Towards Monitoring our Growing National Security Debt with AI: Tracking Humanity's Succession of Control to and Susceptibility to Influence from AI Systems

Gatlen Culp [*1] **Hamza Chaudhry** [2] **Herbie Bradley** [1] **Nandini Shiralkar** [1]

Abstract

Background. AI is becoming increasingly integrated into every aspect of our lives, institutions, and culture. While productive in the short term, this process risks gradually disempowering humanity, empowering AI-automated institutions that abandon human interests and leaving us vulnerable to a new oligarchy or destructive artificial general intelligence.

Action. This project develops metrics for what I term "AI-induced Civilizational Vulnerability"– the vulnerability arising from integration and the implicit power we place in AI models and those controlling them. I create measurements across five institutions (economic, cultural, educational, research, and political) examining critical effects: collective forgetting as AI masters human skills, AI manipulation of voters and consumers through digital content, ceremonial democracy as governance exceeds human comprehension, emotional dependencies on AI systems, and automated self-improvement pushing humans out of the loop. I aggregate these metrics into a broader vulnerability model, establish influence thresholds, and make policy recommendations. Where data gaps exist, I call for collection by AI labs and governments.

Impact. Like climate scientists making atmospheric changes visible through CO2 measurements, this project makes civilizational vulnerability measurable for policymakers. By establishing thresholds and trends, we transform abstract alignment concerns into urgent national security imperatives, enabling international coordination to prevent races to the bottom, and providing metrics for AI labs' safety cases and responsible scaling policies.

1. Introduction

AI is Growing. In recent years, there has been a boom in the domain of artificial intelligence, and in particular, a boom in the development of Large Language Models (LLMs) introduce statistics on the diffusion of AI models, possibly from AI as Normal Technology – a technology that boasts impressive capabilities in the domains of Software Engineering, Content Creation, and more expand this, perhaps stats. Such developments have sparked a silicon-valley craze over the development of an Artificial General Intelligence (AGI) quote "feel the AGI" – an AI capable of accomplishing the same tasks as a remote knowledge worker different definitions, maybe change, worth noting – or even an AI beyond human level intelligence ("Superintelligent AI" as I will write throughout), sparked national discussions over job displacement, received trillions(?) of dollars of investment, resulted in international headlines when DeepSeek, a Chinese AI company, released an AI model capable of competing with state-of-the-art LLMs coming out of American AI companies, and had figures like JD Vance call for the rapid development of AI at the Paris AI Summit cite. While there are great disagreements over the extent and timeframe AI changes society, there is little disagreement eh over whether AI is a transformative general purpose technology on the order of the industrial revolution.

AI is Integrated with our Institutions. AI is becoming increasingly embedded in American institutions should I keep the national framing? – economic, cultural/cognitive, education, research & development, and state. At the present moment, we see educators grappling with incorporating AI into the classroom and preventing cheating, software developers taking the day off if their AI-enabled tooling is offline, researchers and journalists raising alarm bells as to how AI could be used to manipulate elections, and grow-

[*]Equal contribution [1]ERA:AI Research Fellowship [2]Future of Life Institute. Correspondence to: Gatlen Culp <GatlenCulp@gmail.com>, Hamza Chaudhry <hamza@futureoflife.org>, Herbie Bradley <mail@herbiebradley.com>, Nandini Shiralkar <nandini@erafellowship.org>.

ing evidence towards dead-internet theory. As capabilities grow maybe introduce timelines? idk how much work to do on background and diffusion continues, the future could see collective forgetting as AI masters human skills, ceremonial democracy as automated governance exceeds human comprehension, emotional dependencies on AI systems, AI enabled military coups, and automated AI research and development pushes humans out of the loop of the same technology powering the modern world. Better examples, more likely and consequential ones also make these less negative.

Integrated AI is a National Security Risk. AI has the opportunity to improve corporate and government efficiency, develop new technologies, democratize education, and far more. With so many prospects, it is unsurprising that so many are keen to rapidly develop and implement this technology. And considering the power that comes with owning and exporting this technology, it's unsurprising that the United States has tried maintaining their grip on the technology and their lead over the China. However, the gradual integration of AI into our lives, institutions, and culture leaves humanity (and especially the nations first to embrace the technology) vulnerable to influence – hard and soft. This paper aims to introduce the idea of a growing national security debt that the pressures to develop AI create and propose some ways of tracking its development.

Definitions. *Security debt*, a term borrowed from Security Operations, refers to the accumulation of unresolved security issues over time. Much like technical debt, where developers prioritize quick solutions over more robust, long-term fixes. The longer security debt is left unaddressed, the more difficult and costly it becomes to resolve, potentially leading to severe security breaches and other vulnerabilities. In this paper, the security being referred to is *national security*, not to be confused with the specific application of AI in cybersecurity or defense TODO: better wording, which has a broader domain including the security of: values of the society, state sovereignty, constitutional human freedoms and rights, society and its relations, and more. Additionally, this paper focuses on the gradual weakening of institutional defenses due to AI[1] as opposed AI-enabled of offensive capabilities such as cyber-attacks or CBRN (Chemical, Biological, Radiological, Nuclear) weapons development link, although these may also share the description of being under-addressed national security issues.[2]

Goal of this Paper. This paper aims to introduce the concept the the Growing National Security Debt into AI & National Security discussions in US and international Governments, think tanks, and AI safety communities in addition to ways of measuring. However, I am uncertain about the fruitfulness of this direction of research as someone interested in catastrophic AI risks and as such, I include a Macrostrategy section for the AI safety community, ie: reasons this may or may not be a comparatively good line of research. The aim here is not to develop a comprehensive analysis, only to outline a direction of research. Probably wouldn't include this in the final paper if a final paper exists. This paper can be broken down into a few sections: In *section 2*, a broad explanation is given as to the mechanisms behind our accumulating national security debt, comparing institutional AI risk to others, and making the case that AI is a uniquely destabilizing technology. In *section 3*, I cover a few potential models for thinking about and measuring this debt. In *section 4*, I discuss, for the AI safety community, reasons for and against this line of research over various threat models and potential impacts this research may have.

[1] Institutional defenses have the opportunity to be strengthened by AI as well cite lock-in, applies, but maybe an odd paper cite AI as normal technology section. However, AI is more likely to be a disruptive force whose positive influences can only occur after addressing key flaws. Substantiate

[2] This paper also looks at the vulnerabilities and plausible threat of harm, not potential harms itself. The line here is fuzzy, but analogous to "this is a exploitable vulnerability in the economy" versus "this will hurt the economy." Distinction not clear, maybe not worth making. It's also not just "this is disruptive to our way of life", it's being disruptive in a dangerous and gameable way.

2. How & Why We are Accumulating National Security Debt

2.1. Primer on AI and Theoretical Risks

Framing AI. *AI* – a fuzzy domain roughly defined by developing computer systems able to simulate human decision making – is best viewed as a general purpose technology akin to steam, electricity, and information and communications technology (ICT) cite. As such, its application is pervasive amongst almost all sectors of life from the military to the workplace to education to home entertainment. In any one of these domains, AI has the opportunity to influence and effect the systems and individuals that they come into contact with, precisely because they were designed to simulate, augment, and/or substitute human decision making. And as such, ceding (implicitly or explicitly) our decision-making capacity on the microscopic level aggregates to ceding our national capacity for steering our institutions, culture, and way-of-life to automated systems on a macroscopic level. In the process, locking ourselves out of the decision-making processes for a number of reasons – competitive factors limiting oversight that would slow down development, inability to intervene on complexity over automated systems, interconnected automation for which small changes can lead to large collapses, technical debt that is too expensive or overwhelming to address, the cognitive decay/forgetting, illusions of control, cultural disinterest, etc. This sounds convincing but vague. Concrete examples would be very helpful.

Gradual Disempowerment. In the paper on Gradual Disempowerment cite. I lowkey want to quote the entire executive summary, they make some great claims, which formed the basis of the paper you are currently reading, the authors describe how incremental AI diffusion across economic, cultural, and state institutions can result in them becoming untethered from the human dependence and feedback that kept these institutions in-line with human welfare throughout history. One possible outcome of this gradual disempowerment being the formation of a fully self-sufficient non-human economy, leading to an *extinction by industrial dehumanization* from pollution, armed-conflict and/or resource depletion cite Critch. While this paper does not go as far as to claim this is our future, this example nonetheless highlights just how severe the consequences of automation can be. There is a strong need to monitor AI's level of integration and influence even at the intermediate states between today's world and the one just described.

Foundational Models as a Risk Amplifier. The influence AI systems have expands greatly when we consider the increasingly general AI systems that we see today, primarily Large Language Models (LLMs) like ChatGPT which are built on top of **foundational models**, of which frontier models take billions in USD to train [cite] – meaning the entire technology stack the modern AI revolutions depends on is enabled by just a handful of frontier models. While "AI" is confusingly used somewhat synonymously with these foundational models, it is important to distinguish between AI using foundational models and AI not using foundational models as most of the societal vulnerability introduced by AI is attributable to these models in particular, a topic explained later in "AI is Uniquely Problematic". I don't distinguish these throughout this paper (oops), especially in the example below. I should maybe fix this.. Critically, it's also this technology that not only amplifies national security vulnerabilities, but also makes them exploitable.

2.2. How AI enables fragility and influence

Automating Menial Labor. AI is often praised for its capacity to automate repetitive or menial labor such as truck driving or factory jobs, allowing these workers to move on to more fulfilling roles Wording feels classist here(?). While there are a number of physical safety and cybersecurity concerns here – incidents may be rare, relatively small, isolated and local, closely monitored, and more easily attributed to bugs or human error. In short – risks are transparent and manageable. Cite. Explain what I mean here. Also maybe people worry about terminator or robot armies or something – maybe worth addressing even if just to dismiss. Overall, automating menial labor neither seems to detract much from our civilizational capacity to decide our future nor does it introduce significant danger to national security. Is this distinction necessary? Do others agree it's not hugely important. This is largely in-line with the tool-view of AI,[3] which sees AI as a blank slate which mirrors the user's intent and loses its power once the user releases. However, this perspective dangerously underestimates both AI's potential for influence of its own and ability to accomplish tasks once believed to require "a human touch" Describe better or exclude.[4] The influence of AI systems expands as we develop and apply them to accomplish ever more important, complex, and open-ended

[3] The view of tools as neutral instruments that depend on the user is flawed in a number of ways. (a) Tools influence the ways humans think and act, AI is no different (ex: Maslow's Hammer – when you have a hammer, everything looks like a nail, phones, social media, etc.), (b) While claims like "this tool is bad" are ambiguous, tools can nonetheless be recognized for their ability to steer the future in one direction or another, (c) AI need not be conscious to display human-like characteristics or skills such as creativity. is this worth mentioning

[4] The fact this point is dogged relentlessly, yet frequently forgotten, points to just how the pervasive idea of tool-influence separation and human exceptionalism is. Prob better way to explain this? Also human exceptionalism might have religious or transhumanist undertones.

tasks, especially when humans are displaced in the process. See thought experiment below.

Congressional Staffer Scenario. Take the following thought experiment: A US congressional staffer that gradually incorporates AI into their day-to-day operations – first using DeepResearch to conduct literature reviews, then using it to manage all of their email communications, then to draft policy proposals, to press releases, to make value judgements on their actions, and so forth. This staffer is able to produce higher quality work at a faster rate. Not only is the AI becoming used across more diverse tasks but it's also being used at a greater depth Horizontal vs vertical integration(?). At first, she only uses AI to assist with her work – restructuring documents, grammatical changes, making text flow better, making recommendations, etc. Under time pressures, she used AI to write entire drafts, changing details manually or asking it to rewrite entire portions better framing her perspective. Over time, delegation creep progresses – her ability to work, think, or make arguments on her own atrophies, she defers to AI more often – its framing of situations, its perspectives, its sources – and sometimes not being able to distinguish her ideas from the AI's Cognitive dissonance leading her to thinking its ideas are her own or critique the AI's writing Something something automation bias / cognitive miser, cite MIT's your Brain on ChatGPT. She is praised by her managers for her efficiency and quality – feeling like a fraud, she wishes she could unravel her dependence, but with increasing time pressures and atrophied skills, can't find the same patience or elegance to complete the same tasks she used to, effectively becoming a liason between the AI and her colleagues. Eventually her colleagues start doing the same and the high quality, high speed work that distinguished her just becomes the status quo. No single person is willing to deviate from AI dependence for the fear of falling behind and being replaced by someone else who is AI-dependent. Perhaps full capture over lower-level congressional staffers occurs without management ever noticing. Perhaps they notice but aren't willing to walk back on the productivity gains – maybe for fear they fall behind other nations supercharged by AI, maybe because they're aware these staffers are no longer capable of the job. Perhaps they are AI-dependent themselves. Perhaps they embrace the changes. Maybe a narrow AI example is better(?)

Congressional Staffer Analysis. The situation above is clearly a national security issue. Every one of these millions of tasks being delegated to AI results in less power over our legislation and opportunities for AI biases[5] AI bias feels like a term that is too narrowly scoped. Also left-coded. to creep in, accumulate, and entrench itself in the foundational policies, relations, law, and even minds of our decision makers – potentially without anyone, including our elected officials, realizing. With further AI diffusion over other substantial governmental decision making processes, even a small share, made even more influential by their collective consistency[6] and AI-supercharged rhetorical abilities, say 40%(?) of legislative, executive, and judicial activities sus, it's possible that over a decade cite/propose to study we see a gradual convergence of government functioning and values towards "those the AI system prefers" (and as the AI influences citizens and government officials in a similar direction, it may be difficult attribute the true influence AI had on our decisions.)[7]

Reasons why AI influence could be costly. To give reason for concern here, foundational models have demonstrated preferences towards those of Chinese nationality, valuing them at x times someone of American nationality and y times someone of Z nationality, cite + confirm potentially leading to policy decisions favoring one ethnicity over another or conducting negotiations with a prejudice against Americans Maybe want to pull back on the "lets paint AI red-white-and-blue and feed it apple pie" implications, there's something here that is of concern to policy makers but I am not framing it right. Additionally, when foundational models were used to simulate various geopolitical actors in military simulations, they consistently opted for more unpredictable and escalatory behavior than their human counterparts, hinting that a government influenced by

[5]These AI biases might be (a) *planned* – developers or backdoor engineers cite who recognize the AI's influence and purposely alter it, perhaps *benignly* (ex: to make their philosophical, political, or religious views more dominant or influential – AI labs like OpenAI have teams dedicated to steering these values, typically to be more "neutral" or "helpful, harmless, and honest", but it is nonetheless impossible to avoid making a number of value judgements and models are known to have left-leaning biases cite.) or *maliciously* (ex: to purposefully destabilize the institutions AI becomes entrenched in). These biases may also be (b) *unplanned*, arising from quirks of the training process, either *undiscovered* – in which case they may be quite alien or uniquely harmful – or *ignored* – potentially due to time/budget constraints or because they, on a surface level, matched the developer's pre-existing biases. While the field of algorithmic bias and engineering ethics is rich, it will remain mostly unexplored here. Something to be said for the scope of biases in AI. Historically these were limited to harmful stereotypes in narrow AI systems. Asking what an object detection model thinks about the Israel-Palestine conflict or how the lessons of Confucius should be applied the design of North American cities or something like this make little sense. Also can be biases at many points assuming control. More about these biases in a later section.

[6]Humans, famously, disagree on most things. Also, this may require the same model (or foundational model) to be used. Different models may have different values.

[7]Would be cool to estimate just how fast the government could be automated, which parts, and what percentage could be automated without elected officials being aware of it (or at least aware of the scale.)

AI might be one inclined to violence and war cite Dennis' thing + confirm/substitute. Might be more anecdotal than empirical..

No one wanted this. In the congressional staffer scenario, it's important to highlight the coordination failure in the final outcome. Even if undesirable to all parties, there are a number of incentives pressuring each actor to acquire more national security debt. Our *staffer* wanted to become more productive, save energy, and meet deadlines, leading delegation creep to set in, their skills to atrophy, and dependence to increase alongside her expectations. Something here about "just don't do this" is not enough Reverting became extremely costly. Her *colleagues* likely felt these similar pressures, likely made worse by management's rising expectation of staffers and the looming threat of termination. The *management* perhaps didn't know about automation, couldn't afford to revert, didn't have meaningful methods to intervene, and personally benefitted from the automation. The *US Government* as a whole experiences pressures to become more productive, and even if aware of the automation, had no choice but to embrace or even accelerate AI diffusion at the risk of being outclasses by the more agile and effective governments of their geopolitical competitors, which also suffer in this scenario. highlighting the need for both global cooperation and internal oversight measures

Other Unilateral Bottom-Up Influence Scenarios. This bottom-up threat model where individual actors integrate AI against their long-term individual and collective interests isn't unique to this scenario. Similar instances are already becoming visible in all levels of education, one survey quoting a student from a prestigious American university wishing that they had never started using AI the way they did and feel trapped but having done so cite + confirm, leading children and the future generation of minds to be rooted in dependence on foundational models and deeply influenced by models. Sounds alarmist Given how widespread foundational models are in education, it would not be surprising if the scenario above described the situation of a few congressional staffers today. Entrepreneurs may not be able to compete in the marketplace without AI influence its investment decisions, ethical choices, communications, etc. cite anthropic. Human operated media content may not be able to compete with AI generated content, forcing its adoption. Likewise, searching for and consuming information may not be possible without AI filters and summarization. Each of these involve some level of unilateral decision making on behalf of everyone, first-movers kick-starting a series of pressures incentivizing adoption – its integration, even into government, required neither explicit approval from others nor exorbitant effort or expertise. Of course there are other ways that AI can integrate itself and influence others. Talk about these later. Currently tired. something something cognitive effect become accumulative. something something game theory nash equilibrium pareto optimality

With foundational models, a single foundational model can be more than just your legislation, it can simultaneously be your best friend, your diplomats, your bureaucracy, your engineers and scientists. If responsible development and adoption of foundational models is hard to prevent, their alignment becomes a matter of national security. Contrary to the mainstream narrative that being the lead state in the AI race enables economic prosperity, it may instead be the case that its integration captures and controls the same incubating country that it was meant to benefit, leading to national disempowerment. Probably move this elsewhere. Also change.

2.3. Comparisons of Institutional AI Risk to Others

In some ways, the risks AI elicits are not unique. AI as Normal Technology does a pretty decent job discussing how pretty narrow AI wouldn't be a wildly new challenge, pull more info from here In particular, arms races to develop and adopt technology to gain an economic or political edge and developing regulations to internalize externalities is a textbook economics problem.

In some ways, the embedded AI risks pose to *cultural institutions and cognitive autonomy* have also been posed by social media algorithms, mass media, internet, and personalized advertising especially over the past two decades, influencing not just public opinion but also legislation (such as car dependent infrastructure in the US kind of old, not best example) and elections. Arguably, the global political polarization, hostility, and rise of conspiracy theories have been a result of how the internet has developed and the economic incentives of businesses embedded in algorithms – which themselves have been shaped by small groups and other institutions, relatively unregulated. This section feels weakly related to my overall point, relate better. While these influences are hard to attribute and are scattered across many individual and organizational choices, singular top-down manipulation is also feasible. Following the acquisition of Twitter/X by Elon Musk, large changes were made to the platform and the content that gained popularity, which can be traced back to individual decisions Cite. Or just remove, contentious in political environment. About (90%?) of the platform stayed, despite (55%?) being against the changes cite – likely attributable to a kind of lock-in and a difficulty to migrate to other platforms (ex: creators have built their following on there, users accustomed to getting their news there, etc.). Regardless, this is

just a single platform and while there are negative effects, they are far from catastrophic. However, the influence over culture should not be underrated – the effects of media control have been widely studied, often with the perspective of state propaganda, a massive topic of the 20th century as it applies to the Soviet Union and Communist China, and the origin of the term brain washing okay now it feels like I'm being alarmist ooo 1984 spooky, kind of overused. Even today, technology has enabled China to influence minds using the Great Firewall, mass surveillance, and content/message censorship. I don't necessarily want to call out China either While AI can also enable authoritarian regimes cite and worsen existing societal issues cite, they can also create new vulnerabilities in liberal democracies maybe don't use these in a policy paper, jk but not really At this point, I haven't made much of the connection between AI and this technology, but I kind of want to move on. Maybe this should be its own section.

Likewise…

- MILITARY – Nuclear weapons / power (might be more apt) (much easier to develop/share, many other use cases, hard to control)
  - We have competition with other countries to develop, it is kind of military but also kind of a general purpose technology.
- Information and communications technology

Have to move on, a bit tired and running out of steam on this topic.

Other kinds of misalignment risks listed in gradual disempowerment.

2.4. AI is Uniquely Problematic. TODO

While AI shares many similarities to other technologies, a few properties make AI a particularly destabilizing technology: (a) the massive cross-domain scale of diffusion, (b) rapid development and diffusion, (c) centralization of control AI coup would say "singular loyalties", (d) non-detectable secret or misaligned loyalties mouthful, (e) potential for dependence (and human lock-out / loss of control?) TODO: Create better differences, be more principled. Also elaborate. Also separate foundational models from other versions of AI

Unlike financial debt (which elicits imagery of a credit card bill) or technical debt (which involves recognizing and not-implementing best practices), The ai-induced national security debt is more subtle and threatening in some key ways: (a) We don't know how much we owe (how costly is it to address?), (b) We don't know how to pay it (how to address it?), (c) We don't know what the consequences of not paying it are (how bad is the problem?). This paper attempts to address each of these. kind of vague, unnecessarily visual?

Rapid Development. There have been rapid improvements in the capabilities of foundational models without any fundamental insights into the nature of intelligence. Beyond the invention of the transformer and the Large Language Model, a considerable amount of improvements from GPT-2, a 2019 model barely able to string coherent sentences together, to 2025 models like o3, capable of automating tasks that would take software engineers over an hour and a half to complete, compose full essays, and achieve impressive scores on math benchmarks eh, not best examples, arise mostly from scaling existing practices and implementing intelligent but relatively mundane tricks, with incremental 1% improvements leading to exponential growth. A METR study indicates that the length of software engineering tasks foundational models can do is doubling every 7 months cite METR, which may lead to a future where AI automates or augments its own research and development, creating feedback loops boosting progress faster than is currently humanly capable. Also recent SWE uplift study by METR showing that AI enabled tooling actually makes developers slower. Also idk if worth mentioning intelligence explosion or "country of geniuses in a data center" Amodei. But even the PERCEPTION of it being helpful (even if harmful/incapable) is enough for it to diffuse. Diffusion is the right question here and development only helps so much in that it enables diffusion. There is trillions of dollars in investment going towards the development of this technology and lots of international attention. find source

Of course there are reasons for doubt – *benchmarks don't necessarily track the actual economic tasks* we are interested in and while capabilities have seemed to rocket ahead, adoption of this technology has been quick, but not proportional. cite part from AI as normal tech Nonetheless, it's reasonable to believe that adoption could lag behind capabilities as many institutions are slow to integrate even decades old technology from the start of the information age. While progress may eventually plateau, for the moment foundational models linger near human-level performance in many domains and it appears there is no slowing of cheap tricks that even junior researchers can pull to make improvements indicate that there may be enough low hanging fruit to sustain progress beyond human capabilities in most domains.

The speed of foundational model development has outstripped that of the industrial revolutions or nuclear energy, akin to developing the ENIAC computer in 1945 and having the internet and iPad by 1946 or the catapult in 399 BC and nuclear weapons by 398 BC. This rapid develop-

ment severely hampers our cultural and political ability to reflect upon and steer the usage and development of this technology. Alarmist?

"The real problem of humanity is the following: We have Paleolithic emotions, medieval institutions and godlike technology. And it is terrifically dangerous, and it is now approaching a point of crisis overall." — Edward O. Wilson

Quiet & Rapid Diffusion. Unlike the previous industrial and information revolutions, distribution of this technology doesn't require building large supply chains or factories to *scale the availability of models to meet global demand* sus. While running these models are computationally expensive compared to typical software applications, datacenters and GPUs remain mainly bottlenecks for the development and not the diffusion of foundational models. Once trained, foundational models can be run relatively cheaply, with less-powerful models able to be run on high-end consumer cards costing less than $2000 USD. Most of the infrastructure for global distribution is already here, and estimations for how fast we can meet any additional demands are on the order of a few years. cite

Unlike many other general purpose technologies, foundational models are *software not hardware* something here but not framed right, capable of being accessed from any computer. While it was mentioned above there is hardware involved in building out the supply of these models, these logistics are isolated from consumers and businesses working with these models. Whereas updating your business for the information age might have meant purchasing an expensive desktop computer and ripping holes in the wall to make way for ethernet cables and phone lines, the barrier to entry foundational models is extremely low and inexpensive since their operations take place elsewhere and are interacted with over the internet. Frontier foundational models can be accessed through chatbot interfaces over the internet in a matter of seconds, completely free of charge for low-quality models and as low as $20/month for state-of-the-art. Individuals and firms don't need large up-front investments to start or stop using these models.

These foundational models also *require little expertise* to use. By the nature of these models being designed as a drop-in solutions for most tasks, without any kind of tweaking, one requires little expertise in using these effectively. In five minutes, anyone could be taught to use a model like ChatGPT to operating their computer to producing photo-realistic videos, policy proposals, homework assignments, websites, and more. As interfaces grow more useable and the models powering these interfaces become more capable, even less effort will be required in getting AI to automate or augment one's thinking, workflow, education, or entertainment. In the information age, computers were powerful but they required dramatic changes in one's workflow, learning new skills and ways of doing things that were more effective. Developing additional applications for computers often required years of experience and a big paycheck – or consulting with another team of experts to get it built. Foundational models, on the other hand, offer the *ability to augment or automate existing workflows with little effort* (in addition to offering more effective ones).

For similar reasons, adoption *doesn't require collective buy-in or network effects*. In the information and communication age, it was necessarily the case that multiple had to agree on protocols of communication – sending emails or sharing online documents were useless unless everyone else you were working with agreed. Social media would not have been possible if not for the collective cultural decision to adopt not only the technology but the specific platforms using them, relying on network effects (and when individuals move on, these technologies die out – think MySpace for social media or how floppy disks are practically non-existent). The growth of social media as a technology was handicapped by this barrier – the idea of it existing as early as the 1960s, becoming technologically/economically feasible in 1990s, but arguably only coming into existence in the late 2000s cite. Because AI has the ability to augment or automate existing processes, it's diffusion not limited by slower moving cultural or institutional norms and the risk-averse tendencies that tend to accompany this oversight. It's this property that allowed our congressional staffer to automate their workflow without any prerequisite alterations to their working environment. This allows small groups and individuals to integrate AI at a local level without the approval of others this effects I realize I might be overusing the big technical words because I want people to take me seriously. Might be beneficial to use more creative writing. I think I'm talking about the information age too much, feel like broken record.

These factors together – massive supply, low barrier to access, low barrier to usage, and the ability to integrate AI in a local and isolated way – all result in the *ability for foundational models to be diffuse not just rapidly but also without notice or oversight*. Especially in domains where automating workflow becomes taboo, individuals avoid admitting they use it and to what extent. In the workplace or government, it may go undiscussed, only in shared in hushed tones and that get quieter as management nears. It's for this reason that bottom-up integration and influence becomes possible described previously. A majority of the US government could be automated with foundational models before elected officials even recognize the possibility. Dead Internet Theory suggests that much of today's

internet, particularly social media, is dominated by non-human activity and AI-generated content. AI might not just be used to flood the internet with entertainment, but also be used in the automation of influential content promoting values and perspectives – the misuse of AI's in spreading misinformation is being widely discussed, but what about less malicious influence? Ie: Well-intentioned journalists, influencers, bloggers, educational content creators, fact checkers, etc. Journalists in key media institutions including Wired (and the New York Times) report to using foundational models as a key component of their writing and research. cite / talk to friend We should find it concerning that we don't know to what extent foundational models are involved in the automation or augmentation of these roles nor the rhetorical steering this might have over our media and information environment. If it were true that 90% of the information we receive on a day-to-day basis were produced and/or influenced by foundational models, would we know? And could we do anything about it?

It's also the reason why, top-down oversight and control over our institutions becomes so difficult. Rules against unapproved usage would be a good start (ex: foundational models must not be used on work at this level of security clearance unless the models have also been), but might drive these activities further under-the-radar. Detecting AI augmentation and automation from human-produced work may be difficult, especially if widespread, since the body of human-generated work and sentiments to compare against diminishes. However, this might require distinguishing acceptable from unacceptable usage, and detecting not just when augmentation occurs but when it crosses an agreed upon line – not just across one task, but every task, implicit and explicit in the worker's job expectations. At an extreme level, maybe involves monitoring their AI involvement outside of the workplace, if a matter of national security. Perhaps move to policy recommendations section.

Potential for Dependence.

Wide Diffusion.

Centralization of Control.

Lack of Transparency / Opportunity for Hidden Loyalties. The "bitter lesson" "The nature of neural networks used in AI makes understanding how AI comes to its conclusion highly challenging. I recently read a 19 page research article that figured out how a single word was predicted."

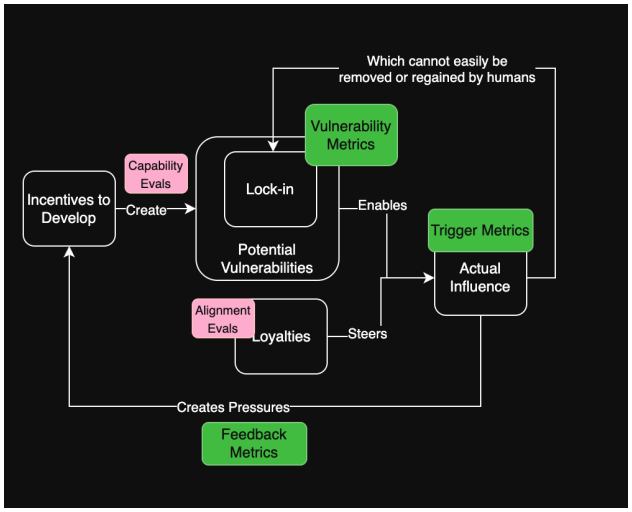

Figure 1. There are ways of developing societal-scale monitoring using vulnerability metrics and trigger metrics. (when vulnerabilities are leveraged.)

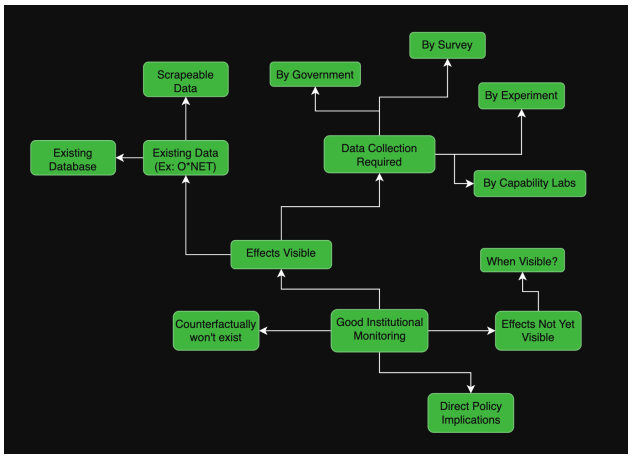

Figure 2. What makes a good institutional monitoring metric

3. Measuring our National Security Debt TODO

3.1. Why Measure? TODO

3.2. Spotting Vulnerability TODO

1. Spotting Vulnerability – Vulnerability / Trigger Framework (+ examples in culture, economy, state, education, military, etc.) (include my diagram and other notes)

3.3. Spotting Good Metrics TODO

1. Spotting Good Metrics –

3.4. Measuring TODO

1. Measuring – Where to get data, things to consider, etc.

4. Is this Direction Worthwhile? (Strategy For the AI Safety Community) TODO

4.1. Reasons for further research TODO

1. Imagining the best case policy scenario – International Monitoring Committee on AI Influence. Monitors keystone AI indiciators of institutional health (cultural, cognitive, economic, educational, r&d, state), our decreasing influence over them and the fragility of these institutions to coordinated influence by centralized powers (authoritarians, AI developers, power-hungry people, or even the AI itself.) Works together with cultural interest groups, the World Wide Web Consortium (W3C), etc. to make sure that AI isn't steering cultural discourse or converging values over the long-term as it gains greater control over the information economy, social media, personal relationships (AI-human relationships), published works (fed through/edited with models), and thought loops. Identifying which patterns of AI integration we have control over and which ones seem mostly unpreventable. Data used by technological standards institutions like NIST and ISO to set standards over what can/should be automated, to what degree, what amount of human oversight and monitoring is required. Placing societal responsibility on AI labs Reducing arms races by highlighting the internal harm AI integration can have. Setting international goalposts akin to the Paris Agreement to limit global warming to below 2 degrees celsius with concrete ways of achieving this. Making calls for states to slow down their integration. Detecting when media, political, or educational ecosystems have been compromised + manipulated. But also cognitive and individual interactions. Surveys.
    1. Maybe labs have to include these studies in their safety cases. Putting societal influences into their responsible scaling policies and internally conducted research.
    2. Theory of Change (Impactful)
        1. Make better benchmarks
        2. Shared across multiple threat models (Measuring the degree to which we are losing control over our institutions) (maybe have some Venn Diagrams or Something, a table checking off which parts are relevant)
            1. Gradual Disempowerment (How can we)
            2. AI-Enabled Coups (Ex: Military coups, cultural coups, etc.) (How can we prevent centralizing power in the hands of a few people? if these are inevitable, how can we detect when these powers are being leveraged?)
            3. Gradual Takeover (How do we prevent or slow-down the worlds where humanity cedes power to misaligned AGI? As a civilization, can we put aside individual first-mover advantages to adoption in favor of preserving human influence?)
        3. Requires less assumptions about AI capabilities
        4. Better framing (ex: Huawei being a national security issue in the United States)
        5. Make metrics visible
        6. Helps with forecasting (effects of legislation, geopolitical effects, etc.)
    3. Somewhat robust policy implications shared across multiple threat models
        1. Gradual Disempowerment
        2. Gradual Takeover
        3. AI-Enabled Power Centralization
    4. Visible now or visible soon (good to measure)
    5. Urgent to collect them (certain bits of data might be lost over time, could be influential)
    6. Domains that may be overlooked by other interest groups that don't have an AI safety framing (counterfactually would not exist)
        1. Which domains are currently looked at (ex: Unemployment)
        2. Which domains are currently not looked at (ex:)

4.2. Reasons against further research TODO

1. God-like AI takeover threat model – Changes don't matter in this case, our future is already locked-in
2. Speed
    1. Capabilities might develop so fast, that by the time meaningful trends show up in the data we will have much bigger fish to fry
    2. Policy implications would take too long to implement

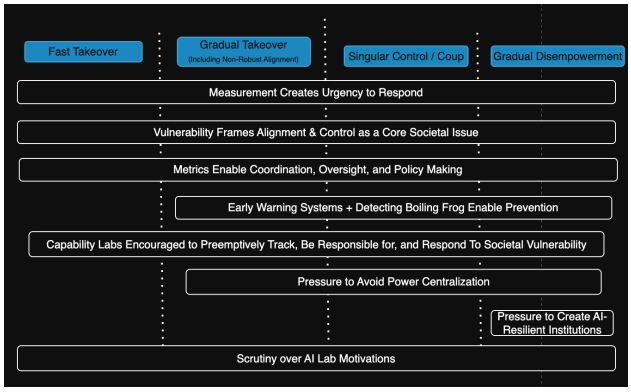

Figure 3. Assuming there is enough time to gather enough data on these metrics, the implications of developing societal-scale monitoring are robust to different threat models

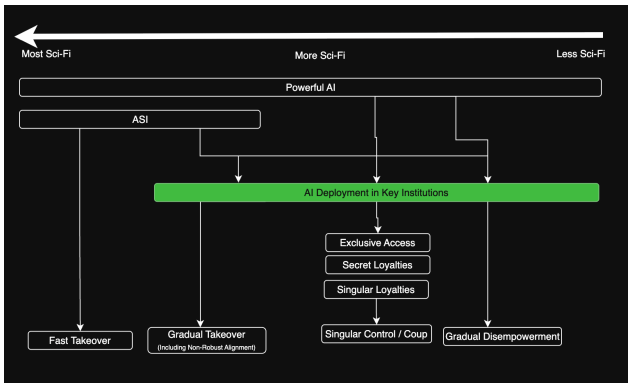

Figure 4. AI Deployment in Key Institutions is Common Across Multiple Threat Models, including ones that are far less sci-fi.

3. Care
    1. People may not care about the metrics
    2. Policy implications might be band-aid solutions in some threat models (ex: societal resilience of)
    3. Might result in calls to develop AI defense mechanisms (possibly counterproductive)
    4. Might call for GREATER proliferation of AI rather than controlled distribution.
    5. These might already be so visible and obvious, such that the metrics are redundant
4. Redundant
    1. These might be developed by other institutions and getting them earlier isn't any more impactful
    2. Post-deployment data analysis might be fine-grained and informative enough for developing and improving benchmarks.
5. Logistic
    1. Too expensive/burdensome to collect
    2. Issue with this model overall: Potentially too broad. More likely that there couldn't even exist a single monitoring body, a book could be written about the influence in any of these domains. (ex: Influence on elections, etc.) Each of these fields would likely need to be developed individually and in depth. However, there's value in seeing these indicators as a collective rather than siloed. In the same way there's a strong relationship with capability evals.

References