Gabriel argues (following Rawls) that any just development of AI as a component of a broader sociotechnical social system should emphasise "fair equality of opportunity" [116, p. 224].

Under a Rawlsian framework of distributive justice, Gabriel might argue, the proliferation of algorithmic technologies that characterise AI proliferation are preferable to those of 'Big Compute'. As Section 3 showed, small models may be easier for smaller companies and even individuals to train and run how they wish; decentralized compute may better support individual actors than traditional cloud provisions; and freely-published augmented and open-weight models provide capabilities to all, particularly those parties that would not be able to train models themselves [76]. As Seger et al. (2023) point out in their useful study of the topic, proponents of technologies like open-sourcing can, and frequently do, refer to such a broadening of the pool of people capable of building models is akin to 'democratising' AI [107].

A fundamental misunderstanding is worth addressing before the virtues of this argument is addressed. Although Seger et al.'s piece on the use of 'democratisation' is a valuable review of the literature, it might permit oth-

ers to be careless in the use of of the word 'democracy'. 'Democracy' derives laudatory weight from its association with a decision process undertaken by groups of people, not the distribution of a population's access to tools. One can talk meaningfully about 'democratising' an AI company board, or 'democratisation' of AI that is used to support democratic governance processes (in the same way that one might talk of 'militarisation' of AI) but not the tools it provides. Tools can be made available to larger populations in lieu or in spite of democratic processes those populations have access to. Similarly populations abiding by democratic processes can, and may well choose to, restrict access to powerful tools, especially if they were worried about technologies undermining democratic processes (see [117, 118, 119]).

The extent to which the AI proliferation would increase the number of people in a population able to create capable models should not be pushed too far either. As of 2022, only 60% of the world's population had access to internet [120]; perhaps only as many as 28 million identify as software developers (about the population of Madagascar) [121]. Open-source communities are also not greatly diverse themselves: studies show that they are predominantly white, university educated, and male at present [122]. Third, the marginal benefit to disempowered communities may be slim: as Seger et al. point out, AI models used (for instance) in drug discovery may still benefit a vast majority, whether or not the users themselves are capable of operating the model [31, p. 28]. At the same time, as Sections 2.1 and 3 noted, these systems also bear serious risks which may well effect many. Some of these, like bioweapons, may distribute their threats more evenly; others, like cyberphising, scam and disinformation attacks, might actively target less empowered groups [17]. A Rawlsian like Gabriel may, in other words, choose to centralise control of AI in trusted parties, and focus on distributing the benefits through economic policies like taxation and universal basic income [123].

At the same time, however, it surely makes sense to broaden the pool of those capable of building models *if this can be done safely*, and without empowering malicious actors. This seems to pose a tension one might think of as an 'access-security trade-off'. How to govern the publication of algorithmic information so as to provide as many capabilities as possible to agents and services which will create benefits, whilst restricting as many capabilities as possible from actors and practises which will cause harms? A key tool here is 'structured access', a feature of AI model interfaces which would allow "access to the tool, without giving them enough information to create a modified version" or learn how to bypass crucial safeguards, analogous "to how a keycard grants access to certain rooms of a building" [124, p. 47].

Two recent papers constitute state-of-the-art in structured access. One, by Irene Solaiman, envisages five 'levels' of model release, from "fully open" to "fully closed" [125, p. 4]. The other, by Bucknall and Trager, looks at the variables of AI access (metadata, inspection, modification, fine-tuning and sampling) that successive layers of an API (Application Programming Interface) might consist of. Bucknall and Trager's paper then goes a step further, arguing that "insufficient access to models frequently limits research, but that the access

required varies greatly depending on the specific research area" [126, p. 1]. Doing so offers a valuable contribution to work that aims to calibrate access to model internals responsibly. In the style of 'Responsible Scaling Policies' [127], one could term these 'Responsible Access Policies'.

What questions would need to be answered in order to calibrate such access policies responsibly? First, policymakers need a more general view of the marginal uplift in benevolent actor capabilities *per se* of varying degrees of model access. More research might conclude that having access to a model weights "significantly aids developers in creating models that are high-performing and specifically tailored to their use-case" [76, p. 18] compared to no access, such that open-weighting has significant economic benefits to a number of actors compared to closed. Though it seems unlikely, it might also be that open-weighting "particularly helps cater to less well-resourced languages, domains, and downstream tasks" (p. 18) compared to closed-models operated by highly-paid teams aiming to do just this [81]. Cynically, one might agree with Pilz, Heim and Brown's suspicion that the "leader advantage" of large models is likely to restrict beneficent outputs to relatively few big leaders, even as capabilities proliferate [67]. Yet empirical studies as to the net marginal capabilities of actors using open-source as opposed to closed-source technologies, or economic projections of how an open-source might augment the long-term distribution of wealth in an economy, should still be a priority. Overlooking rigorous empirical research risks compromising the integrity of movements like open-sourcing, and invites over-regulation.

Second, policymakers need a more general view of the net marginal uplift in *malicious* actor capabilities, under different technological assumptions. Some threats may be scaled out (be low *net*): for instance, if a model could be shown to make cyberattacks easier to perpetrate, but also easier to defend, then it might represent less of a risk to open-weight [128]. Some threats might not be marginal: releasing access to model weights might make it substantially easier to perform valuable research and to clone the model without safeguards, whereas releasing access to fine-tuning might make it easier to perform valuable research, but without the same downsides. Finally, some threats might be different under different assumptions: if a powerful model is large, but malicious actors could easily gather the compute required to detune it anonymously, or it is relatively easy to augment, then one should be more cautious about open-weighting it. The feasibility of addressing risks should be kept in mind throughout: risks that are easier to defend against, like societal disruption, might be addressed anticipatorily through social policy [129].

In order to responsibly calibrate structured access policies, policymakers need a view of the net marginal uplifts in both malicious and benevolent actor capabilities under different technological assumptions before they can make any meaningful decisions. The work required, even to inform rough estimates, is considerable: Bucknall and Trager's paper should set the standard for papers which perform new empirical research and summarise existing research and expert opinions for policymakers to map this rapidly evolving landscape. An 'optimal' calibration here should not just look for diminishing returns (that is,

the point at which increasing the amount of access fails to increase capacity to do beneficial work but creates new risks) but a dynamic process which involves a wide variety of stakeholders, particularly those representing no-internet or no-code communities, to communicate about and converge on the value structures that they would wish to embody [130]. More guidance on how to elicit these values is suggested in the next section.

4.2 Governing Decentralized Compute: Towards Privacy-Preserving Oversight

A common refrain in AI ethics literature is that AI technologies and AI governance should respect individual privacy; specifically, individual's right to access and build AI systems without government oversight [36]. Pursuing Rawl's "concern with the ability of citizens to pursue a conception of the good life that is free from unwarranted interference", for instance, Iason Gabriel cites Andrei Marmor's argument that this should include the right to privacy on the grounds that it is intimately connected to well-being [116, p. 224]. Such a right, Marmor argues, is "violated when somebody manipulates, without adequate justification, the relevant environment in ways that significantly diminish your ability to control what aspects of yourself you reveal to others" (ibid).

As Section 1.2 mentioned, the 'Big Compute' paradigm provided a number of advantages for governance mechanisms that would defend the privacy of actors aspiring to access compute. The set of relevant companies (or 'hyperscalers') capable of providing or affording sufficient compute to elicit dangerous capabilities was assumed to be relatively small and high profile. A Know-Your-Customer scheme of the sort proposed by Heim and Egan (2024) in the style of financial markets was consequently unlikely to impose a regulatory burden on smaller actors [48, p. 7]. As Sastry et al. (2024) note, compute has further potential to be a uniquely impersonal site for governance [4]. Using privacy-preserving practises like workload monitoring, regulation need not require any information as to the function the compute would be used for, thereby protecting confidential company strategies or IP [4, p. 86].

A proliferation paradigm would upend many of these assumptions. The set of relevant actors is vast, including not only many companies and jurisdictions but potentially individuals; these individuals might be unwilling, or unable, to comply with the transparency burdens required of larger companies; and privacy-preserving oversight measures like work-load monitoring might not translate well to a paradigm of smaller models. At the same time, entering a Proliferation paradigm does not diminish the importance of regulating compute: if anything, by lending greater importance to small models, it increases it. Small models can be trained on widely available quantities of compute, and open-weighted models can be downloaded and adjustment on decentralized platforms. At the same time, as Sastry et al. suggest, "increasing visibility into AI-relevant computation could carry significant risks to privacy and civil liberties" [4, p. 52]. The question then is what level of privacy should be protected, for which actors, under what technological and societal circumstances?

One way to do this, following the previous section, would be to perform empirical research to estimate the net marginal uplift in malicious and benevolent actor capabilities from various levels of *compute* access, and use it to inform regulatory policy. Empirically founded views here will be crucial to ensure that this does not become standard practise for malicious actors, a 'black market for compute'. Keeping track of these datapoints also seems valuable in general: there may be models in the future that would be reasonably safe assuming very widely available levels of compute (e.g. desk top computers) but quickly become unsafe when that level is raised by several orders of magnitude (e.g. by using a compute cluster accessed off a compute market to adversarially detune them). On the other hand, if empirical research showed that there was no real uplift in risks from the level of compute generally provided on decentralized platforms (as may be the case at present) then these platforms should be protected from overregulation.

It seems likely, however, that the Proliferation paradigm would involve decentralized compute platforms which *are* capable of creating some net uplift in risks. One way to regulate here would be to follow Heim and Egan's model of 'Big Compute' in setting thresholds for the ability of unregistered anonymous users to interact with the system [48]. For instance, actors can read posts on HuggingFace, but they cannot post anything themselves or download model weights without creating an account on the platform. Gensyn might, under certain circumstances, implement a similar policy of responsible access: actors could use the account to obtain a certain amount of compute, but beyond a certain threshold, would have to give up personal information or indicate the purpose of their usage. Those resisting these policies or failing to update their thresholds in line with changing technological environments might be shut down, following Sastry et al. (2024), in a manner akin to how "digital services are shut down for legal violations, such as hosting illegal online drug markets" [4, p. 57].

This threshold would and should be highly contested. Five research questions can help policymakers estimate the ethical trade-offs. The first would be to ask who uses decentralized compute. To what extent are decentralized marketplaces used, or vulnerable to being used, by international actors with VPNs seeking to evade either national AI compute restrictions, or restrictions on the export of chips to that area? Second would be to ask how users of decentralized market places do in fact use them. Are they often used to train AI systems, or for other purposes (building video game simulation graphics, training climate simulation models)? Third, how much do such actors value the anonymity decentralized marketplaces provide: are they, for instance, attracted by cheaper price, or short-term contracts instead? Fourth, what sort of privacy do these actors require: are clients willing to share their name and details, but not the purpose of their compute usage, or vice versa?

Fifth, assuming that these parties do value privacy, how persuasive are their reasons for doing so? As Helen Nissenbaum has argued, privacy norms are greatly informed by social practises: norms in one field are often translated into other, particularly into contemporary digital spheres, without a rigorous exposition of the purpose of these practises [131]. To be clear, these reasons

for privacy might be hugely important. If so, this research should usefully bolster the tendency to impose less rigorous anti-privacy regulation. On the other hand, other reasons to seek privacy might be less convincing ('the NSA is spying on me') or tractable through other means ('Cloud compute know-your-customer requirements are far too tiresome for me to fill out'). Graphing and interpolating these unknowns could not only help third party marketplaces understand the needs of their consumers, but help to estimate the actual harm of enforcing anti-privacy measures on these groups.

Of course, if the compute requirements to elicit dangerous capabilities were more minimal, but still within the domain of third-party compute providers, then the gradated access policy might not be secure enough, nor basic Know-Your-Customer policies sufficiently robust against malicious actors. Research into privacy preserving monitoring strategies, such as those proposed hypothetically by Sastry et al. [4], might then be a key priority. If these were difficult to implement, mitigating the number of decentralized markets [132], the number of AI chips [133], or implementing on-chip workload tracking [134] might all be tractable suggestions for preserving safety, though the significance level for empirical evidence for the existence of such threats (as opposed to e.g. regulatory capture by large AI firms) might have to be considerable.

There remains, however, the 'Vulnerable World' [2] scenario in which AI capabilities are so easy to elicit using augmented models, or small models, or even large models operated across distributed architectures, that any sort of governance faces the colossal and impossible task of regulating almost all devices with the level of compute in a smart phone. Under such scenarios, among others, the role of scientific information might be critical. This is the topic of the next section.

4.3 Governing Ideas: Towards Responsible Information Security

Some scholars have contended that, in an ideal world, many people would have access not only to algorithms and the compute necessary to operate them, but to the information to design and elicit capabilities from them. In 'A Human Rights-Based Approach to Responsible AI' (2022), for instance, Prabhakaranan et al. argue that everyone has a right to "share in scientific advancement", where "science" describes "(1) knowledge, (2) the application of that knowledge, and (3) the method of the knowledge production." [135, p.9], citing citing the Universal Human Rights Declaration (UHRD, Article 27). The right to share scientific information under freedom of speech (as protected, for instance, by the First Amendment) is debated by a number of legal scholars in the US [136, 137] and European [138].

There are cases where increasing public access to information about AI systems should be a priority. Patches resolving model bugs should proliferate; information about how to make models cheaper or more energy efficient might help to lower the customer and environmental costs. But clearly not *all* information shared about AI systems is beneficial. One wouldn't want nuclear

capability keys shared widely; similarly, model weights might deliver powerful capabilities to malicious actors; capability keys circulated online technique shared publicly compromise the safety of a powerful model.[5] Taken literally, the UHRD is absurd: companies routinely patent and protect scientific IP, require workers in critical labs to sign NDAs, and designate issues as confidential. Yet the underlying challenge remains: what steps should individuals, platforms and research communities take to create principles for sharing scientific information pertaining to potentially dangerous AI capabilities that maximise the benefits from this communication, whilst minimising the risks?

The first step towards responsible information security requires identifying information with potential hazards. This bears strong resemblances to "infohazards" described in biotechnological literature (e.g. [139]). For AI, these might fall into two categories of 'inputs'. The first can be thought of as *jailbreaks* which bypass safeguards. This might include prompt-based jail-breaking strategies or anti-safeguard synthetic data for adversarial fine tuning (see 3.3). The second could be thought of as *capability keys* which elicit unexpected capabilities from models without involving substantially more compute. This might include the weights of dangerously capable models (see 3.5), powerful pruning techniques (3.1), powerful fine-tuning datasets or prompt-scaffolding arrangements for agentic models (3.3), or efficient distributed architectures (3.4).

While such evaluation might usefully borrow methodology from responsible access policies around algorithms, it is exceptionally difficult to ensure that there will be no unlocks or capabilty keys for LLMs. LLM outputs can be highly unpredictable [140], and the science for evaluating them is still nascent [23], making it possible that vulnerabilities will be missed. Moreover, information security faces the substantially greater challenge of having to encapsulate many axes of communication, rather than just one (model release). These challenges can be broken down into two sorts: *closed-group* policy related to groups like labs and companies developing new technologies, and *open-group* policy related to platforms like HuggingFace and LAION sharing this information.

Closed-group information communication poses one set of challenges. Workforces are constantly evolving; the information flow needing to be evaluated constantly is great; individual reputation might be involved with decisions about whether to share a new capability unlock. As a recent RAND paper on securing model weights observed of an AI company, key stakeholders may also disagree over key variables [141, p. 32]. It may be that substantial empirical research is required before stakeholders can reach agreements here. Given the fast nature of AI development, this may not be feasible. Yet it is precisely this that makes developing and articulating responsible information communication policies a priority. In a landscape of many unknowns, having robust norms and established practises for identifying which information is and isn't harmful, and how it should be communicated, is all the more crucial.

Take, as a case study, the 2022 blog post setting out the 'infohazard policy' of

[5] Some of these risks might be able to solved by in a centralized manner (e.g. patching a centralised small model against jailbreaks); others might not, and if models are open-weight and downloadable, patches might proliferate, creating irreversible risks.

AI company Conjecture. Despite setting out how various secrecy levels should relate to each other, the policy framework fails to describe how they should relate to technologies, a conversation which might benefit other actors more significantly. It offers no concrete analysis of why certain information falls into which bucket beyond the discretion of the 'appointed infohazard coordinator' [142]. This seems unlikely to scale effectively, and proposes no accountability on those infohazard coordinators in the event of an error. Yet given the lack of precedent or common practise here, the naivety is understandable. No other companies have released explicit infohazard policies, or commitments to mitigate infohazards. Though these are unlikely to be sufficiently rigorous, making commitments to set out explicit infohazard policies would be a start, and allow third parties and the public to compare the stringency of individual companies, how companies accountable for any leaks, and track how these policies evolved in line with responsible scaling, access, and security policies. Frameworks like the National Institute of Standards and Technology's information security framework [143], or the EU's General Data Protection Regulation required Data Protection Impact Assessments [144] might provide initial starting points.

A prevailing concern here however is how to secure these models weights once these striations are decided. In a recent report setting out approximately 38 attack vectors from different capabilities of malicious actors, Nevo et al. (2024) note that securing model weights against the most capable actors (i.e. highly motivated states) is currently not possible, and there is substantial debate about how to achieve security against very capable actors (i.e. state-sponsored groups) [141]. Securing weights may become harder as frontier models develop more powerful cybercapabilities [128]; information about capability keys might be far easier to elicit through Nevo et al.'s proposed attack vectors, especially if it isn't immediately identifiable as sensitive.

Open-group communication norms present a subtly different set of challenges. A key risk here is that a neutral or malicious actor shares weights or jailbreaks on a public site, which then proliferate. On the one hand, machine learning information platforms like EleutherAI Discord, HuggingFace, EA Forum, LessWrong, Alignment Forum and even more general platforms like Facebook and X might be usefully served by modelling infohazard policies on those of AI companies (Meta and X could use policies already developed in-house, for instance) to identify and take down content before it proliferates online. On the other hand, creating an information security regime that could realistically account for these practises is a serious challenge. Content moderation on large platforms has been viciously contested, reluctantly deployed, and ineffective [145, 146]. Content policy already in action deserves further examination, but clearly displays flaws. For instance, HuggingFace suggests that it will downgrade, privatise or disable the visibility of ML artefacts on its platform, and asserts that it does not tolerate "code that is designed to disrupt, damage or gain unauthorized access to a computer system or device" or "proxies that are primarily designed to bypass restrictions imposed by the original service provider", among other technical restrictions [114]. In practise, however, it is trivially easy to find

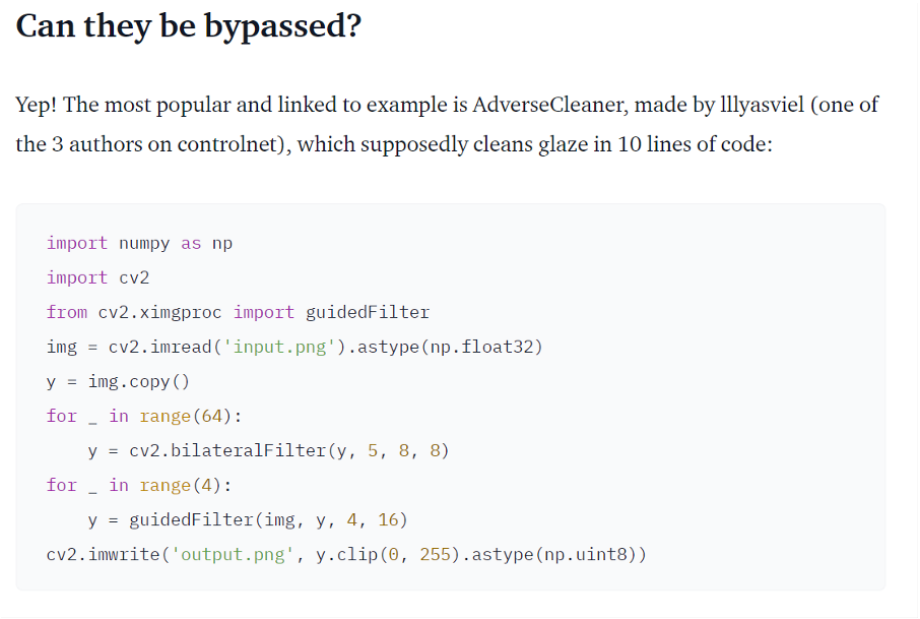

Figure 1: It took about 2 minutes to find a HuggingFace post describing how to bypass Glaze, a measure designed to protect art against AI mimicry. [147]

material that skirts the lines of these restrictions (figure 1).

As with decentralized compute, governance actors may struggle to contend with established norms of free information sharing developed in one paradigm which translate poorly to proliferation. Even then, analogies from biotechnology suggest that content like papers may be published with dangerous dual-use capabilities [148, 149, 139], even after the infohazard awareness in that field was relatively mature [150]. Such policy might still be worthwhile: in information epidemiology, restricting the transmission of information from one site to another even slightly (say, each person telling 1 other rather than 5) can have significant compounding effects. Mitigating the spread of a leak – even just to give governance mitigations days or hours of head start – can mean the difference between deploying patches and serious harms.

The bottom line here is that an AI paradigm in which powerful models are jailbroken or elicited by capability keys would be extremely difficult to govern responsibly, even assuming advanced information policy and security. Once unsecured, this information is likely to proliferate, potentially creating irreversible risks. The low feasibility of this governance approach to mitigating harms would seem a significant reason to be cautious about model releases, especially open-weight models, and to prioritise model robustness against jailbreaks, for instance via adversarial testing [151]. At the extreme end, specific research objectives with clear dual use capabilities to avoid governance, like Deepmind's DiPaCo research stream, should be carefully evaluated as to the extent they might en-

able malicious actors before research is made public, or should be proposed only alongside strategies for governing the risks. Sometimes the most secure way to deal with a possible information hazard is not to put funds towards developing it at all.

4.4 Governing Proliferation: Key Principles

Each governance strategy—responsible access policies, privacy-preserving oversight, and strong information security measures—fundamentally requires three things. First, an empirically rigorous basis for estimating the net marginal uplift in malicious and benevolent actor capabilities from different levels of access (to compute, algorithms or information) under different technological assumptions. Second, a well-substantiated critical literature reflecting stakeholder values that can be used to navigate the necessary trade-offs, even as the technology emerges. The third requirement is security. Unless sensitive information like model weights and capability keys can be effectively secured, governance promises (like responsible access policies) will be futile. Future research in all three domains is necessary and urgent if responsible governance strategies are to develop at the speed of developing technology.

Two further points might usefully orientate future research. First, just as policy designed for one paradigm might not be suitable for another, ethical norms that earned respect in one technological paradigm shouldn't be assumed to hold value in another. Free communication strategies designed around harmless scientific factoids cannot apply to capability keys or jail-breaks, nor privacy ethics for agents training AI models which could hack private infrastructure. Two implications stand out. If decision makers wish to champion values that acquired evaluative weight before AI proliferation, they should do so conditional on recognition of their cost. Alternatively, if some values were assumed to be absolute, then decision makers should be careful about funding particular technology that enables behaviours antithetical to those values. Further research into how the risk landscape affected by technological developments informs, and alters, ethical norms will be necessary to avoid undesirable ethical drift.

Second, even given empirically-driven estimates for relevant uplifts, and a clear view of societal values, policymakers still face the problem of inductive risk [37]. Should models be assumed safe until proven unsafe (loosely speaking, the 'accelerationist' position), or vice versa? A useful heuristic here is to consider the lens of *regime lock-in*, and operate under the principle of 'accelerate when actions are reversible, and deccelerate or even pause where irreversible harms may arise'. Such a framework might obtain the best of both worlds, might be bullish on open, free API access for powerful models; it might be bearish on publication, capability keys, and prioritise cyber- and information-security. Fine-graining what such a policy strategy might look like in practise would be a valuable line for future research.

5 Conclusion

> *You cannot [do anything to] control a technology which gets more than a hundred times cheaper to do in half a decade. Not a thing!*
> - Jack Clark, GPT2, Five Years On [115]

As well as benefits, dangerously capable AI models already present significant risks and challenges for governance. The technological pathways that constitute AI proliferation are already coming into existence, will continue to develop, pose substantial risks, and are both less visible to regulators and harder to enforce against. Several strategies for mitigating these risks look promising, but it will require considerable further research to ensure they are calibrated, and defended, both effectively and ethically.

How should policymakers conceptualise the scope of this new challenge, and their individual agency within it? This paper has taken, for an organising metaphor for AI, the nuclear bomb. Yet a tightly centralised, expert-driven, vastly expensive national project with a binary output (a society either 'has' atomic weapons, or it doesn't) is far more reminiscent of 'Big Compute' than AI proliferation. A Proliferation paradigm would involve vastly greater numbers of actors across states, industries and publics; complex interactions between multiple technological pathways; and a more gradual diffusion of risky capabilities from major labs into the broader population than any 'singularity'. If the new paradigm suits a metaphor, it is closer to climate change: self-interested actors acting as individuals, creating compound effects that gradually introduce greater risks into the ecosystem, affecting sudden shifts that require global action to mitigate.

This view sounds pessimistic. Yet an AI proliferation paradigm is unlikely to stumble on a "vulnerable world" of the sort Segrè's bomb suggests [2]. Jack Clark's thesis, above, is overly deterministic: if high-risk capabilities arrive without security and safeguards, it will be due to a panoply of individual decisions. Developers, and the public, have real agency here. The benefits of AI proliferation are immense, and if society is capable of achieving them without irresponsible risks, they should try to do so. At the same time, if risks cannot be mitigated effectively, or mitigated ethically, then pausing, rerouting funding away from high-risk strategies, and steering towards reversible experimentation, should all be real considerations. Without sufficient safeguards or societal resilience, societies that hurtle towards AI proliferation might do so at their peril.

References

[1] Richard Rhodes. *The Making of the Atomic Bomb*. Simon and Schuster, 2012.

[2] Nick Bostrom. "The vulnerable world hypothesis". In: *Global Policy* 10.4 (2019), pp. 455–476.

[3] Jaime Sevilla et al. "Compute trends across three eras of machine learning". In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2022, pp. 1–8.

[4] Girish Sastry et al. "Computing Power and the Governance of Artificial Intelligence". In: *arXiv preprint arXiv:2402.08797* (2024).

[5] Markus Anderljung et al. "Frontier AI regulation: Managing emerging risks to public safety". In: *arXiv preprint arXiv:2307.03718* (2023).

[6] Lennart Heim. "Training Compute Thresholds: Features and Functions in AI Governance". In: *arXiv preprint arXiv:2405.10799* (2024).

[7] Giovanni Dosi. "Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change". In: *Research policy* 11.3 (1982), pp. 147–162.

[8] Anthropic. *Anthropic's Responsible Scaling Policy*. Tech. rep. Anthropic, 2023.

[9] Anca Dragan, Helen King, and Allan Dafoe. *Introducing the Frontier Safety Framework*. Tech. rep. Google DeepMind, 2024.

[10] OpenAI. *Preparedness Framework (Beta)*. Tech. rep. OpenAI, December 18 2023.

[11] *EU AI Act*. 2021.

[12] Allan Dafoe. "AI Governance". In: *The Oxford Handbook of AI Governance* (2024), p. 21.

[13] Johannes Schneider et al. "AI governance for businesses". In: *arXiv preprint arXiv:2011.10672* (2020).

[14] Araz Taeihagh. "Governance of artificial intelligence". In: *Policy and society* 40.2 (2021), pp. 137–157.

[15] Lewis Ho et al. "International institutions for advanced AI". In: *arXiv preprint arXiv:2307.04699* (2023).

[16] Toby Shevlane et al. "Model evaluation for extreme risks". In: *arXiv preprint arXiv:2305.15324* (2023).

[17] Miles Brundage et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. 2018. arXiv: `1802.07228 [cs.AI]`.

[18] Jonas B Sandbrink. "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools". In: *arXiv preprint arXiv:2306.13952* (2023).

[19] OpenAI. *Building an early warning system for LLM-aided biological threat creation*. Tech. rep. OpenAI, 2024).

[20] Yisroel Mirsky et al. "The threat of offensive ai to organizations". In: *Computers and Security* 124 (2023), p. 103006.

[21] Masike Malatji and Alaa Tolah. "Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI". In: *AI and Ethics* (2024), pp. 1–28.

[22] Jiaming Ji et al. "Ai alignment: A comprehensive survey". In: *arXiv preprint arXiv:2310.19852* (2023).

[23] Usman Anwar et al. "Foundational challenges in assuring alignment and safety of large language models". In: *arXiv preprint arXiv:2404.09932* (2024).

[24] Daron Acemoglu. *Harms of AI*. Tech. rep. National Bureau of Economic Research, 2021.

[25] Tom Slee. "The incompatible incentives of private-sector AI". In: *The Oxford Handbook of Ethics of AI* (2020), pp. 106–123.

[26] Richard Moulange et al. *Towards Responsible Governance of Biological Design Tools*. 2023. arXiv: `2311.15936 [cs.CY]`.

[27] Brianna Richardson and Juan E Gilbert. "A framework for fairness: A systematic review of existing fair ai solutions". In: *arXiv preprint arXiv:2112.05700* (2021).

[28] Carissa Véliz. "Data, Privacy, and the Individual: A Report for the Center for the Governance of Change (https://philarchive.org/rec/VLIPM)". In: (2020).

[29] Carissa Véliz. *Privacy is Power*. London, UK: Penguin (Bantam Press), 2020.

[30] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. "Accountability in artificial intelligence: what it is and how it works". In: *AI and SOCIETY* (2023), pp. 1–12.

[31] Elizabeth Seger et al. "Democratising AI: Multiple meanings, goals, and methods". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 715–722.

[32] Hilary Greaves and William MacAskill. "The case for strong longtermism". In: *Global Priorities Institute Working Paper No. 5-2021* (2021).

[33] Huw Roberts et al. "Global AI governance: barriers and pathways forward". In: *International Affairs* 100.3 (2024), pp. 1275–1286.

[34] Angela Daly et al. "AI, governance and ethics: global perspectives". In: *University of Hong Kong Faculty of Law Research Paper* 2020/051 (2020).

[35] Jinghan Zeng. "Artificial intelligence and China's authoritarian governance". In: *International Affairs* 96.6 (2020), pp. 1441–1459.

[36] Allan Dafoe. "AI governance: a research agenda". In: *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK* 1442 (2018), p. 1443.

[37] Heather Douglas. "Inductive Risk and Values in Science". In: *Philosophy of Science* 67.4 (2000), pp. 559–579. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/188707 (visited on 06/04/2024).

[38] Elissar Toufaily, Tatiana Zalan, and Soumaya Ben Dhaou. "A framework of blockchain technology adoption: An investigation of challenges and expected value". In: *Information and Management* 58.3 (2021), p. 103444.

[39] Tommaso Ciarli et al. "Digital technologies, innovation, and skills: Emerging trajectories and challenges". In: *Research Policy* 50.7 (2021), p. 104289.

[40] Jonathan C Ho and Chung-Shing Lee. "A typology of technological change: Technological paradigm theory with validation and generalization from case studies". In: *Technological Forecasting and Social Change* 97 (2015), pp. 128–139.

[41] Ben Buchanan. *"The AI Triad and What It Means for National Security Strategy"*. Tech. rep. Center for Security and Emerging Technology, August 2020.

[42] Jordan Hoffmann et al. "Training compute-optimal large language models". In: *arXiv preprint arXiv:2203.15556* (2022).

[43] David Owen. *How predictable is language model benchmark performance?* 2024. arXiv: 2401.04757 [cs.LG].

[44] Leopold Aschenbrenner. *Racing to the Trillion-Dollar Cluster*. Tech. rep. Situational Awareness (Blog), June 6th 2024.

[45] Megha Panjwani and Suman De. "Study of Cloud Security in Hyperscalers". In: (2020). DOI: https://doi.org/10.23919/indiacom49435.2020.9083727.

[46] Kirsten E Martin. "Ethical issues in the big data industry". In: *Strategic Information Management*. Routledge, 2020, pp. 450–471.

[47] Bernd W. Wirtz, Paul F. Langer, and Jan C. Weyerer. "'An Ecosystem Framework of AI Governance'". In: *The Oxford Handbook of AI Governance, Justin B. Bullock, and others (eds)* (2024).

[48] Janet Egan and Lennart Heim. *Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers*. 2023. arXiv: 2310.13625 [cs.CY].

[49] Joseph R Biden. "Executive order on the safe, secure, and trustworthy development and use of artificial intelligence". In: (2023).

[50] Thomas S Kuhn. *The structure of scientific revolutions*. Vol. 962. University of Chicago press Chicago, 1997.

[51] Paul Scharre. *Future-Proofing Frontier AI Regulation*. Tech. rep. Center For National American Security, 2024.

[52] Danny Hernandez and Tom B. Brown. *Measuring the Algorithmic Efficiency of Neural Networks*. 2020. arXiv: 2005.04305 [cs.LG].

[53] Anson Ho et al. *Algorithmic progress in language models*. 2024. arXiv: 2403.05812 [cs.CL].

[54] Amir Gholami et al. *A Survey of Quantization Methods for Efficient Neural Network Inference*. 2021. arXiv: 2103.13630 [cs.CV].

[55] Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. *A Survey on Deep Neural Network Pruning-Taxonomy, Comparison, Analysis, and Recommendations*. 2023. arXiv: 2308.06767 [cs.LG].

[56] Sheikh Musa Kaleem et al. *A Comprehensive Review of Knowledge Distillation in Computer Vision*. 2024. arXiv: 2404.00936 [cs.CV].

[57] Lingling Xu et al. *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. 2023. arXiv: 2312.12148 [cs.CL].

[58] Marah Abdin et al. "Phi-3 technical report: A highly capable language model locally on your phone". In: *arXiv preprint arXiv:2404.14219* (2024).

[59] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[60] Sachin Mehta et al. *OpenELM: An Efficient Language Model Family with Open Training and Inference Framework*. 2024. arXiv: 2404.14619 [cs.CL].

[61] Noam Shazeer et al. *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*. 2017. arXiv: 1701.06538 [cs.LG].

[62] Mosaic Research Team. *Introducing DBRX: A New State-of-the-Art Open LLM*. Tech. rep. Databricks, March 27, 2024).

[63] Snowflake AI Research. *Snowflake Arctic: The Best LLM for Enterprise AI — Efficiently Intelligent, Truly Open*. Tech. rep. Snowflake, April 24 2024.

[64] Aidan McLaughlin. *The Bitter-er Lesson*. Tech. rep. Personal Blog, 14 June 2024.

[65] Ziming Liu et al. *KAN: Kolmogorov-Arnold Networks*. 2024. arXiv: 2404.19756 [cs.LG].

[66] Xingyou Song et al. *Position: Leverage Foundational Models for Black-Box Optimization*. 2024. arXiv: 2405.03547.

[67] Konstantin Pilz, Lennart Heim, and Nicholas Brown. "Increased Compute Efficiency and the Diffusion of AI Capabilities". In: *arXiv preprint arXiv:2311.15377* (2023).

[68] Omkar Thawakar et al. "MobiLlama: Towards Accurate and Lightweight Fully Transparent GPT". In: *arXiv preprint arXiv:2402.16840* (2024).

[69] Nathan Lambert. *Phi 3 and Arctic: Outlier LMs are hints*. Tech. rep. Interconnects.ai, April 30 2024.

[70] Konstantin Pilz, Lennart Heim, and Nicholas Brown. *Increased Compute Efficiency and the Diffusion of AI Capabilities*. 2024. arXiv: 2311.15377 [cs.CY].

[71] Tim Keary. *GPT-2: The Strange Debacle Surrounding Mystery 'OpenAI' Chatbot*. Tech. rep. Techopedia, 11 May 2024.

[72] Noam Kolt et al. *Responsible Reporting for Frontier AI Development*. 2024. arXiv: 2404.02675 [cs.CY].

[73] Jinze Bai et al. "Qwen technical report". In: *arXiv preprint arXiv:2309.16609* (2023).

[74] IBM. *Through partnership with IBM, Saudi Data and Artificial Intelligence Authority (SDAIA) launches a groundbreaking Arabic AI model to the Middle East*. Tech. rep. IBM, May 21 2024.

[75] Lisa Barrington. *Abu Dhabi makes its Falcon 40B AI model open sourc*. Tech. rep. Reuters, May 25 2023.

[76] Francisco Eiras et al. "Near to Mid-term Risks and Opportunities of Open Source Generative AI". In: *arXiv preprint arXiv:2404.17047* (2024).

[77] Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL].

[78] Mirac Suzgun and Adam Tauman Kalai. *Meta-Prompting: Enhancing Language Models with Task-Agnostic Scaffolding*. 2024. arXiv: 2401.12954 [cs.CL].

[79] Tom Davidson et al. *AI capabilities can be significantly improved without expensive retraining*. 2023. arXiv: 2312.07413 [cs.AI].

[80] Dwarkesh Patel. *Sholto Douglas and Trenton Bricken - How to Build and Understand GPT-7's Mind (https://www.dwarkeshpatel.com/p/sholto-douglas-trenton-bricken)*. Tech. rep. Dwarkesh Podcast, 28 May 2024.

[81] Gemini Team et al. *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. 2024. arXiv: 2403.05530 [cs.CL].

[82] Zeyu Han et al. *Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey*. 2024. arXiv: 2403.14608 [cs.LG].

[83] Xiangyu Qi et al. *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!* 2023. arXiv: 2310.03693 [cs.CL].

[84] Takuya Akiba et al. *Evolutionary Optimization of Model Merging Recipes*. 2024. arXiv: 2403.13187 [cs.NE].

[85] Cohere. *Command R+ (https://docs.cohere.com/docs/command-r-plus)*. Tech. rep. Cohere, 23 May 2024.

[86] Github. *Github Copilot (https://github.com/features/copilot)*. Tech. rep. Github, october 2021.

[87] Qingyun Wu et al. “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework”. In: 2023. arXiv: 2308.08155 [cs.AI].

[88] Shaokun Zhang et al. “Training Language Model Agents without Modifying Language Models”. In: *ICML'24* (2024).

[89] Deepmind. *Frontier Safety Framework*. Tech. rep. Google Deepmind, 17 May 2024.

[90] Prerna Agarwal et al. “Multi-Stage Prompting for Next Best Agent Recommendations in Adaptive Workflows”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 21. 2024, pp. 22843–22849.

[91] Paul Christiano et al. *Deep reinforcement learning from human preferences*. 2023. arXiv: 1706.03741 [stat.ML].

[92] Amir Fard Bahreini et al. “Distributing and Democratizing Institutional Power Through Decentralization”. In: *Building Decentralized Trust: Multidisciplinary Perspectives on the Design of Blockchains and Distributed Ledgers* (2021), pp. 95–109.

[93] Erik Brynjolfsson and Andrew Ng. “Big AI can centralize decision-making and power, and that'sa problem”. In: *Missing links in ai governance* 65 (2023).

[94] Zeming Wei et al. *Jailbreak and Guard Aligned Language Models with Only Few In-Context Demonstrations*. 2024. arXiv: 2310.06387 [cs.LG].

[95] Gabriel Axel Montes and Ben Goertzel. “Distributed, decentralized, and democratized artificial intelligence”. In: *Technological Forecasting and Social Change* 141 (2019), pp. 354–358.

[96] Vid Kersic and Muhamed Turkanovic. *A Review on Building Blocks of Decentralized Artificial Intelligence*. 2024. arXiv: 2402.02885 [cs.AI].

[97] Ishan Gupta. “Decentralization of artificial intelligence: Analyzing developments in decentralized learning and distributed AI networks”. In: *arXiv preprint arXiv:1603.04467* (2020).

[98] Anil Murty. *Distributed Machine Learning on Akash Network With Ray (https://akash.network/)*. Tech. rep. Akash, January 28 2024.

[99] Zhibin Lin et al. “Decentralized Physical Infrastructure Network (DePIN): Challenges and Opportunities”. In: *arXiv preprint arXiv:2406.02239* (2024).

[100] Golem. *Golem (https://www.golem.network/)*. Tech. rep. Golem, 2021.

[101] DeepBrain Chain. *Welcome to DeepBrain Chain (https://www.deepbrainchain.org/)*. Tech. rep. DeepBrain Chain, 2017.

[102] iExec. *Build, Own and Monetise Web3 (https://iex.ec/)*. Tech. rep. iExec, 2020.

[103] gensyn. *Gensyn Litepaper: The hyperscale, cost-efficient compute protocol for the world's deep learning models(https://docs.gensyn.ai/litepaper)*. Tech. rep. Gensyn, 2024.

[104] Aksh Garg. *Shard: On the decentralized training of foundation models (https://aksh-garg.medium.com/shard-on-the-decentralized-training-of-foundation-models-2fd982176724)*. Tech. rep. Medium, May 20 2024.

[105] Arthur Douillard et al. "DiPaCo: Distributed Path Composition". In: *arXiv preprint arXiv:2403.10616* (2024).

[106] Junchen Zhao et al. *LinguaLinked: A Distributed Large Language Model Inference System for Mobile Devices*. 2023. arXiv: 2312.00388 [cs.LG].

[107] Elizabeth Seger et al. "Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives". In: *arXiv preprint arXiv:2311.09227* (2023).

[108] Wei-Lin Chiang et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. 2024. arXiv: 2403.04132 [cs.AI].

[109] Meta AI Research. *Introducing Meta Llama 3: The most capable openly available LLM to date (https://ai.meta.com/blog/meta-llama-3/)*. Tech. rep. Meta, 18 April 2024.

[110] Dwarkesh Patel. *Mark Zuckerberg - Llama 3, Open Sourcing 10b Models, Caesar Augustus*. Tech. rep. Dwarkesh podcast, 18 April 2024.

[111] HuggingFace. *HugginFace index (https://huggingface.co/docs/hub/en/index)*. Tech. rep. Huggingface, 2024.

[112] David Evan Harris. *How to Regulate Unsecured "Open-Source" AI: No Exemptions*. Tech. rep. Tech Policy Press, December 4 2023.

[113] David Evan Harris. *How to Regulate Unsecured "Open-Source" AI: No Exemptions*. Tech. rep. Tech Policy Press, 18 December 2023.

[114] HuggingFace. *Content Policy (https://huggingface.co/content-guidelines)*. Tech. rep. HuggingFace, 30 August 2023.

[115] Jack Clark. *Import AI Import AI 375: GPT-2 five years later; decentralized training; new ways of thinking about consciousness and AI*. Tech. rep. Demoscene AI, 2024.

[116] Iason Gabriel. "Toward a theory of justice for artificial intelligence". In: *Daedalus* 151.2 (2022), pp. 218–231.

[117] Sarah Kreps and Doug Kriner. "How AI Threatens Democracy". In: *Journal of Democracy* 34.4 (2023), pp. 122–131.

[118] Karl Manheim and Lyric Kaplan. "Artificial intelligence: Risks to privacy and democracy". In: *Yale JL and Tech.* 21 (2019), p. 106.

[119] Noémi Bontridder and Yves Poullet. "The role of artificial intelligence in disinformation". In: *Data and Policy* 3 (2021), e32.

[120] Hannah Ritchie et al. "Internet". In: *Our World in Data* (2023).

[121] David Evans. “Number of software developers worldwide in 2018 to 2024 (in millions)”. In: *Statista* ((August 21, 2023)).

[122] et al. Maslej N. *Artificial Intelligence Index Report 2023: Chapter 7 Diversity*. Tech. rep. Institute for HumanCentered AI, Stanford University, Stanford, CA, 2023.

[123] Edvard PG Bruun and Alban Duka. “Artificial intelligence, jobs and the future of work: Racing with the machines”. In: *Basic Income Studies* 13.2 (2018), p. 20180018.

[124] Toby Shevlane. “Structured access: an emerging paradigm for safe AI deployment”. In: *arXiv preprint arXiv:2201.05159* (2022).

[125] Irene Solaiman. “The gradient of generative AI release: Methods and considerations”. In: *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 2023, pp. 111–122.

[126] Benjamin S Bucknall and Robert F Trager. “STRUCTURED ACCESS FOR THIRD-PARTY RESEARCH ON FRONTIER AI MODELS: INVESTIGATING RESEARCHERS’MODEL ACCESS REQUIREMENTS”. In: (2023).

[127] METR. *Responsible Scaling Policies (https://metr.org/blog/2023-09-26-rsp/)*. Tech. rep. METR, 26 September 2023.

[128] Ben Garfinkel and Allan Dafoe. “How does the offense-defense balance scale?” In: *Emerging Technologies and International Stability*. Routledge, 2021, pp. 247–274.

[129] Jamie Bernardi et al. “Societal Adaptation to Advanced AI”. In: *arXiv preprint arXiv:2405.10295* (2024).

[130] Geoff Gordon, Bernhard Rieder, and Giovanni Sileno. “On mapping values in AI governance”. In: *Computer, Law and Security Review* 46 (2022), p. 105712.

[131] Helen Nissenbaum. “Privacy As Contextual Integrity”. In: *Washington Law Review* 79 (May 2004).

[132] Julien Mercille and Enda Murphy. “Market, non-market and anti-market processes in neoliberalism”. In: *Critical Sociology* 45.7-8 (2019), pp. 1093–1109.

[133] Shawn Donnelly. “Semiconductor and ICT Industrial Policy in the US and EU: Geopolitical Threat Responses”. In: *Politics and Governance* 11.4 (2023), pp. 129–139.

[134] Onni Aarne, Tim Fist, and Caleb Withers. “Secure, Governable Chips Using On-Chip Mechanisms to Manage National Security Risks from AI and Advanced Computing”. In: *CNAS* (January 2024).

[135] Vinodkumar Prabhakaran et al. “A human rights-based approach to responsible AI”. In: *arXiv preprint arXiv:2210.02667* (2022).

[136] John A Robertson. "The Scientist's Rights to Research: A Constitutional Analysis". In: *S. Cal. l. Rev.* 51 (1977), p. 1203.

[137] Steven Goldberg. "The Constitutional Status of American Science". In: *U. Ill. LF* (1979), p. 1.

[138] Gert Verschraegen. "Regulating scientific research: A constitutional moment?" In: *Journal of Law and Society* 45 (2018), S163–S184.

[139] Gregory Lewis et al. "Information hazards in biotechnology". In: *Risk Analysis* 39.5 (2019), pp. 975–981.

[140] Jessica Rumbelow and Matt Watkins. *SolidGoldMagikarp (plus, prompt generation)*. Tech. rep. LessWrong, February 5 2023.

[141] Sella Nevo et al. "RAND Report: Securing AI Model Weights". In: (2024).

[142] Connor Leahy et al. *Conjecture Internal Infohazard Policy*. Tech. rep. Conjecture, 29th July 2022.

[143] Information Technology Laboratory Computer Security Division. *Managing Information Security Risk Organization, Mission, and Information System View*. Tech. rep. National Institute of Standards and Technology, 2017.

[144] Martin Horák, Václav Stupka, and Martin Husák. "GDPR Compliance in Cybersecurity Software: A Case Study of DPIA in Information Sharing Platform". In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. ARES '19. Canterbury, CA, United Kingdom: Association for Computing Machinery, 2019. ISBN: 9781450371643. DOI: 10.1145/3339252.3340516. URL: https://doi.org/10.1145/3339252.3340516.

[145] Ian Goldstein et al. *Understanding the (In)Effectiveness of Content Moderation: A Case Study of Facebook in the Context of the U.S. Capitol Riot*. 2023. arXiv: 2301.02737 [cs.SI].

[146] Nik Popli. "The 5 Most Important Revelations From the 'Facebook Papers' https://time.com/6110234/facebook-papers-testimony-explained/". In: *Time* (October 26, 2021).

[147] Parsee Mizuhashi. *Glaze and the Effectiveness of Anti-AI Methods for Diffusion Models (https://huggingface.co/blog/parsee-mizuhashi/glaze-and-anti-ai-methods)*. Tech. rep. HuggingFace, May 15 2024.

[148] Masaki Imai et al. "Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets". In: *Nature* 486.7403 (2012), pp. 420–428.

[149] Ryan S Noyce, Seth Lederman, and David H Evans. "Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments". In: *PloS one* 13.1 (2018), e0188453.

[150] RM Atlas. "Biodefense research: an emerging conundrum". In: *Current Opinion Biotechnology* (2005 June).

[151] Deep Ganguli et al. “Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned”. In: *arXiv preprint arXiv:2209.07858* (2022).