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. How AI opens paths to influence

AI – a fuzzy domain 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 or cultural decay over influencing these changes, etc. This sounds convincing but vague. Concrete examples may be helpful.

This influence 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. Most of the societal vulnerability introduced by AI is attributable to these models in particular, a topic explained later in "AI is Uniquely Problematic". While these systems are distinct

2.2. Comparisons of Institutional AI Risk to Others

In some ways, AI and the risks it 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.

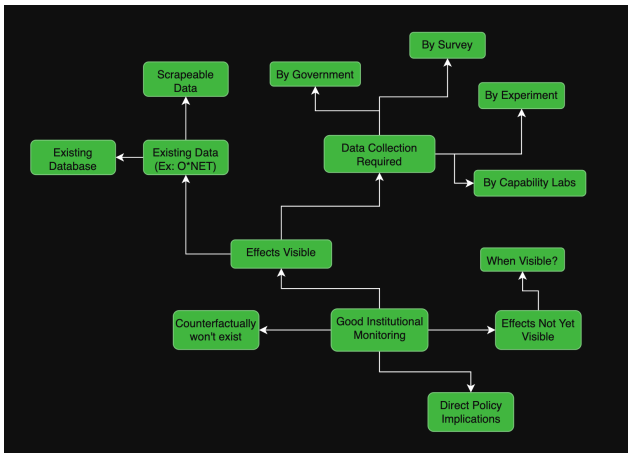

Figure 1. What makes a good institutional monitoring metric

2.3. AI is Uniquely Problematic.

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.

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?

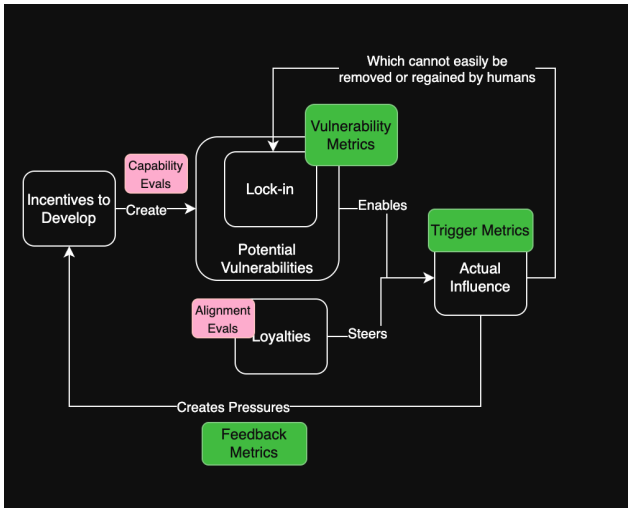

Figure 2. There are ways of developing societal-scale monitoring using vulnerability metrics and trigger metrics. (when vulnerabilities are leveraged.)

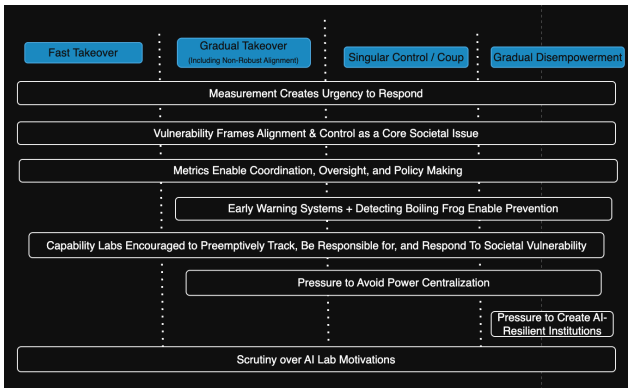

Figure 3. Assuming there is enough time to gather enough data on these metrics, the implications of developing societal-scale monitoring are robust to different threat models

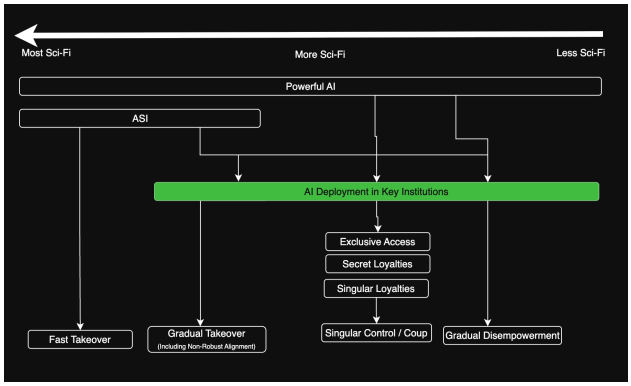

Figure 4. AI Deployment in Key Institutions is Common Across Multiple Threat Models, including ones that are far less sci-fi.

3. Outline

Kind of like societal impacts team of UK AISI: "Societal impacts: evaluating the direct impact of advanced AI systems on both individuals and society—including the extent to which people are affected by interacting with such systems, as well as the types of tasks AI systems are being used for in both private and professional contexts. Chris Summerfield, Oxford University's Cognitive Neuroscience Professor"

This is a strategy piece for the AI safety community, might be developed to be more later.

0. Goal of paper overall – Not to develop specific societal scale indiciators or even "introduce them as a new thing", but instead to frame them in the context of AI safety and identify their use cases and a potential methodology of developing them. This may be a blindspot for the AI safety community, which may focus too much on in-lab capabilities and alignment metrics (ex: measuring bias in the lab) of models rather than the downstream slow-moving civilizational effects. These effects are likely influential to policy makers and the public, already present in a meaningful way as LLMs have, likely won't be developed by other interest groups especially in the domains highlighting value misalignment in the wild and power centralization.
1. What is societal scale monitoring
    1. Extension of pre-deployment and post-deployment metrics
        1. How post-deployment tends to effect the development of models
        2. Measuring the relationship between benchmarks, post-deployment studies, and actual societal effects
            1. Gap between these benchmarks and actually accomplishing the required tasks (capabilities tests fail to measure the outward effects to some extent. Allows us to gather some feedback on whether our benchmarks are accurate)
    2. Possible classes of monitors
        1. Vulnerability (How much influence does AI introduce in this domain?)
        2. Trigger/Influence (Is the vulnerability above being leveraged?)
        3. Human Lock-out (How does AI integration eliminate the power humans have to change this vulnerable regime?)
        4. Integration Reinforcement (How strong are the pressures to increase the integration of AI)
        5. Cross-Domain Influence (How does this influence other metrics you care about? Ex: Culture influencing politics)
        6. Power Consolidation (How are these vulnerabilities concentrated in a single source of)
    3. Existing research in this domain
        1. UK AISI societal impacts
        2. Anthropic Economic Index (ish, more post-deployment monitoring)
        3. Research into AI information economy, influence over elections and jazz
        4. idk, others.
    4. Structure of the paper
2. Why might this be important (Impactful, Tractable, Neglected)
    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.

2. Maybe labs have to include these studies in their safety cases. Putting societal influences into their responsible scaling policies and internally conducted research.
3. 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.)
4. Somewhat robust policy implications shared across multiple threat models
    1. Gradual Disempowerment
    2. Gradual Takeover
    3. AI-Enabled Power Centralization
    5. Visible now or visible soon (good to measure)
    6. Urgent to collect them (certain bits of data might be lost over time, could be influential)
    7. 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:)
3. Why might they not be important
    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
    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.

4. How can we develop them
    1. Spotting Vulnerability – Vulnerability / Trigger Framework (+ examples in culture, economy, state, education, military, etc.) (include my diagram and other notes)
    2. Spotting Good Metrics –
    3. Measuring – Where to get data, things to consider, etc.
5. Further work
    1. I'm applying this framework to measuring automated AI R&D, may be updated later with what I find.
6. Conclusion

4. Outline 3 (Security Debt)

1. Intro
    1. Background (AI is growing, AI is integrated)
    2. AI is a National Security Risks
    3. Defining National Security Debt
    4. Goal of the paper
    5. How to Read this Paper
2. How and Why is National Security Debt being Accumulated?
    1. Decomposing National Security
    2. Applying various AI threat models to National Security feels contrived
    3. Comparisons with other Institutional Risks
    4. Why AI is uniquely disruptive
3. Measuring our National Security Debt
    1. Spotting Vulnerability – Vulnerability / Trigger Framework (+ examples in culture, economy, state, education, military, etc.) (include my diagram and other notes)
    2. Spotting Good Metrics –
    3. Measuring – Where to get data, things to consider, etc.
4. (AIS Community) Reasons for further research
    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.
    2. Maybe labs have to include these studies in their safety cases. Putting societal influences into their responsible scaling policies and internally conducted research.
    3. 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.)
4. Somewhat robust policy implications shared across multiple threat models
  1. Gradual Disempowerment
  2. Gradual Takeover
  3. AI-Enabled Power Centralization
5. Visible now or visible soon (good to measure)
6. Urgent to collect them (certain bits of data might be lost over time, could be influential)
7. 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:)

5. (AIS Community) Reasons against further research
  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
  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.
6. Further Work & Research Agenda
7. Conclusion

References