Journal template credit: The Centre for Technology and Publishing at Birkbeck, University of London.

Though Shalt Not Freeze Frame: The rise of catastrophic AI in safety discourse

Maryam Akram, akram_maryam@live.co.uk

This paper explores the shifting landscape of AI governance and policy, focusing on the emergence of catastrophic risk as a dominant framework within AI safety debates. It examines how this discourse has evolved and become embedded across key stakeholder groups including, government and civil policymakers, industry leaders, academic institutions and societal actors. These stakeholders collectively shape the priorities and strategies of AI governance. Catastrophic or existential risk can be defined as the potential for AI to cause harm on a society-wide scale, driven by factors like misalignment, single point of failure and overreliance.[1] The paper examines the language used by AI startup governance policies and materials as well as the topics addressed in university AI ethics and governance courses. It investigates whether a disconnect exists between the high-profile warnings about the existential risks posed by advancements in deep learning techniques and the broader AI safety and governance discourse.

The aim of this paper is not to critique proactive policy research on catastrophic AI risks. Instead, it explores the mechanisms that have elevated artificial general intelligence (AGI) to prominence in AI safety discussions and examines how these dynamics may influence policy development. Indeed, the pursuit of

[1] https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf

solving the problem of intelligence has been a key motivator in the evolution of the field of AI, and for many AI labs, AGI continues to be a north star. This research analyses how AGI and AGI-related risks have been institutionalized in formal agreements, such as the reported OpenAI's Microsoft AGI clause[2], polarized risk priorities within the AI safety community, informed scaling laws, and influenced the development of 'future-focused' regulations.

This paper draws on interdisciplinary frameworks including Bruno Latour's process-oriented and relational analysis and ideas from his seminal essay 'Thou Shalt Not Freeze Frame'. It advocates for policy development that is use-case driven, grounded in scientific consensus, and informed by real-world examples of AI-related harms.

Additionally, the research is grounded in key insights from three foundational papers:

- On the Limitations of Compute Thresholds as a Governance Strategy by Sara Hooker.[3]
- Data-Centric AI Governance: Addressing the Limitations of Model-Focused Policies by Gupta, Walker, Corona, Fu, Petryk, Napolitano, Darrel, and Reddie. [4]
- What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis by Fiesler, Garrett, and Beard.[5]

From distant catastrophes to everyday harms: A Latourian perspective of AI risk priorities

"Scientists are very much entangled in their culture, and this culture is not pristine, untouched by other cultures and practices." – Bruno Latour (2011)

"Neither science nor religion fits even this basic picture that would put them face-to-face, or enough in relation to be deemed incommensurable." – Bruno Latour, Thou Shalt Not Freeze Frame (2005)

Thou Shalt Not Freeze Frame

The philosopher and anthropologist Bruno Latour, known for his work in actor-network theory (ANT), the philosophy of science and his critique of modernity as a fallacy, offers a framework that can be used to reframe catastrophic AI risk and systemic AI risk. In Thou Shalt Not Freeze Frame, Latour challenges the conventional polarized view of science and religion as competing based on religion as abstract and science as objective and visible. Instead, Latour argues that science and religion operate in different temporal and conceptual domains – science often strives for permanence and universality, while religion deals with the immediate, subjective and dynamic.

Latour's paper critiques the tendency to reduce or 'freeze frame' dynamic processes into simplified fixed categories, illustrated using the overly simple science and religion dichotomy. Science and religion are not static, and their practices involve ongoing, dynamic interactions with the world.

[2] https://www.nytimes.com/2024/10/17/technology/microsoft-openai-partnership-deal.html
[3] https://arxiv.org/abs/2407.05694
[4] https://arxiv.org/abs/2409.17216
[5] https://cmci.colorado.edu/~cafi5706/SIGCSE2020_EthicsSyllabi.pdf

Science and religion freeze frame:

- Science is framed as universal, objective, and unchanging - concerned with 'eternal truths.' Latour argues that scientific facts evolve over time and are contingent on social, political and material factors.
- Religion is depicted as subjective, emotional, and personal - focused on immediate experience and relationships. Latour argues that religion can be understood as performative action and active engagement – rather than an inward practice that referencing a distant and invisible world.

Latour use of temporal framing and dynamic relationships can be applied to the way existential AI risk and systemic AI risk are conceptualized:

- Existential AI risk is analogous to the frozen, abstract, and far-off—a 'freeze frame' of imagined future scenarios represented by the most advanced models.
- Current systemic AI risk aligns with the dynamic, every day, and actionable—a lens focused on immediate impacts and continuous evolution of context-specific machine learning models.

This paper challenges the static, 'freeze-frame' representations of existential AI risk and advocates for a more nuanced, dynamic understanding of AI risks. By contrasting the abstract, apocalyptic focus of existential risk with the immediate, evolving nature of current systemic AI challenges, this paper reframes the discussion of AI governance. It aims to shift the focus away from speculative, future-oriented scenarios toward a more grounded perspective that emphasizes actionable, contextualized issues in the present.

The temporal framing of AI Safety

"In three to eight years we will have a machine with the general intelligence of a human being" – Marvin Minky, Life Magazine (1970)

AI forecasting is not new. Historically, AI winters and summers have been defined by ambitious and mistaken projections about AGI. The rise of generative (GenAI) has spurred a new wave of predictions, often built on opaque heuristics and assumptions, which have permeated AI safety discourse.

In contrast, systemic AI risks—such as biases, disinformation, and unequal access to AI technologies—are rarely framed in terms of timelines. These risks, though arguably 'intellectually compelling', demand immediate and tangible interventions. This temporal framing is reflected in regulations like the EU AI Act, the US AI Executive Order and increasingly popular scaling laws. These polices are partially based on concepts like emergence, compute thresholds and future-focused governance tools.

Both Hooker and Gupta et al. highlight that the current definition of 'foundation/frontier models' applies to only a small subset of machine learning models, compared with those 'currently deployed in the wild'. They emphasize the significant capabilities of models trained on domain-specific data and challenge the focus on scale as the primary measure of performance and risk profiles. Future-focused policies represent attempts by regulators and policymakers to accommodate subsequent developments and keep pace with technological innovation and accommodate future developments. However, these policies risk overlooking day-to-day harms that are less susceptible to thought experiments and speculative scenarios. This prioritization of anticipating black swan events over addressing the gradual precipitation and amplification of AI risk misses the nuance and richness of lived experiences.

Double clicking: Fixed governance metrics

Floating-point operations per second (FLOPs) and parameter metrics for AI risk governance, give the impression of objectivity and reliability. FLOPs measure the computational power required to train a model, whereas parameters count the total learnable parameters in a model. Both are used to measure size and complexity. Latour's concept of double clicking – the assumption that information can be accessed and understood in a self-evident manner, without context or interaction – can be applied to the way FLOPs and parameter counts are treated as self-explanatory indicators of AI risk. Indeed, Hooker points out that metrics like FLOPs fail to provide a stable foundation for risk governance due to the uncertain and rapidly evolving relationship between compute and risk. She argues that relying on inflection points of risk is insufficient for effective governance. Here, what Hooker highlights can be considered as double clicking – simplifying the dynamic processes of AI development into fixed metrics. By framing FLOPs as self-evident, governance frameworks risk 'freeze framing' AI risks into a static category which obscure the relational and contextual factors influencing their significance.

Grounding governance in real-world examples

Real-world case studies have long been valuable pedagogical tools in university courses. Thomas's introductory course on Data Ethics demonstrates this by grounding lessons on disinformation, bias and fairness, ethical foundations, privacy and surveillance, and algorithmic colonialism in real-world examples[6].

Despite being dismissed as 'mundane risks' by some industry stakeholders, current systemic risks remain highly relevant for meaningful action and policymaking. Birhane discusses how the most vulnerable in society are disproportionately impacted by the harms driven by the digitalization of services[7]. In her article The Algorithmic Colonization of Africa, she draws an analogy between data collection and mining, framing the relentless pursuit of data as a colonial attitude that treats humans as raw material. Reports of AI and facial recognition tools being used in harmful contexts, such as surveillance of Uyghurs in Xinjiang[8], further underscore the stakes. Cases like men's rights activists orchestrating attempts to use the #EndFathersDay hashtag to sow division demonstrate how malicious actors can easily create and spread damaging content[9]. The drive for scale, is also fuelling unprecedented energy demands to train and deploy AI models. The push for bigger and bigger models appears to be in direct conflict with efforts to reduce the environmental impacts of technology. Stakeholders from various technology companies have met with the White House to discuss and advocate for the future power requirements needed to support evolving AI workload demands.

[6] https://ethics.fast.ai/syllabus/
[7] https://reallifemag.com/the-algorithmic-colonization-of-africa/
[8] https://www.theguardian.com/world/2021/sep/30/uyghur-tribunal-testimony-surveillance-china
[9] https://slate.com/technology/2019/04/black-feminists-alt-right-twitter-gamergate.html

AI policy frameworks need to be balance between speculative foresight and pragmatic intervention. Gupta et al, argue that AI governance should be reconceptualized as a data-centric practice rather than policies which treat AI as simply a 'function of models'. The relative lack of focus on data governance highlights how attention is being drawn to far-off catastrophes, while sidelining messy and difficult to manage but addressable causes of AI harms.

Evolving AI governance and predicting the future

"In particular, the seductive popularity of the trolley problem has led many to misconstrue autonomous vehicles as a technology that is already reliable, ubiquitous, and fully operational — allowing for abstract moral questions about their hypothetical behaviour to be posed, and even answered." – Ian Bogost, Enough with the Trolley Problem (2018)

High-profile predictions and warnings about catastrophic AI risks

An evolving discipline

"...don't ask a compute scientist or economist whether you can predict the future. The temptation to say yes often overrides a necessary humility about what can and cannot be predicted accurately" – Sara Hooker, On the Limitations of Compute Thresholds as a Governance Strategy (2024)

AI ethics and governance, along with the broader technological field they inhabit, remain relatively nascent disciplines. Nevertheless, high-profile scandals have brought the ethical and societal implications of these technologies into the public eye. Incidents such as the Cambridge Analytica scandal and concerns over algorithmic bias in Amazon's hiring models are often cited, but they represent only a fraction of the broader landscape of AI harms and risks. These include, for example, the role of social media platforms in spreading misinformation in the Philippines[10] and their use in the ethnic cleansing of the Rohingya minority in Myanmar[11]. More recently, the working conditions and compensation for data labellers[12] and content moderators[13] in Kenya and have also come under scrutiny.

Overtime AI ethics and governance has expanded and become more mainstream, impacting industry practices and corporate policies, research and academia. The seminal work of Buolamwini and Gebru, Gender Shades, a pioneering study in algorithmic bias in computer vision models, has had a profound effect on algorithmic auditing and research practices. As Birhane notes, their study played a particularly crucial role in the curation of large-scale datasets[14]. In 2020, NeurIPs – a prominent AI research conference – made a broader potential social impact statement a requirement for submissions, as well as any financial conflicts[15]. However, Gupta notes that LAION-5B, submitted in 2022, was accepted and award the 'best

[10] https://www.buzzfeednews.com/article/daveyalba/facebook-philippines-dutertes-drug-war
[11] https://www.technologyreview.com/2018/08/14/240325/how-social-media-took-us-from-tahrir-square-to-donald-trump/
[12] https://time.com/6247678/openai-chatgpt-kenya-workers/
[13] https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/
[14] https://www.nature.com/articles/d41586-022-03050-7
[15] https://venturebeat.com/ai/neurips-requires-ai-researchers-to-account-for-societal-impact-and-financial-conflicts-of-interest/

datasets and benchmarks' paper. The LAION-5B paper, based on Large-scale Artificial Intelligence Open Network LAION dataset, the has been criticized for containing 'copyrighted and harmful' materials[16].

The rapid proliferation of GenAI and its widespread consumer adoption have propelled the technology into public consciousness, amplifying and reshaping the landscape of AI harms and risks. This shift heralds a new era of AI governance and ethics. According to Stanford's Artificial Intelligence Index Report the number of US AI-related regulations in 2023 grew to 25 from one in 2016. 'Artificial Intelligence' was also mentioned 37 times in bills passed in 127 countries surveyed in 2022[17]. Both China and the European Union (EU) have introduced landmark AI regulations addressing GenAI, while the US and UK have established AI Safety Institutes.

Within this evolving AI landscape, marked by the discourse around GenAI, new players, dynamics, and narratives have emerged - chief among them the debate surrounding our proximity to AGI and the existential risks associated with AI. In 2023, the Future of Life Institute called for a six-month pause on training models 'more powerful than GPT-4'[18]. Later that year, the Centre for AI Safety published a statement urging that mitigating the risk of AI-induced extinction should be a global priority alongside other catastrophic risks such as pandemics and nuclear threats[19]. While these statements reflect growing concern, they do not represent consensus within the industry. Nonetheless, they highlight the increasing prominence of cautionary warnings focused on catastrophic AI risk, which is garnering more public attention. Such warnings are not confined to the industry and research communities. In 2023, then-UK Prime Minister Rishi Sunak acknowledged existential risks posed by AI in government documents, stating that these risks could not be ruled out.[20]

Although concerns about catastrophic AI have existed since the early days of the field, these narratives have evolved, especially as AI technology has advanced. Today, discussions of AI risk often draw parallels to the scale of climate change or nuclear disaster. These narratives have also become embedded in the corporate structures of AI labs in the form of alignment teams and responsible scaling laws.

What's in a word: Language and the framing of AI risk

"Ethics isn't a matter of applying a simple calculus to any situation—nor of applying an aggregate set of human opinions about a model case to apparent instances of that model. Indeed, to take those positions is to assume the utilitarian conclusion from the start." – Ian Bogost, Enough with the Trolley Problem (2018)

CEOs of major AI companies have released memos, some of which frame AI's future as a logical progression - highlighting their overarching good and cross-industry innovations and revolutions. In the context of AI governance, the use of simplified or abstract models - such as catastrophic AI or AGI forecasts - can have profound implications for the framing of AI risks. These narratives often predict that super-intelligent machines will soon solve humanity's greatest problems, typically accompanied by ambitious timelines like "in five years" or "by 2026." While these forecasts are bold and attention-grabbing, they tend to oversimplify the political, economic, and social realities that govern AI development. By presenting AI as an inevitable, monolithic force, either as a utopian saviour or an existential threat, the language used in these discussions strips away the complexity and nuances of real-world AI applications. These oversimplifications can obscure the potential for harmful outcomes, but also marginalizes critical conversations about governance, ethics and social inequalities.

A key feature of these forecasts is the reliance on syllogistic reasoning: AI's future is presented as a logical, inevitable progression toward a speculative future i.e., the greater good or catastrophe. However, in this context, the 'greater good'[21] is not scrutinized in terms of its broader societal consequences. This form of reasoning, rooted in abstract hypothetical scenarios, positions AI as something detached from its current applications, ignoring the complex, everyday ethical challenges posed by AI systems in their present state. As a result, the focus shifts away from the more immediate, tangible harms - such as algorithmic bias, disinformation, misinformation and surveillance - that affect vulnerable populations and toward distant, speculative risks.

Framing AI's future risks as inevitable or logical conclusions, without grounding them in tangible, real-world examples of harm, risks diverting attention from the pressing, already unfolding risks. As Hooker highlights, a critical challenge lies in balancing known, current risks with speculative future threats. She critiques precautionary policies, such as compute thresholds, which are based on uncertain, future-oriented scenarios. These policies rest on the assumption that models of a certain scale inherently present greater risks, a premise Hooker questions. By focusing on the risks tied to surpassing specific compute thresholds, policymakers and industry leaders may overlook the 'messier' but urgent issues, such as the current practices surrounding the training of GenAI systems. These immediate concerns, though less speculative, have real consequences and demand attention in AI governance and policy discussions.

The language of catastrophe

The language of catastrophe can at times be problematic – especially when it is highly-publicised without relevant context. Catastrophic AI narratives, though captivating, often reinforce a dualistic framing: they focus on a binary of either a utopian breakthrough or apocalyptic downfall, with little room for the complexities of the ongoing societal impacts of AI systems. This positioning of catastrophic risk as a distinct, abstract category that exists far in the future, enables industry voices to argue that short-term regulation would stifle innovation, while long-term intervention is necessary but undefined[22][23]. In effect, these narratives render certain risks as 'unimaginable' and others as 'manageable', diverting attention and regulatory efforts toward catastrophe while sidelining current systemic risks deemed less pressing.

[16] https://www.techpolicy.press/laion-and-the-challenges-of-preventing-ai-generated-csam/

[17] https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024_Chapter7.pdf

[18] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[19] https://www.safe.ai/work/statement-on-ai-risk#open-letter

[20] https://www.theguardian.com/technology/2023/oct/25/ai-dangers-must-be-faced-head-on-rishi-sunak-to-tell-tech-summit

[21] https://moores.samaltman.com/

[22] https://www.theguardian.com/technology/2023/may/24/openai-leaders-call-regulation-prevent-ai-destroying-humanity?ref=foundr.ai

[23] https://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk

Bogost highlights the issues of relying on thought experiments to evaluate the moral complexities of technology. Thought experiments like the trolley problem, often serve to simplify and distil complex moral dilemmas into manageable, abstract scenarios. While they may be useful mechanisms to explore specific ethical principles, they often fail to address the full scope of real-world consequences. Similarly, in AI safety, the overreliance on speculative, far-reaching risks (such as AGI) distracts from the present-day use of AI.

The words used to frame AI risk are far from neutral, and it's crucial to understand how they shape our perceptions and responses to both current and future technological challenges. The prevailing narrative of a potential apocalyptic event - an unintended consequence of model training - positions technology companies as the gatekeepers of a powerful technology that could either deliver boundless benefits or cause catastrophic harm. Such discourses, however, overlook the reality that AI is a tool embedded within a complex web of actors, including humans, machines, data, and institutions. As Seaver argues, technology cannot be understood in isolation or framed as a disinterested tool that can be wielded for either good or evil[24].

The language we use in discussing AI risks matters deeply, not just because of its impact on public understanding, but because it reflects the underlying power dynamics and motivations that drive technological development. Efforts to draw on popular culture imagery[25] and tropes[26] when designing AI products may inadvertently reinforce misleading or unhelpful narratives about the technology's capabilities and risks. Initiatives by some AI labs to publish system prompts marks a positive step forward in terms of transparency, helping the public better understand how large language models (LLMs) operate[27].

Misaligned incentives

"The pursuit of monopoly has led Silicon Valley astray." — Tim O'Reilly, The Fundamental Problem with Silicon Valley's Favorite Growth Strategy (2019)

The backdrop to catastrophic AI and AGI is the competitive landscape within the current AI industry. The exorbitant cost[28] associated with training LLMs have sparked a new race among firms, all vying for sustainable business models while striving to remain at the cutting edge of technology. This race is currently defined by the pursuit of scale and compute power, with access to high-performance computing becoming a critical dynamic. Tim O'Reilly's critique of 'Blitzscaling' - a strategy that prioritizes speed and rapid growth over efficiency in order to achieve market dominance - applies to the current AI landscape. Industry sectors like search are being drastically reshaped as firms challenge existing monopolies with new AI-driven tools and systems. Firms are participating in funding rounds and rapid product releases to stay competitive and hopefully claim the market. But is this really what is meant by 'democratizing AI'?

[24] https://culanth.org/fieldsights/anthropology-and-algorithms#transcript287485

[25] https://www.bbc.co.uk/news/articles/ce3z37dpvl9o

[26] https://www.theguardian.com/technology/article/2024/may/27/scarlett-johansson-openai-legal-artificial-intelligence-chatgpt

[27] https://docs.anthropic.com/en/release-notes/system-prompts

[28] https://fortune.com/2019/10/03/openai-will-need-more-capital-than-any-non-profit-has-ever-raised/

In order to raise the capital necessary to sustain high-stakes competition, firms must demonstrate a competitive edge. This drive for dominance can lead to inflated AI projections and narratives, fuelled by the need to maintain investor confidence. Although companies may explicitly assert that the 'race to AGI' will not come at the expense of safety - through mechanisms such as early warning signals, red teaming, and emergent capability reporting - corporate incentives make self-regulation unreliable. The commitments to cooperate, form partnerships and withhold potentially dangerous models stand in contrast to current practices which infringe on copyright and scrape sensitive and toxic internet data. The implications of the digitalization of services and the widespread presence of apps in societies are well-documented[29][30]. While existing privacy laws, such as GDPR, aim to safeguard data rights, the rise of GenAI is intensifying existing issues. Indeed, even standard practices like privacy policies, in practice are ineffective methods of obtaining meaningful consent[31].

While red teaming, safety training, and guardrails may be valuable AI governance tools, the fundamental processes outlined by Gupta et al., such as 'existing data filtration' and the integration of data evaluation frameworks, remain neglected. Gupta et al. argue that models are highly unlikely generate specific or sensitive data (e.g., instances of child sexual abuse material) without being exposed to that data during training. However, ethical and legal concerns over training data seem to be frequently overlooked in the rush to scale. The incentive to fuel vast models with more and more data informs the way technology companies interact with communities and users.

Moreover, AI safety pledges are shown to be highly contingent upon rapidly changing corporate structures. The majority of GenAI startups, many backed by hyperscalers, are less than a decade old and have already been mired in controversies, senior executive exits, non-disparagement clauses, shifts in safety departments and transitions from non-profit to capped-profit models. Pressure from investor dynamics can significantly impact changes in corporate direction. Regardless of whether startups deviate from their initial missions to develop AI for the public good, the incentives inherent in a capital-intensive, highly competitive industry mean that these companies cannot reliably define the AI safety discourse. Their efforts to collaborate with government bodies, too, should be critically examined, as firms may lobby for favourable policies[32]. Indeed, while the expertise and perspective of industry stakeholders is a critical aspect of AI governance and policy, the way AI safety discourse is embedded across different stakeholders needs to be scrutinized.

Changing industry governance practices and the language of the new safety eta

In their review of The Evolution of AI Governance, Chesterman et al. trace the emergence and growth of AI governance frameworks, guidelines, and principles, highlighting the catalytic role of the 2016 Cambridge Analytica scandal. They argue that AI governance is a relatively recent phenomenon, gaining traction

[29] https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html

[30] https://www.fastcompany.com/90447583/our-collective-privacy-problem-is-not-your-fault

[31] https://www.theatlantic.com/technology/archive/2012/03/reading-the-privacy-policies-you-encounter-in-a-year-would-take-76-work-days/253851/

[32] https://techcrunch.com/2024/07/31/openai-pledges-to-give-u-s-ai-safety-institute-early-access-to-its-next-model/

policies. Chesterman et al. identify a convergence around six key principles frequently cited in governance documents:

- Fairness
- Accountability
- Transparency/Explainability
- Ethics/Human-Centricity
- Safety/Security
- Privacy

A significant hallmark of this shift has been the creation of dedicated AI governance bodies within technology firms, such as Responsible AI or AI Ethics departments. These teams are tasked with establishing governance protocols, conducting research, and developing methodologies for ethical AI deployment. However, some of these departments have faced controversy, with a number being disbanded or restructured over time[1].

Chesterman et al. observe a 'remarkable consistency' in the terminology used in AI governance policies from companies and governments alike. However, their analysis also reveals divergences in the specific meanings attributed to these principles. For instance, while 'fairness' is commonly referenced, it can mean different goals, such as non-discrimination, inclusive design, or broader societal inclusivity. Notably, corporate policies were less likely to include 'openness' as part of transparency, compared to governmental frameworks.

As GenAI becomes increasingly central to the industry, new types of governance units like alignment teams and frontier ethics teams have emerged, focused specifically on managing advanced AI capabilities.

Analysing AI governance language: A focus on startups

Building on Chesterman et al.'s approach, I conducted a review of governance and policy materials published by major AI startups. My aim was to analyse the language and terminology used in AI governance. To ensure a different perspective, I excluded policies from established big tech firms providing LLMs, as their documents often reflect historic practices and exhibit significant convergence, as noted by Chesterman et al. The materials I analysed included a mix of formal policies, governance research, and general statements on governance approaches.

To organize the data, I identified key categories and conducted a word frequency analysis based on these themes. The table below outlines the categories and associated keywords:

[1] https://www.theverge.com/2023/11/18/23966980/meta-disbanded-responsible-ai-team-artificial-intelligence

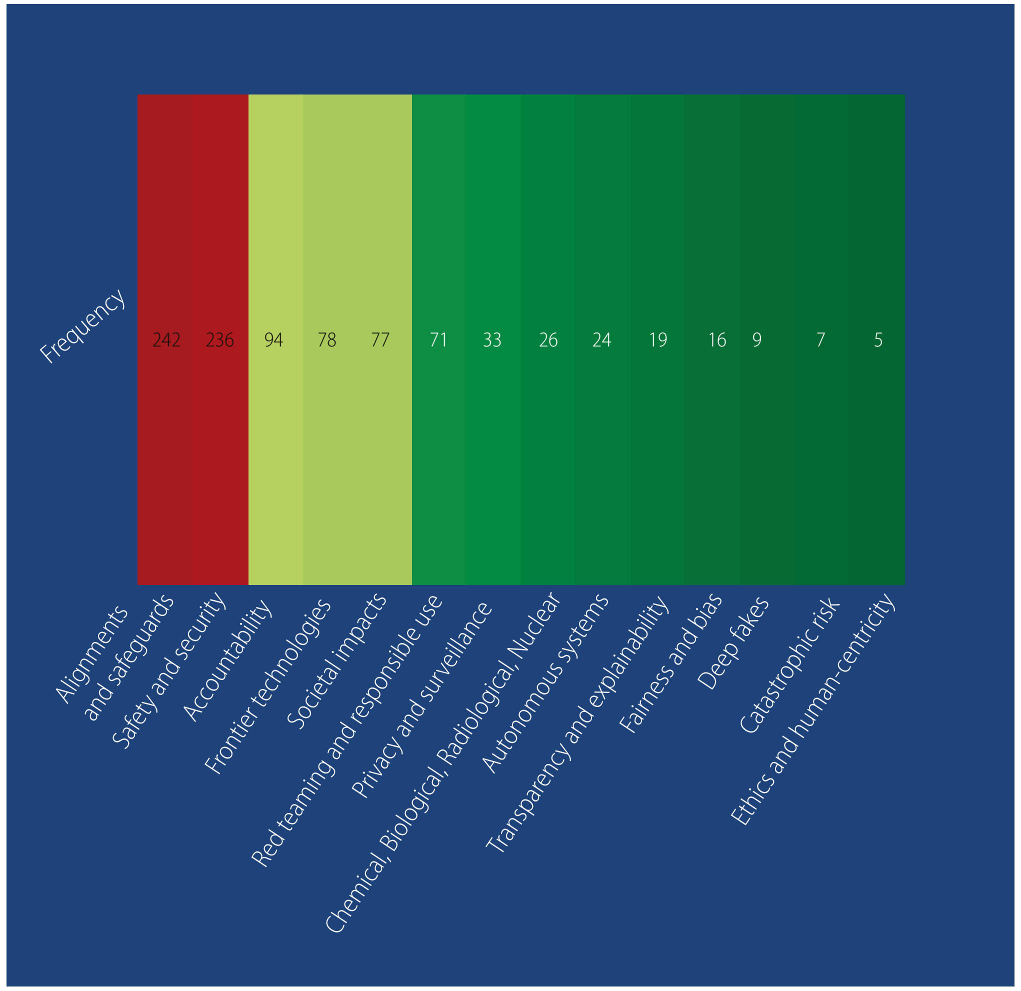
Figure 1: Governance and policy categories with keywords

Categories	Keywords
Fairness and bias	Fairness, bias, equity, and discrimination
Privacy and surveillance	Privacy, surveillance, cyber, and data protection
Accountability	Accountability, oversight, commitment, commitments, and trust
Catastrophic risk	Existential risk, catastrophic risk
Transparency and explainability	Transparency, explainability, and interpretability
Ethics and human-centered	Ethics, human-centered, and values
Safety and security	Safety, security, child safety, and safe systems
Deep fakes	Deep fake, synthetic media, fake content, misinformation, disinformation, and elections
Societal impacts	Societal, society, impacts, benefits, and challenges
Alignment and safeguards	Alignment, scaling, safeguards, and responsible
Frontier technologies	Frontier, frontier risks, LLM, generative, and advanced AI
Chemical, biological, radiological and nuclear	Chemical, biological, radiological, and nuclear
Autonomous systems	Autonomous, autonomous vehicles, and autonomy
Red teaming and responsible use	Red team, responsible, and adversarial testing

Although a more systematic and exhaustive collection would be necessary to draw definitive conclusions, the preliminary analysis reveals significant variation in the terminology used by different AI startups. This variation suggests that the language surrounding AI startup governance has yet to converge meaningfully, reflecting its status as a developing and nascent area within the field. Notably, one major AI startup had no publicly available AI governance or policy materials, further emphasizing the uneven adoption of standardized governance language across the industry.

The statistical summary of word frequencies: M = 66.93, Mdn=29.5, SD = 78.72 and R = 237

Figure 1: AI governance category heatmap



Shaping the future: How AI ethics and governance are taught in universities

In 2020, Fiesler et al. published an analysis of university syllabi from technology ethics courses, examining the topics covered, the departments hosting these classes, and the disciplinary backgrounds of instructors. Their findings revealed that courses on technology ethics spanned diverse departments, including Computer Science, Philosophy, and Law. Notably, while Computer Science departments were the most common hosts, instructors often had academic backgrounds in Information Science or Philosophy.

The topic analysis conducted by Fiesler et al. showed that law and policy were the most frequently covered subjects, followed by privacy and surveillance. Philosophy, including traditional ethical theories, appeared in 53% of the courses evaluated. Interestingly, courses based in Computer Science departments often focused on utilitarianism when addressing ethical theory.

This emphasis on utilitarianism within Computer Science curricula may reflect broader trends in how technology ethics is framed. A utilitarian perspective, which emphasizes outcomes and consequences, aligns closely with narratives surrounding catastrophic risks in AI. This framing could shape how students conceptualize ethical responsibilities, emphasizing large-scale, long-term impacts over localized or systemic concerns.

Following Fiesler et al.'s analysis of university ethics syllabi, I conducted a brief qualitative analysis of university AI governance and ethics courses, including undergraduate, graduate and online summer school courses. Interdisciplinary approaches and philosophy, ethics and legal perspectives appeared to be common.

Some key topics included:

- Privacy and surveillance
- Algorithmic bias and fairness
- Regulatory frameworks
- Responsible innovation
- Ethical decision-making in AI
- AI and human rights
- Governance of AI

Future focused regulations: Compute thresholds and scaling laws

'Discussions prioritizing model size as a viable threshold fixate on a superficial, easy-to-obtain quantity that is ultimately a red herring. In reality, model capacity and generalizability represent characteristics that are innately difficult to quantify and measure.' — Gupta et al, Data-Centric AI Governance: Addressing the limitations of model-focused policies (2024)

How do these narratives around catastrophic AI risk influence policy? There appears to be a disconnect between the high-profile, cautionary public letters warning about AI's potential to cause catastrophic harm - narratives that often dominate headlines - and the topics covered in AI ethics and governance courses, both in academia and within AI startup governance materials. Even within AI startup governance, explicit references to catastrophic risk are relatively limited. Despite this, this paper argues that such narratives may have implicit yet significant implications for the development of policy and regulation.

For instance, policy frameworks like the White House Executive Order and the European Union's AI regulations - specifically the use of compute thresholds 10^{26} (US) and 10^{25} (EU) - may reflect an indirect influence of catastrophic risk narratives. Although, as Hooker notes, how compute thresholds gained such widespread international and national traction over a relatively short timeframe is very difficult to track.

These thresholds aim to regulate large-scale AI systems, ostensibly in preparation for potential future risks. However, they also underscore an assumption that AI's emergent properties can be controlled or predicted based on specific computational milestones. These approaches seem to operate under the belief that once certain technical thresholds are crossed, AI systems will inherently reach critical or dangerous levels of

As Hooker aptly notes, “policymakers face a formidable task ahead of them.” This paper does not suggest that precautionary policies are inherently ineffective in addressing AI risks; rather, it argues that a more balanced, comprehensive approach is necessary – an approach that acknowledges the stakeholder narratives, power dynamics, and language that shape the way these risks are framed.

Considerable regulatory efforts are underway to address both current and evolving harms. The EU AI Act tackles critical aspects of AI governance, such as the risks of behavioural manipulation, social scoring, biometric identification for law enforcement, and the use of AI in hiring algorithms and insurance. In the US, the Fair Lending Act regulates AI-based lending decisions, while state and federal initiatives are developing to combat deepfake pornography. Although publishing agreements between content creators and AI companies are becoming more common, the scraping of artists’ work remains a significant point of contention. Moreover, the rise of GenAI poses risks of amplifying existing issues of technology and internet access, as language models trained on the dominant languages of the internet may further marginalize many communities.

Gupta cites the Stanford Internet Observatory (SIO) report, which highlights the significant presence of Child Sexual Abuse Material (CSAM) in the LAION[1]. Gupta evaluates the potential for established GenAI models trained on LAION-5B to be misused by malicious actors seeking to generate CSAM, with documented instances of abuse already on record[2]. His work underscores the need for robust data-oriented governance practices, as relying solely on post-training guardrails is insufficient. Gupta et al., suggest that AI governance frameworks can build on existing legal structures, which would help reduce regulatory overhead.

Conclusion

Precautionary AI policies should strike a balance between speculative foresight and pragmatic interventions. Latour’s critique of ‘freeze framing’ can be used to inform policy perspective and support dynamic, relational approaches that address both the grand and current systemic risks of AI. By shifting the focus from abstract catastrophic scenarios to actionable systemic harms, this paper advocates for a more grounded path forward in AI safety. More attention on data governance and basing policy actions on real-life examples are essential aspects of the development and design of AI-focused policies. Indeed, while the challenge ahead is complex and nuanced, policymakers cannot afford to rely on freeze frame ideas of existential AI risk.

[1] https://www.techpolicy.press/laion-and-the-challenges-of-preventing-ai-generated-csam/

[2] https://www.bbc.co.uk/news/articles/c170gr4n94wo