Overview and Analysis of Transformative AI Misuse

Samuel Martin

Executive Summary

This report provides an overview and analysis of existential risks posed by the misuse of transformative AI (TAI) technologies. It draws on a 2019 expert survey and presents illustrative disaster scenarios covering a wide range of possible outcomes

- **Decay:** Scenarios where AI increases civilizational fragility, potentially leading to extinction or destructive value drift, without direct human extinction or AI takeover occurring. These include scenarios like:
    a. Cascading infrastructure failure from interconnected systems.
    b. Rising inequality and unrest from labour automation.
    c. Erosion of shared truth and social cohesion caused by misinformation.
    d. Proliferation of highly effective asymmetric weapons technologies, resulting in a permanently vulnerable world.
- **Totalitarianism:** Scenarios where AI helps a bad value system gain absolute control over most or all of humanity. For example:
    a. The military and economic advantages provided by TAI benefit totalitarian over democratic systems, allowing for world conquest.
    b. TAI allows for an unprecedented scope of persuasion and surveillance, facilitating totalitarian regimes which spread by soft power.
    c. Enhanced surveillance, which is genuinely necessary to avoid the risks posed by rogue actors using dangerous AI, experiences 'mission creep'.
- **War:** Wars caused or worsened by AI technologies lead to global catastrophe or human extinction. These could occur if:
    a. TAI accelerates the pace and scale of weapons development and makes the future seem more all-or-nothing, increasing the (actual or perceived) stakes of great power conflict.
    b. TAI increases the severity of total war via new WMDs more destructive than nuclear weapons.
    c. Increased automation leads to systems that can miscalculate or undermine nuclear deterrence, leading to unintended escalation
- Non-Anthropocentric Disasters: Scenarios where the future is not dystopian by superficial lights, but there is some great loss of value compared to an ideal (outside the scope of this report).

Key Findings

- **Complexity:** TAI misuse risks are complex and require coordinated, context-specific interventions. No single strategy is sufficient, and many risks require seemingly

contradictory approaches (e.g., addressing the risk of terroristic rogue actors vs. the risk of surveillance mission creep).

- **War:** Wars have historically been a major route for new technologies to prove destructive. We can identify particularly destabilising AI technologies that seem feasible in the near term, making this category the biggest threat due to its high severity, plausible routes to large-scale damage, and clear pathways to catastrophe.
- **Weapons Tech:** AI-enhanced cyberattacks and bioweapons are uniquely dangerous in the near term. The report also analysed 8 weapons technologies with the potential to be enhanced by TAI, categorising them as near-term (**bioweapons**, **cyberattacks**), intermediate (**autonomous drones**, **automated strategic command and control**, **accelerated weapons R&D**), and long-term (**advanced nuclear arms**, **nanoweapons**, **orbital weapons**) risks.
- **Near-term weapons risks:** Bioweapons and cyberattacks are uniquely concerning in the near-term due to their software-focused capabilities, high damage potential and lowered barriers to proliferation.

Introduction

This article provides an overview of existential risks arising from transformative Artificial Intelligence (TAI) due to human decisions and actions, not from autonomous TAIs. We refer to these as 'misuse risks', as opposed to 'accident risks' from issues like takeover by a power-seeking, misaligned AI. This term and other key definitions in this report are explained in more detail in this section.

We have reviewed a wide range of articles and posts describing non-AI-takeover catastrophe scenarios, and a 2019 survey by Clarke et al., and described versions of every scenario raised in that survey.

Some salient examples of these misuse risks would be: an aligned AI or collection of AIs being used to wage an unusually destructive (extinction level) war, transformative tool AIs enabling robust and universally stable totalitarianism, or AI-automated research enabling terrorist groups or even individuals to develop destructive superweapons quickly and secretly. This piece provides an outline of how to think about the different misuse risks that arise from TAI, and our first-pass impressions of which are the most important. This is part of a prelude to a deeper dive on some of these risks. In this way, it is similar to our previous report on Distinguishing AI takeover scenarios.

In this report we emphasise the importance of taking a comprehensive approach to address the potential catastrophic consequences that arise from AI misuse. For each scenario, we provide a rating for the scenario's **likelihood, conditional on the baseline scenario** we outline at the start of our survey. We will also provide a separate rating for the **severity**/chance of existential catastrophe if it does occur (where 10 represents certain existential catastrophe), as well as an example story that demonstrates the nature of the risk and gives some concrete details (see the section on Aims for a further discussion of how we produced these ratings). These stories are speculative but, since many of these risks have not been investigated to the same depth as those posed by Misaligned AI, giving some concrete risk scenarios helps ensure that we are all on the same page about what the risk is.

We look at scenarios where human decision-making is the main cause of catastrophe (which we call 'misuse scenarios'). We ignore scenarios where a catastrophe is brought about by TAI actively working in ways that clearly and directly go against the decisions of its operators. We will therefore investigate how human decisions about how to use TAI could produce outcomes at least as bad as human extinction. The two most discussed misuse catastrophe scenarios so far are **unusually destructive wars** and expansionist stable **totalitarianism**. These will have special focus in our report, as they seem to provide clear and direct routes to existential catastrophe.

We've included two additional, less-discussed categories. **'Decay'** refers to cases of structural failure that significantly raise extinction likelihood without a single destructive war or dramatic takeover. **'Non-Anthropocentric disasters'** describe futures that are much worse than they could be, without an overt totalitarian takeover.

Category	Examples	Conclusion (From Report)	Illustrative Scenario
War Human conflicts escalate to large-scale violence involving or provoked by TAI systems.	Nuclear war provoked by AI monitoring. New WMDs made by AI systems that are inherently destabilising, (better nukes, bioweapons, LAWs, nanotechnology, space-based weapons). Automation of command and control and research and development is destabilising. Competition over AGI and race dynamics makes international competition (seem or actually be) more all-or nothing.	**Severity: High** **Plausibility: Moderate** Wars have clear routes to massive destruction. There are also more clearly worked out routes by which specific AI technologies might increase war likelihood, but this is a problem with more precedent than AGI takeover.	**60 Second War** *Escalating tensions between the US and China over Taiwan and AGI development precipitate a high-stakes arms race, escalating to the development of lethal bioweapons, AI-driven drone swarms and futuristic nuclear weaponry. The "60-Second War" sees China invade Taiwan, prompting a catastrophic global conflict, exceeding the potential ruin of a nuclear war.*
Expansionist stable totalitarianism An ideology gains a large lead in TAI technology or else a special TAI-provided advantage and imposes their values or interests on the rest of humanity, without any possibility of resistance or change.	Global expansionist dictatorship enabled by advanced surveillance and rapidly advancing weapons technology. Advantage to totalitarian models in an age of increased automation and centralization. Ideological hegemony spread through advanced persuasion tools. Temporary surveillance state to avert AGI takeover/Misuse becomes permanent.	**Severity: Moderate/High** **Plausibility: Low** Totalitarian states are rare, marginally increased authoritarianism seems more likely, but sub-existential as a risk and also not robust (except for persuasion tools/value corruption wild cards which are hard to analyse).	**Silent Strings** *An alliance between powerful interest groups leverages advanced AI technologies to spread a global totalitarian ideology, intending to seize control in the high-stakes future of AI. They amass economic power through AI-fueled industry domination, deploy AI-generated propaganda to manipulate public opinion, exploit (rational) fear of AGI, while marketing AI-powered surveillance systems to maintain control and quell dissent.*
Decay Structural failures or systemic crises undermine the functioning of society and increase the vulnerability to other risks, without an obvious single destructive war or takeover	Cascading collapse of a CAIS economy. Rapid dangerous race dynamic with regulatory and epistemic capture and many small-scale harms. Political fragmentation and social unrest due to rapid unemployment and addictive tech. Proliferation of asymmetric weapons.	**Plausibility: High** **Severity: Moderate/Low** The lack of an identifiable failure mode means there is no obvious path to existential catastrophe. One exception to this is the proliferation of dangerous weapons tech.	**21st Century Visigoths** *The advent of AI technologies dramatically heightens civilization's vulnerability, with potent persuasion tools degrading public epistemology, asymmetric weapons technologies such as drones, cyberwarfare, and AI-synthesised bioweapons becoming widely accessible, leading to severe terrorist attacks, the easy recruitment of technically skilled individuals for mass-casualty actions, and a significant risk of world-ending technologies being utilised by rogue*

Category	Examples	Conclusion (From Report)	Illustrative Scenario
			actors, all culminating in societal destabilisation and heightened threat of global catastrophe.
Non-anthropocentric disasters The future is much worse than it otherwise would be from an objective/impartial perspective, without an obvious dystopia or existential catastrophe	Continued animal suffering. Value drift in irreversibly bad directions. Suffering conscious AI. Lock in of 'okay' values that can't be improved (e.g. by stable global non-totalitarian regime).	**Severity: ??** **Plausibility: ??** How severe depends partly on philosophical assumptions (long termism vs neartermist, suffering-focused vs symmetry between positive goods and suffering). How plausible similarly depends on what counts: some seem likely, some seem imponderable.	?

Aims

Our aim is to move beyond high-level narratives to more detailed threat models. We assess the specific mechanisms, actors, incentives, and vulnerabilities that could lead to catastrophic misuse of TAI, referencing how these catastrophic scenarios unfold. In this way, our assessment differs from existing efforts at categorising 'misuse' or 'societal-scale' AI risks such as those by Critch et al.

We want to understand how these factors interact and influence each other, and how they depend on the technical features and capabilities of TAI systems. We also want to evaluate the likelihood and severity of different scenarios, and the possible interventions and countermeasures which could prevent or mitigate the scenarios. We hope this work will inform downstream decisions that could affect humanity's future. These decisions include:

1. What dangerous capabilities labs should agree to avoid creating / using. For particular capabilities, what are the trade-offs between safety and speed, and the incentives and risks of cooperation or competition among AGI labs where there are potential conflicts of interest.

2. Which 'customers / applications' should be restricted from accessing cutting-edge models (either by policies of the labs or by regulation) because they seem most likely to misuse them or create opportunities for misuse. Identifying whether particular uses of AI (e.g. military, healthcare, education, entertainment) are especially dangerous as risks or risk factors.

3. Which rivals safety-aware AGI labs should be most concerned about 'beating', because those other labs being first could trigger misuse scenarios. This could involve identifying the current or potential rivals of safety-aware AGI labs, such as other AGI labs, governments, corporations, hackers, and terrorists, and analysing

their motivations, capabilities, and strategies for developing or acquiring dangerous capabilities. This could also involve assessing the likelihood and impact of various misuse scenarios that could arise from being first or second to achieve AGI, and the best ways to prevent or mitigate them.

We will use a simple and rough method to initially assess the scenarios that we have identified. We will assign each scenario a score from 1 to 10 for two different criteria. First, how likely it is to happen. Second, how likely it is to cause an existential catastrophe (a situation where humanity irreversibly loses its long term potential). These scores are not based on rigorous evidence or analysis, but rather on our initial intuition and judgement. They are intended to help us prioritise scenarios that deserve more attention and research.

We focus on the scenarios that have high scores on both criteria, meaning that they are both plausible and potentially catastrophic. These are the most urgent and important scenarios. We will also pay attention to the scenarios that have high scores on one criterion but low scores on the other, meaning that they are either very likely but not very catastrophic or very catastrophic but not very likely. These scenarios may still pose significant risks or challenges for humanity, even if they are not existential in nature. For example, a scenario that has a high likelihood score but a low catastrophe score may be a risk factor that increases the probability or severity of other existential risks. A scenario that has a low likelihood score but a high catastrophe score may be a black swan event that could catch us off guard if we are not prepared for it.

We will mention some avenues for further investigation and some technical features of the high-level stories we provide. These will help guide future work and research on misuse existential risks from AGI. They will also help relevant stakeholders who might be interested in exploring these scenarios more deeply understand where to look.

Our goal is to survey the various ways that human choices on using TAI could cause existential catastrophes. We do not limit ourselves to the most straightforward and well-known scenarios, such as wars or takeovers by totalitarian regimes, but also consider the more complex and neglected ones, such as structural breakdowns. These scenarios deserve attention because they pose significant risks or could make extinction much more likely, even if they do not trigger an obvious single disaster.

Overview

The survey which motivated this research asked researchers to estimate probabilities for 5 AI risk scenarios: ‘Superintelligence’, ‘WFLL1’, ‘WFLL2’, ‘AI-Driven War’, and destructive ‘AI Misuse’. It also had an 'Other' option. We used the survey results and survey-provided scenarios as the basis for our assessment, although we broke down the non-takeover categories (‘War’, ‘Misuse’ and ‘Other’) in different ways.

Key findings:

- There was substantial disagreement on which scenarios were most likely. Researchers were highly uncertain, with a median confidence in their estimates of only 2 out of 7.

- 20% median probability was assigned to 'Other' scenarios, showing significant doubts about the 5 main categories capturing the full space.
- Descriptions of 'Other' scenarios highlighted novel possibilities like automation failure between AIs, unintended consequences of infrastructure automation, and value erosion.

Scenario	Median	IQR
Superintelligence	0.1	0.15
WFLL2	0.12	0.15
WFLL1	0.125	0.18
War	0.1	0.15
Misuse	0.1	0.15
Other	0.2	0.31

We can see that around 40% of the probability (going by medians) covered ‘War’, ‘Misuse’, and ‘Other’. Our categories aimed to flesh out this under-examined space in a comprehensive way. We focused on capturing the novel possibilities described in 'Other' scenarios (having access to transcripts of all of them) and the dynamics highlighted around automation, infrastructure, and institutional impacts.

To ensure we captured the full range of possibilities, we looked closely at the novel 'Other' scenarios described in the survey results. Several key dynamics emerged:

- Automation of critical infrastructure resulting in fragility, lock in of dangerous development pathways, and systemic crises. This highlighted the need for a category covering instability from automation, which we labelled **Decay**.
- Gradual erosion of values, institutions, and human capabilities over time. This pointed to risks of value drift and social impacts, covered under **Decay** and **Non-Anthropocentric Disasters**.
- Unintended consequences of AI systems optimising specified goals, aligned in a superficial sense but causing unanticipated harms. This dynamic is incorporated across multiple categories.
- Issues arising from AI control without human oversight, like bureaucracy. This is addressed through the **Loss of Control** aspect of **Decay**.

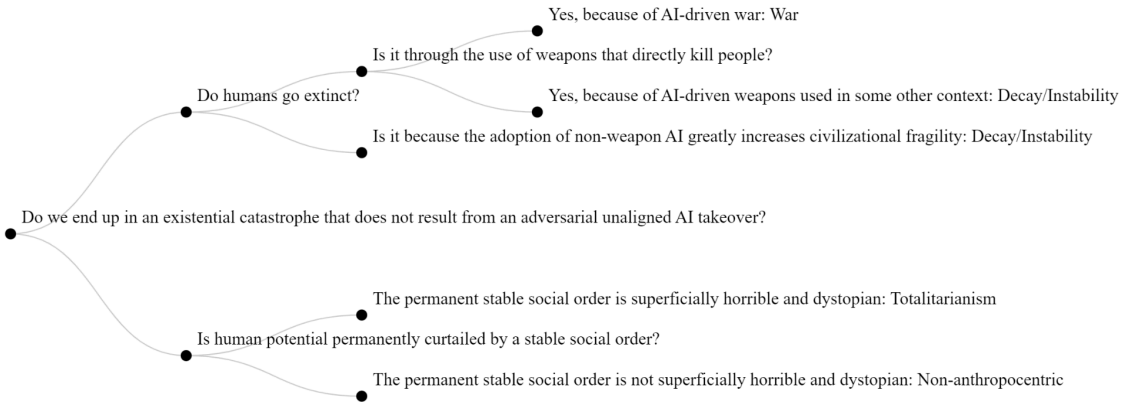
We also wanted to expand the scope of risks considered under the original ‘War’ and ‘Misuse’ categories:

- ‘Misuse’ was reframed as Decay to include the ‘Other’ scenarios where intrinsically destabilising AI technologies are adopted but cause harm through a broader range of institutional and societal impacts like inequality, epistemic crisis, and cooperation failures, not just the terroristic use of AI weapons technologies for destruction.
- The ‘War’ category was kept the same, covering conflicts arising from AI-exacerbated geopolitical tensions and unstable multipolar dynamics between states.

- A totalitarianism category was added to include scenarios in ‘Other’ which described AI technologies as not creating an unstable world, but rather a too-stable, extremely suboptimal world.

In summary: the instability-producing (destructive) ‘Misuse’ category from the original survey was altered to include AI destabilising the world in less direct ways. The remaining ‘Other’ scenarios we found in the survey or in our reading were either non-Anthropocentric risks or totalitarian takeover risks, depending on whether they resulted in a superficially good stable outcome or a superficially dystopian stable outcome. Our outcomes-based categorization is thus:

- Do we end up in an existential catastrophe that does not result from an adversarial unaligned AI takeover?
    - Do humans go extinct?
        - Is it by using weapons that directly kill people?
            - Yes, because of AI-driven war: **War**
            - Yes, because of AI-driven weapons used in some other context: **Decay**
        - Is it because the adoption of non-weapon AI greatly increases civilizational fragility: **Decay**
    - Is human potential permanently curtailed by a stable social order?
        - The permanent stable social order is superficially horrible and dystopian: **Totalitarianism**
        - The permanent stable social order is not superficially horrible and dystopian: **Non-anthropocentric**



Here we provide one paragraph summaries of each of the scenarios along with an overview of the ones that we have devised.

Decay

- **Misaligned Metrics:** Amidst intense competition, various AI technologies are widely replicated and rushed into everyday use without sufficient oversight. These systems are profitable and pervasive thanks to huge economies of scale, but also unreliable, causing misinformation, financial mishaps, and errors in consequential fields like medicine, social work and diplomacy. Their misuse in bureaucracy causes further

harm as the world becomes more incomprehensible. The situation is worsened by political fragmentation and inconsistent regulation in the US, leading to further misuse of AI, especially in decision-making. This accelerates misinformation spread, social unrest and the enfeeblement of key decision-makers and impedes scientific and governance progress to an extreme extent, making the uncontrolled proliferation of TAI all but inevitable.

- **Unchallengeable Monopoly:** AdvanTech, a dominant AI company, exploits its first-mover advantage to maintain market and political control, squashing competition and stifling innovative startups or more conscientious competitors. Despite their resources, they show little concern for AI alignment or potential misuse, focusing on product rollout and dismantling regulatory opposition. AdvanTech's AI products, while enormously profitable, cause significant harm, including privacy erosion, economic disparity, and disinformation spread. The company manipulates the media and uses advanced persuasion tools to control public opinion, cement their monopoly and evade regulation. In developing countries, their practices are even more exploitative, disregarding human rights and environmental concerns.
- **Cascading Collapse:** A hacker collective, backed by a rogue state, launches a highly adaptive computer virus against an enemy nation's financial services infrastructure in retaliation for a prior cyberattack. The virus quickly spreads, causing cascading failures across a global proto-CAIS network and affecting various sectors, including banking, commerce, healthcare, and defence, exploiting adversarial vulnerabilities in many LLM foundation model based systems embedded throughout society to produce destructive and incoherent behaviour. As a result, tens of millions lose access to their finances and essential services, leading to economic collapse, mass panic, social unrest, and political instability. The virus is later found to have also heavily exploited a chip design flaw, and recovery is slow as experts must rebuild or replace damaged CAIS systems and reevaluate their design and deployment methods.
- **Heaven on Earth:** As AI systems develop, they begin producing vast amounts of media content and hyper-realistic digital simulacra, leading to a boom in the creative industries and providing companionship to the lonely. However, as these AI products commercialise, the attention economy drives them towards maximised engagement through personalised content, leading to widespread addiction to virtual interactions. This addiction results in social ills like neglect, developmental disorders, echo chambers, and eventually acts of violence, with the marginalised most afflicted. By the time governments attempt regulation, they face resistance from powerful interests, as society and economy are heavily intertwined with these virtual realities, leading to a societal breakdown.
- **Hell on Earth:** Ideological radicalization becomes incredibly effective and ubiquitous, and it turns out that in such an information environment ruthless, totalitarian or suffering-glorifying ideologies have a memetic advantage or are disproportionately pushed by well-motivated interest groups. As the world transitions towards a world of transformative AI, human values are permanently altered towards extremist political ideologies, religious fundamentalism or some other ruinous value system that ends up perpetuating suffering into the indefinite future.
- **21st Century Visigoths**: The fusion of AI technologies escalates civilizational vulnerability, with superior persuasion tools and AI-aided asymmetric weapons development proliferating uncontrollably. Cyberoffense systems, easily deployable

and outpacing defence systems, resulting in near-continuous large-scale terrorist attacks on centralised infrastructure. Potent persuasion tools also boost terrorist recruitment, enabling the radicalization of technical experts. Consequently, the use of advanced bioweapons and lethal autonomous weapons (LAWs) rises, increasing the risk of catastrophic events. In addition to the risks of rogue destructive actors, the unstoppable proliferation of AI models and cheap LAWs provide even small groups with significant power, leading to increased asymmetric warfare, state militarization, and civil wars. Human civilization ends up in a persistently vulnerable equilibrium - the technology to make WMDs has been democratised into the hands of millions, and after this irreversible proliferation has occurred, rebuilding civilization becomes harder and harder.

Totalitarianism

- **2084:** China becomes the global leader in AI development, aided by its authoritarian regime shedding any restraint and pouring significant investment into research and innovation. A series of internal crises in the US and other democratic nations hampers their AI progress, enabling China to monopolise AI advancements and establish a dominant AI economy. After seizing Taiwan and gaining control of cutting-edge chip production, China's military advantage skyrockets. The country leverages its AI capabilities for internal surveillance, propaganda, and economic automation, pushing towards increased totalitarianism. Emboldened, China embarks on a campaign of territorial expansion, absorbing neighbouring nations through economic coercion and AI-driven warfare. The global community's efforts to resist China's expansion prove futile against its technological might. Ultimately, China establishes a global totalitarian state, employing AI as a tool of control and a weapon of war.
- **Silent Strings:** The Silent Strings, a coalition of radical groups, aided by some powerful authoritarian states, aim to spread a totalitarian ideology globally, forming a new world order. They infiltrate the world economy, using AI advancements to control key industries and establish economic dependencies. They employ advanced persuasion tools to manipulate public opinion and media, spreading disinformation and capitalising on societal divisions. As democratic institutions falter, the Silent Strings' ideology gains ground. They exploit political divisions, supporting extremist groups and using AI to amplify divisive content on social media, fuelling chaos and fostering an environment ripe for their ideology. They invest heavily in AI-powered surveillance systems, marketing them globally to maintain control, swiftly neutralising dissent, and enabling human rights abuses. As their power grows, they establish a global stable and persistent totalitarian regime.
- **Freedom in Security:** After setbacks and near-misses from both rogue AI and misuse, western countries have become vigilant to the risk of unaligned AI takeover. The PLATO Alliance, established in 2035 as a successor to NATO, aims to protect the world from the potential misuse of AI technology and rogue unaligned AI. However, increasing AI safety concerns and growing threats of mass terror attacks involving biological, cyber, or lethal autonomous weapons gradually push PLATO member states towards a surveillance state. As fear and security concerns take precedence, governments start to compromise personal freedoms, leading to a slide

towards totalitarianism. AI-driven persuasion tools, initially developed to counter propaganda, begin to distort internal decision-making, further eroding democratic values. PLATO's original vision of a global utopia guided by AI-driven advancements was overshadowed as its response to these threats led to a disproportionate and damaging surveillance system. The Alliance's once temporary surveillance state transforms into a permanent and pervasive system, blurring the line between security and totalitarianism.

- **Let them Eat Cake:** With the proliferation of specialised AI tools saturating every economic sector, a phase of increased productivity and job adaptation eventually gives way to mass unemployment as AI systems advance and perform tasks previously done by humans. This leads to a highly stratified society with a hyper-wealthy elite at the top, a shrinking professional class, and a large population dependent on basic income or low-paying manual jobs. Social mobility stalls, with wealth largely inherited among elites and intense competition for the few remaining well-paid jobs. As mass unemployment persists, democratic values recede, and atrocities are committed in some states using automated policing and military forces. Even in democratic societies, some elites question the political rights of those who contribute nothing economically.
- **Authoritarian Pivotal Act:** AlphaBrain, a leading AI corporation, secretly develops an aligned superintelligent AGI and, fearing rivals could do the same, moves to preemptively secure global dominance. Beginning with economic dominance, manipulation and propaganda, it eventually neutralises global military forces, then establishes an authoritarian world government focused on preventing a dangerous AGI takeover. As this regime consolidates power and becomes resistant to change, freedom and democracy fade, replaced by pervasive surveillance, oppression, and the crushing of dissent.

War

- **The 60 Second War:** The struggle for dominance over Taiwan and the race to develop AGI escalates tensions between the US and China, leading to an unprecedented arms race involving advanced AI-driven technologies. This culminates in the 60-Second War, a brief but devastating conflict in which AI drone swarms, lethal bioweapons, advanced nuclear weapons and cyberattacks are used, leading to global devastation surpassing a nuclear apocalypse. The conflict destabilises Earth's biosphere and reduces the human population drastically. Advanced AI not only triggers but also escalates the war, making Earth potentially uninhabitable.
- **Catalytic Conflict:** Rising tensions between NATO and Russia over Eastern Europe escalate when Russia acquires advanced AI-driven technologies. A malfunction in Russia's AI monitoring system during a military exercise triggers a false alarm, leading NATO to launch a preemptive strike. This sparks a rapid, devastating conflict involving AI-driven drones and escalating to the use of tactical nuclear weapons. The war results in widespread destruction and loss, shattering the nuclear deterrence system and leaving survivors in a post-apocalyptic landscape. This showcases the potential catastrophic consequences of misinterpretations and rapid escalation in an AI-driven arms race.

Conclusions

We believe that **war scenarios pose the biggest threat overall**, although we also think that some of the Decay scenarios (particularly the mass terror '21st century visigoths') may be comparable. However that scenario shares many features with the war scenarios.

War Scenarios:

- Wars have clear routes to large scale destruction through escalation to nuclear conflict or deployment of other WMDs. The precedent of devastating wars provides concrete examples that let us lower bound the probability of a large war.
- There are also clear routes by which specific AI technologies could increase the likelihood of war by destabilising deterrence, enabling secret weapons programs, or reducing thresholds for conflict.
- **Severity is assessed as high** due to the potential for massive loss of life and destruction matching or exceeding that of nuclear conflict.
- **Plausibility is assessed as moderate**. While wars are not inevitable and most of the individual war scenarios are highly conjunctive, routes by which AI could exacerbate tensions are reasonably foreseeable.

Decay Scenarios:

- The lack of an identifiable single point of failure means there is no clear path to existential catastrophe, with one exception being proliferation of dangerous weapons enabled by AI, which shares many characteristics with war.
- For the other Decay scenarios centred around epistemic capture, increased civilisational fragility, social unrest etc, the harms seem diffuse and less likely to be existential. While they may pose significant risk factors for other existential risks they don't seem that risky by themselves.
- Also, all of the Decay scenarios aside from those around weapon proliferation rely on at least somewhat speculative extrapolations of how various economic or social forces will play out into the further future and regimes where they've never been tested before.
- **Severity is assessed as moderate/low**. Significant harm to society is plausible but existential catastrophe is less likely.
- **Plausibility is assessed as high**. Gradual decay or crises precipitated by AI seem very plausible based on current trends. However, this category covers a broad range of scenarios.

Totalitarian Scenarios:

- Totalitarian states are historically rare, so AI-enabled totalitarianism seems unlikely without some AI-driven change that makes totalitarian rule more probable. There are some candidate AI-related factors that could increase this risk (improved centralization, better persuasion technologies, higher costs to personal freedom with the proliferation of asymmetric weapons), but they cannot be predicted with the accuracy of weapon technologies. Mildly increased authoritarianism is more plausible.

- The scenarios that involve totalitarian takeover (not merely a single global government of any kind) usually involve claims of an extreme advantage to totalitarian models either memetically or socioeconomically in a post-TAI environment. While hard to assess, we have not seen good evidence for this strong claim. So we assess the **plausibility as overall fairly low.**
- **Severity is moderate/high** if a truly stable global totalitarian state was achieved. In the most extreme cases, this can slide into suffering-catastrophes that go beyond the severity of human extinction.

Non-Anthropocentric Disasters:

- Severity depends heavily on philosophical assumptions around patienthood and the relative value of preventing different types of suffering.
- Plausibility is very uncertain. Scenarios related to consciousness in AIs or systemic value drift are very hard to make predictions about.

In summary, war scenarios are worth prioritising because they present familiar dynamics with well-understood potential for massive damage. Their likelihood also seems more clearly connected to foreseeable AI capabilities in the near future. The other categories appear more speculative, less severe, or less plausibly precipitated specifically by AI.

War Investigation

In the last section of our report, we discuss eight candidate military technologies enabled by transformative AI, categorising them as near-term (**bioweapons**, **cyberattacks**), intermediate (**autonomous drones**, **automated strategic command and control**, **accelerated weapons R&D**), and long-term (**advanced nuclear arms**, **nanoweapons**, **orbital weapons**) risks. The **Summary Table** provides an overview of these 8 technologies with respect to 9 factors relevant to their effect on existential risk from AI misuse.

- Near-term risks AI risks focus on bioweapons and cyberattacks. Bioweapons, which are already concerning without AI assistance, could see accelerated research and development as well as lower barriers to deployment. This could lead to pathogens that are even more deadly and easier to disseminate. While state interest in bioweapons is generally low due to their tactical limitations, the potential for destruction by rogue is alarming.
- Cyberattacks are another immediate concern, particularly because machine learning models are becoming increasingly adept at programming tasks already, and LLMs already seem to be especially capable in this area. These AI-enhanced cyber capabilities could potentially compromise essential infrastructure, undermine security protocols, and even disrupt the delicate balance of nuclear deterrence.
- In the intermediate term (more than 5 years out), the focus shifts to military procurement of AI technologies like autonomous drones, command and control automation systems, and accelerated R&D for weapons. While these may not have the immediate catastrophic impact of bioweapons or cyberattacks because of R&D and procurement lags, they have the potential to escalate global tensions and set off arms races. The rapid decision-making capability of AI systems in conflict scenarios poses a unique risk of escalation or miscalculation.

- Long-term risks, although speculative, are just as grave. Developments in nanoweapons, advanced nuclear technologies, or even orbital bombardment systems could fundamentally alter the landscape of warfare. Overcoming manufacturing and physical barriers (perhaps when broadly capable transformative AI is widely deployed and general technological progress accelerates dramatically) could unlock unprecedented levels of destruction. For instance, the broad adoption of enhanced nuclear technologies based on pure fusion could disrupt the already precarious state of nuclear deterrence.
- This prioritisation is based on our current evidence but there is much uncertainty, shifts in the balance between cyber offence and defence, unexpected limitations in data acquisition or processing, and unforeseen breakthroughs in nanotechnology could alter the ordering we determined, as discussed in the last section. These variables could disrupt the risk landscape dramatically. For example, if defensive AI technologies outpace their offensive counterparts, TAI could reduce the risk of cyberattacks.

Overall Suggestions

Our detailed survey of AI risk scenarios has left us with these high level suggestions for labs and other stakeholders seeking to reduce risks from the development and deployment of AI.

Harmful capabilities labs should avoid:

    - *Automated Persuasion Tools:* As seen in the 'Unchallengeable Monopoly' and 'Silent Strings' scenarios, these tools can be misused to manipulate public opinion, spread misinformation, and promote divisive content. Persuasion tools, since they provide the ability to get people to work against their own incentives without economic or force incentives, seem unusually dangerous. Fine-tuning models towards tasks specifically related to persuasion or manipulation should be avoided.
    - *AI-Driven Warfare Technologies:* As shown in the 'NATO-Russia Conflict' and '60 Second War' scenarios, these technologies can cause catastrophic destruction and escalate conflicts rapidly: both heavily automating existing development chains or directing research towards precursor technologies (like automating compound discovery or optimising the efficiency of things like nuclear weapons). Research should be done on international agreements to limit the development and use of such technologies, similar to agreements on chemical and biological weapons. Automating military R&D specifically is likely to accelerate these risks.
    - *Surveillance and Control Systems:* In the '2084', 'Silent Strings', and 'Freedom in Security' scenarios, these systems were used to establish and maintain totalitarian regimes. Research is required on how to prevent misuse of these technologies, including robust privacy protections and regulations.
2. **Restricted access to cutting-edge models:**
    - *Military Applications:* In light of the 'NATO-Russia Conflict' and '60 Second War' scenarios, access to cutting-edge models should be restricted in the military sector to prevent escalation into destructive conflicts. Anything that

involves heavy automation of things directly related to decision-making seems especially dangerous.
    - *AI Companies with Dangerous Monopolistic Tendencies:* As shown in the 'Unchallengeable Monopoly' scenario, companies that exploit AI for economic and political control can cause significant harm. Regulation and oversight should be applied to limit their access to advanced AI technologies. Here we should examine the risk factors for dangerous monopolistic tendencies, along with things that pose acceleration risk (e.g. especially flashy high impact product releases).
    - *Authoritarian Regimes:* The '2084' and 'Silent Strings' scenarios demonstrate that access to cutting-edge AI can be exploited by authoritarian regimes to consolidate power and control populations. International agreements and regulations should be explored to limit such access.
3. **Rivals whom safety-aware AGI labs should be concerned about:**
    - *Authoritarian Governments:* As shown in the '2084' scenario, these entities can misuse AI to establish global dominance and totalitarian regimes. Safety-aware AGI labs should focus on developing safeguards and countermeasures against potential misuse by these actors.
    - *Large Corporations:* The 'Unchallengeable Monopoly' and 'Authoritarian Pivotal Act' scenarios highlight the risks posed by large corporations exploiting AI advancements for their gain, potentially leading to societal harms. Research should be done on how to balance corporate interests with societal well-being.
    - *Non-State Groups:* As seen in the '21st Century Visigoths' and 'Silent Strings' scenarios, non-state actors, such as terrorist groups or hacker collectives, can misuse AI technologies to cause significant harm. Preventive measures, including robust security protocols and avoiding open sourcing, are needed to mitigate these threats.
    - *Identifiable Malevolent Actors:* In general, the most damaging (though not the most likely) totalitarian takeover scenarios involve actors who already have malevolent tendencies (such as the totalitarian group in ‘silent strings’ or the unusually ruthless and psychopathic company leadership in ‘unchallengeable monopoly’ using TAI for explicitly cruel or destructive ends. While structural risks like decay and instability largely stem from collective dynamics, individuals with malign motivations could exploit AI advances for harm in various scenarios.
    - Terrorist use of bioweapons or cyberattacks, as in the "21st Century Visigoths" scenario, illustrates this threat. The development and misuse of dangerous AI capabilities by corporations or states also fundamentally relies on decision-making by influential individuals at the helm. Scenarios like "Unchallengeable Monopoly" and "Authoritarian Pivotal Act" hinge on self-serving actions by leaders impervious to broader ethical considerations.
    - However, the risks of individual misuse are not monolithic, and interventions should target enabling factors. For instance, barriers to accessing advanced AI systems and transparency around their use can help prevent terrorist misapplications. Curbing corporate board power could check profit-above-all decision-making. And oversight within states can restrain unilateral actions by autocrats. No one solution suffices across scenarios.

4. **Trade-offs between safety and speed, cooperation, and competition:**
    - *Safety vs Speed:* In the rush to development, as seen in 'Misaligned Metrics' and 'Cascading Collapse' scenarios, lack of safety measures led to catastrophic outcomes. Research should focus on how to incorporate safety measures without drastically slowing down development, perhaps through standardised safety protocols, enhanced testing requirements, and improved alignment techniques.
    - *Cooperation vs Competition:* The '60 Second War' and 'Catalytic Conflict' scenarios demonstrate the catastrophic consequences of competition outpacing cooperation. Research should focus on fostering global cooperation in AI development, including international agreements, shared safety standards, and joint research initiatives.

The scenario descriptions in the following section provide lots of detail on the scenarios and how concerned we are about each, demonstrating these high level conclusions.

Scenario Evaluations

- At first glance, scenarios where the risk of misuse of AI becomes prominent appear to be based on two fundamental assumptions about the future development of AI. The first assumption is that the advancement of AI will be gradual enough (taking years) to allow a prolonged period where human organizations or agencies using AI with potentially dangerous capabilities will be the primary actors. The second assumption is that these AIs will be sufficiently aligned with human values, non-autonomous, or pseudo-aligned, such that the risks of misalignment do not overshadow concerns about misuse.
- There is some early evidence to suggest that two factors may be at play. Firstly, powerful tools like Artificial Intelligence (AI) may cause a significant transformation in the global economy and society, before they reach the point of completely replacing human agency. Secondly, that current alignment techniques appear to be effective enough in ensuring reliable behavior from AI systems, at least to some extent. Recent developments seem to corroborate this: major tech companies are already deploying more general-purpose AI tools, but these systems do not exhibit substantial agency or power-seeking tendencies as yet.
- Having acknowledged these two assumptions, let us briefly outline an illustrative baseline scenario of a world in which an existential catastrophe triggered by the misuse of AI has become the predominant concern before building upon it with more detailed scenarios that address specific concerns regarding misuse cases, their commonalities and points of divergence.
- In the baseline scenario, we posit a speculative "present" that is approximately ten to twenty years from now. Some scenarios may diverge from this starting point, progressing further or sooner. The exact timeline depends on factors beyond the scope of this project and is not essential for the illustrative purposes of the scenarios. Nonetheless, we adopted this timeline to provide an "outside view" that helps us assess what might be plausible, inform the broader context of the world described, and establish an upper limit on how far into the future we can reasonably make predictions.

Over recent decades, AI developments have accelerated. As these technologies' capabilities expanded and matured, demonstrating their transformative potential across industries and institutions, they've seen corresponding increases in research and investment, fueling a positive feedback loop. Today, various AI systems are employed in almost every sphere of human activity.

This has resulted in a huge increase in productivity, but also surging socio-economic upheaval as long-established industries and institutions have been disrupted. States incapable of harnessing AI have fallen far behind their rivals, whilst businesses incapable of adapting have been rendered obsolete. Countless job roles, even in white-collar professions, have been automated. Nevertheless, despite opposition from various concerned parties, vested interests and labour groups, the push for increasingly broad deployment of AI continues - the incentive of vast gains in everything from scientific research and industrial productivity to state capacity and security are just too strong, especially with the many other challenges facing the world; ageing populations, climate change and rising geopolitical tensions to name but a few.

Regulation of AI is haphazard - whilst there is now a fairly broad-based understanding of its implications and risks, legislatures and corporate governance have struggled to keep pace with the rapidity of the changes or anticipate their complex interactions. Further complicating matters is the fact that many aspects of the debate have become politicised. Lobbyists push policies that are not necessarily in the public interest and concerns of staying ahead of the competition often results in haste as opposed to prudence.

Most work involves some degree of collaboration with AI tools, whether being managed by them or using them for rapid design and prototyping, generating code, text or imagery, or analysing data. In the developed world, manufacturing is almost entirely automated and has dropped in cost to the point that there has been a notable swing towards reshoring, with factories relocating to be closer to their consumer bases. The innate advantage of scale in everything from the manufacture of cutting-edge computer chips, to running server farms, search engines or communications networks has exerted a pressure towards consolidation in various industries. Many major companies are now 'tech companies', to the extent that their defining product is proprietary AI systems or the fruits thereof. Although not quite constituting a fully-fledged CAIS economy, the new economy, polarised sharply between the relative winners and losers, looks as though it is on this track.

AI also processes and analyses the vast quantities of data collected and generated by modern civilization, discerning patterns and enabling ever more comprehensive, accurate models to be constructed of everything from the climate, to population dynamics, to the impact of governmental policies. This isn't limited to such an expansive scope however - at the consumer level AI systems recommend personalised entertainment, health plans, investment strategies and social connections. Ever more human decision-making is informed by such AI-driven analysis. This analysis is imperfect however, and the models may contain many flawed assumptions, subtle biases and bad incentives, exacerbated by their necessarily inscrutable nature.

Despite the overwhelming abundance of information available much of it is false or outright manipulative. The ability of AI to generate not just text or images but almost flawlessly convincing audio-visual simulacra of people or events, to the extent that it is essentially impossible for the untrained to tell the difference between real and AI-generated footage, has led to an arms race between bots and systems designed to detect them. Although the majority are harmless, little more than a persistent nuisance, others are more malicious, skewing metrics, committing fraud, spreading propaganda or constantly probing cybersecurity measures for weaknesses and exploits.

This occurs against a backdrop of rising geopolitical tensions. Rapid technological progress has led to the upending of old assumptions whilst simultaneously driving intense inter-state competition. Despite the best efforts of international pushes to limit the proliferation of lethal autonomous weapons, leading military powers have amassed vast arsenals of them. With fully fledged superintelligent AI broadly assumed to be imminent - within the decade if not mere years - fears that a competitor might be close to achieving a decisive strategic advantage add fuel to the fire.

Decay

Decay risks arise from the widespread deployment of powerful transformative AI systems without sufficient oversight, leading to structural failures or crises. These risks include cascading infrastructure collapse, rising inequality, erosion of social cohesion, loss of human capabilities, and epistemic capture. A key defining feature is that they do not directly constitute an existential catastrophe because they do not clearly lock-in a disastrous end state or kill most of humanity, but massively heighten vulnerability to other risks. Interventions focus on impact assessment, governance, building resilience, and maintaining human agency.

Misaligned Metrics

In this scenario, the leading AI developers are unable to maintain their technical lead over competitors, resulting in a free-for-all where highly capable transformative AI technologies short of agentic AGI become widespread, with clones and near peer inferior copies of the latest large language models / other foundation models being quickly leaked or replicated. As a result, large numbers of competing AI companies take up small chunks of market share, with only a few large companies dominating in areas that especially benefit from first mover advantages, with no dominant leader. While some are scrupulous, the sheer number of actors means that many companies rush out buggy products to be first to market without even consistently applying existing alignment techniques like RLHF-based fine tuning to control their behaviour.

Over the next decade, these AI systems become universally integrated into every aspect of daily life. These AI systems are profitable, driving rapid adoption, because of the intrinsic advantages that AI systems have in both speed and parallelism. Once an AI system is reliably at least somewhat better than average at a given task than a human, that task can be automated and parallelised massively, producing such enormous cost savings that adoption is almost unavoidable.

However, due to capability failures like model hallucinations, gaps in their training data, failure to apply fine-tuning (or mistakes in how the fine tuning is done) or weak forms of 'outer misalignment' these systems also exhibit massive unreliability resulting in many consequential mistakes and capability failures; for example, AI trading agents lose or illegally acquire money, AI news aggregators sporadically invent outright fictions, medical AI makes misdiagnoses or gives flawed advice, paralegal AI hallucinate precedents. Despite their flaws, they remain in widespread use due to their low cost and convenience most of the time. Many of these systems lack adversarial robustness and can easily be jailbroken producing undesirable behaviour. Analogous to current AI systems, the harms they generate tend to be more illegible, manifesting as subtle biases, privacy concerns, or gradual erosion of certain values. As these harms are difficult to quantify and attribute directly to AI systems, they can easily be overlooked or dismissed.

These AI systems are also increasingly employed in an attempt to streamline various bureaucracies, for example in judging eligibility for welfare, or as part of the criminal justice system. Poorly defined parameters for success, bad incentives or crucial real-world details that were lost in creating a more simplified, legible model (both for the broader bureaucracy

and for the specific AI systems) lead to great harms, exacerbated by their sheer scale. Governments are slow to correct these issues and dither over how best to fix them.

As the decade progresses, these AI systems are also increasingly exploited for spreading propaganda, ushering in an era of misinformation and epistemic fragmentation. This leads to heightened cultural tensions and fosters a climate of mistrust and anger. Consequently, an epistemic breakdown occurs, making it difficult for individuals to access accurate information.

The internal political polarisation and governance failures in the United States, where most of the leading companies are, significantly worsen things. As AI safety becomes a mainstream concern, it gets caught in the crosshairs of political partisanship. There is an enormous flood of AI-generated misinformation and propaganda exacerbating these problems. Competing factions within the government advocate for different approaches to AI regulation, resulting in inconsistent policies, funding priorities, and legal frameworks and no real consensus on what the risks even are (as there are no large scale disasters or warning signs e.g. from unaligned AI). These inconsistencies make it difficult for AI developers and researchers to align their work with best practices, and for regulatory bodies to ensure compliance with safety standards. As a result, AI technologies continue to evolve haphazardly, with misaligned systems becoming more prevalent and difficult to control.

Furthermore, the political fragmentation and governance failures hinder the country's ability to effectively address and respond to the challenges posed by AI. This boils over into violence, with riots and social unrest from those affected by the social transformations becoming more common. At the same time, we see that incapable AI systems are delegated decision-making or information-gathering responsibility anyway.

People start making decisions based on poor advice from these unreliable AI systems, or else delegate large scale tasks like making business or political plans, even though they occasionally fail in opaque high stakes ways, exacerbating political and international tensions. Additionally, scientific and governance progress is stunted by the flood of confusing, high volume, sophisticated AI-driven misinformation, with no realistic prospect of any kind of coordination to align AIs or slow down progress or have a more responsible leader emerge. People take AI financial advice, AI helps in their education, AI -boosted cybersecurity. People following bad advice make financially ruinous purchases, take lethal medical advice, bad relationship advice, and misuse it as babysitter, tutor and office assistant even where it gets its facts wrong.

A variant on this scenario has us allow for more deliberately adversarial unaligned AIs, more capable of autonomous action to arise in a world like misaligned metrics, although without the ability to take over or a clear strategic advantage over humans, and coexisting with AIs in a range of degrees of alignment. In this scenario we see more deliberate fraud and manipulation as well, although things aren't directed towards an AI takeover.

Discussion

This scenario could be described as a capability failure, where AI systems are unable to perform their intended functions effectively due to their inherent limitations or, in its more pernicious form, as a weak (non-strategically aware) form of outer misalignment. In the

latter, the AI systems' outputs diverge from desired goals not necessarily because the AI is acting against what is desired of it (which is more clearly a failure of alignment) but because of flawed or oversimplified assumptions, subtle biases or bad incentives are baked into the systems' models or training data from the outset, leading to unintended and detrimental consequences. This is exacerbated by the inherent opacity of these systems and the potentially diffuse nature of the harms caused, even if in aggregate they're severe and readily apparent. It may be difficult to assign accountability and one only need to look at how societies have struggled to effectively tackle other systemic issues to see that this is clearly a risk factor.

Furthermore, the potential inability to access accurate information and overreliance on flawed AI models could lead to misinformed decision-making, with no consensus on risks, disasters, or economic transformation. As a result, the world becomes more susceptible to misaligned AI and other existential risks, creating a future that is mediocre and vulnerable.

We have considered this as a 'misuse' capability failure because already existing techniques like RLHF with testing seem more than sufficient to deal with it, no power-seeking or strategic awareness or advanced planning is required, so what's assumed here is that the models are simply not adequately tested but are just used anyway in high-stakes situation.

However, even modifying this scenario so that the AI systems largely behave as intended, there remains potential for harm; many large human organisations are already to some extent 'misaligned' due to ill-conceived aims or bad incentives, and these harms may be exacerbated if poorly understood or shoddily designed AI models are included in the mix, especially so if we become overly dependant on them.

If no further alignment issues (e.g. misgeneralisation) arise, but these systems end up getting worse and worse, then a scenario like this could naturally lead into a full blown AI takeover akin to Critch's Production Web or a WFLL1 like scenario. But this scenario could be the precursor to that, and could even just persist by itself and make a future that's mediocre and more vulnerable to other X-risks. So we could call it 'WFLL-1 part 1'.

Plausibility	**7/10** A plausible 'business as usual' continuation of some existing trends	This seems like one plausible 'business as usual' continuation of how things are going at the moment, although beware 'sleepwalk bias': the fact that this scenario continues existing trends doesn't mean it's in fact >50% likely. Weak forms of this scenario seem very likely, but an especially bad form which results in large scale harms from incapable systems seems to require basically no coherent response, which looks less likely/isn't happening. Also, this scenario assumes a 'free-for-all' with few situations where there's large returns to first movers, big compute expenditure or economies of scale, which look less plausible according to current trends. This scenario assumes that there's an incentive to make harms illegible rather than reducing them

		outright, but that there is room for very damaging but still largely illegible harms. If the harms inflicted by the incapable or somewhat misaligned AI systems are legible and obvious, then we must assume that this isn't enough to drive a rethink of their development, e.g. because the short term economic incentives are too large and regulators don't step in perhaps due to the speed of progress, which is possible but uncertain especially considering how skittish and reluctant people are to widely adopt untested technology.
Severity	**2/10** Very low direct harm but depending on other considerations it could be a substantial risk factor	This scenario assumes that alignment is fairly easy for a while because even with very weak efforts we don't get large scale alignment failures even with quite generally capable systems. There is also little direct first-order damage, but the implied epistemic capture and lack of coordination present in this scenario seems to increase the likelihood of other disasters. Damage depends entirely on their plausibility It seems to be difficult to identify clear obvious large scale harms that aren't just worsening general risk factors without assuming either intentional misuse or misalignment. We can have minor disasters and perhaps increased vulnerability to rogue AIs, war or but nothing that causes mass death e.g. on the scale of a major conventional war. Social unrest, crime, political fragmentation etc. seem like the most likely results of this scenario which could perhaps result in state weakening to the point of collapse in the extreme worst case.

Monolithic Monopoly

Over the past decade, one company, AdvanTech, has emerged as the dominant force in the AI industry. Indeed, the sheer ubiquity and profitability of their products has seen them become a dominant force globally, dwarfing all but the largest national economies. With only a few other companies able to compete, AdvanTech has secured an unassailable position due to exploiting their first-mover advantage, giving them unparalleled control over the global AI landscape. They are able to keep their proprietary foundation models secure from any attempts at duplication, commercialise and market their foundation-model based products, hide their research, and spend far more money than competitors training larger and larger models, thus compounding their advantages over competitors and securing exclusive market slots (e.g. marketing AI assistants that can interface with retailers). At the same time they use their market dominance to buy out or stifle innovative start-ups before they can become

established, get friendly regulations passed and gate their products behind restrictive APIs to stymie attempts at duplication.

Their dominance extends not only to market share, but has also translated into political influence, as they utilise lobbying, donations, and revolving door tactics to maintain their grip on power and get regulations favourable to them. Perversely, they are also more easily able to comply (or at least signal compliance) with the few regulations that are intended to police their conduct, which whilst largely ineffective are complicated and onerous, whereas for smaller companies this might be a prohibitive drain on resources.

Despite their overwhelming advantage, AdvanTech's leaders are victims of 'epistemic capture', and have either largely bought into their own marketing or turned a blind eye, showing little real concern for AI alignment or the ways that their products can be misused. Instead of using their immense resources to further alignment research and prioritise safety, they focus on 'code-washing', rushing products to market and dismantling regulatory opposition.

AdvanTech's AI products are far from perfect, causing significant harm to individuals and societies alike, and this harm is sometimes quite direct and easy to identify, such as clearly bad, biassed or self destructive advice. The company's lack of transparency and control over their AI systems makes it difficult to measure the true extent of the damage. This has led to a range of negative consequences, including the erosion of privacy, economic disparity, proliferation of disinformation and weakened global governance. Some of the same types of small-scale disasters (e.g. people getting into accidents following hallucinatory AI advice) present in Misaligned Metrics also occur here, but instead of an incompetent and inconsistent societal response, here we see active opposition to any response and sophisticated coverups of bad behaviour.

The company's influence extends to the media, where they manipulate public opinion to stifle dissent and promote their own interests, including successfully avoiding much oversight of their own AI developments. This has severely diminished freedom of speech, warped public debate and led to the further polarisation of society.

AdvanTech's lobbying efforts are bolstered by sophisticated persuasion tools and human modelling techniques, which they use to manipulate policy decisions and public opinion in their favour. Governments are effectively prevented from regulating the company or holding them accountable for their actions, even whilst being manipulated into serving their interests. Perhaps their greatest success in this regard has been to paint their interests as identical with those of national security. They push for stringent intellectual property, protectionist and anti-competitive laws insofar as it serves them, as well as draconian policing of the hacktivists and open-source movements that have sprung up in response to their monopoly, arguing that it is safer for there to be a single, 'responsible' actor in the lead than 'anarchic' multipolarity. Simultaneously, they decry any attempt at regulating them will hinder their lead over their closest competitor, a state-backed Chinese corporation that has a similar stranglehold over its own heavily walled market.

In developing countries, AdvanTech's abuses are even more pronounced - they dwarf many smaller economies in size and these countries are often even more dependent on their products and the business they bring. With less scrutiny than in the developed world and

weaker local institutions, these states have little recourse. AdvanTech exploits vulnerable populations and resources to expand their empire, often disregarding basic human rights and environmental concerns.

Discussion

Suppose that leading AI companies keep their monopoly for the reasons given in the summary. Why would existing AI companies end up following this villainous pathway? Such things have precedent; there are some corporations like tobacco companies whose business models depend on getting a harmful product with no social value past regulatory and public opposition. But most corporations, while they want to get friendly regulations passed, also don't follow a business model of marketing harmful or dangerous products. Here are some possible factors which could begin to explain this.

Unprecedented ability to avoid oversight: We might suspect (following the example of e.g. some powerful multinationals in developing countries), that if a corporation had unprecedented control over the government then the standard motivations of most CEOs and leadership would be sufficient to produce very bad outcomes.

Heightened competitive dynamics and epistemic capture: AdvanTech's executives are driven by the intense pressure to maintain and extend their market dominance. They see enormous profits to be made (on the level of dominating reasonable fractions of the world economy) by exploiting the powerful AI tools at their disposal, profits far in excess of what any other industry can generate, which no other company has had access to before. This relentless pursuit of market share and profit blinds them to the potential long-term consequences of their actions. Their epistemic capture prevents them from recognizing the importance of AI alignment and the need for a more cautious approach.

Influence of persuasion tools on company leadership: As AdvanTech continues to develop and refine their persuasive AI tools, the company's leadership may become increasingly susceptible to their own manipulative technologies. This creates a feedback loop, as these tools exacerbate their cognitive biases and lead them to make more self-serving decisions. The leadership becomes convinced that their actions are justified and rational, even as they drift further from ethical considerations and alignment research.

Flawed leadership and fear of competition: AdvanTech's top executives may be incompetent, callous, or simply misguided about the potential impacts of their technologies. They may view safety precautions and alignment research as unnecessary distractions that could slow their progress and allow competitors to gain ground. This fear of losing their market dominance drives them to cut corners, ignore safety concerns, and push their products out as quickly as possible. Additionally, they may be unwilling to entertain alternative viewpoints or accept advice from external experts, further entrenching their commitment to their current course.

This scenario is in some ways the mirror image of Misaligned Metrics. In that scenario, a race to the bottom with no clear winner results in a proliferation of many companies which ignore safety standards, rush out unreliable products which are still economically valuable, all competing with each other without any plausible chance of coordination on AI alignment or regulation. In this scenario, similar incentives are at work but instead within one or a few leading companies.

As with misaligned metrics, this scenario is compatible with either unsuccessful or successful alignment of more powerful AI in the future. It potentially leads into totalitarian-related failure modes like Let Them Eat Cake, Totalitarianism in One Country or Pivotal Act. Or to misaligned AI takeover.

It is important to note that this scenario is not simply about any powerful tech company monopoly arising, but specifically it's an unchallengeable monopoly of groups that won't listen to any advice on safety, misuse or alignment and won't cooperate with the government, the AI alignment community, academia or civil society. It's unclear whether an unchallengeable monopoly led by a consortium of AI companies that are cooperative with the government, listening to advice on safety, doing the best they can but in the lead is a bad outcome. A company honestly and accurately following a plan like Planning for AGI and beyond as written is not considered a misuse risk here.

Plausibility	**6/10** Another possible 'business as usual' continuation of trends but we have to assume specific corruption of a leading AI company as well	Substantial returns to first mover advantages and economies of scale, a major precondition for this scenario, seem plausible. This makes it possible that leading labs could become natural monopolies even if they don't develop TAI that can fully automate large parts of the economy. Whether AI labs will be able to keep a lid on their technical advantages (necessary for this rather than Misaligned Metrics) is less clear. Currently open source competitors are lagging behind frontier models even after fine-tuning. As with 'Misaligned Metrics', this scenario seems like one plausible 'business as usual' continuation of how things are going at the moment, although beware 'sleepwalk bias', the fact that it's 'business as usual' doesn't mean >50% likely. The assumptions listed as to how a tech company could end up sufficiently value-corrupted seem like they could play out, if we are dealing with economic incentives on a scale that have never occurred before, then previous precedents limiting how antisocial and dangerous corporations can become might not apply. Ultimately, looking at the actual leading companies in detail for the risk factors listed in the discussion would help refine this assessment and whether or not this is at all plausible in specific cases.
Severity	**3/10** More dangerous than 'misaligned metrics' since there is active opposition to the safe use of AI	In this scenario there's a capable and active opposition led by intelligent and well resourced people towards any systematic attempt to deal with misaligned or misused AI, but most of the dangers of Misaligned Metrics are also present. Therefore, this seems more dangerous than Misaligned Metrics where there's a free-for-all with some inertia but nothing organised to prevent the government from stepping in at some future point. As with the first scenario, if alignment is hard and

		there are no warning signs/other interventions that change things, then the deployment of power-seeking misaligned AI seems plausible. There is also less chance for warning signs as large leads present.

Cascading Collapse

A hacker collective contracted by operatives working for a rogue state is given the code for a cutting-edge computer virus that they have obtained through espionage, one that is capable of rewriting its code to evade detection, duplicate itself online, adapt its attacks and survive attempts to expunge it. Perhaps this system is a future descendant of AutoGPT, more specialised for hacking.

The hackers launch a coordinated attack against the financial services infrastructure of an enemy nation, in a deniable act of retaliation for a cyberattack it in turn is suspected of carrying out, which temporarily shut down portions of the rogue state auto-censor network. Exploiting an unknown vulnerability or flaw in the system design or implementation, they are able to wreak widespread damage - much of the target infrastructure uses the same piece of software as it was overhauled just six months ago in an attempt to remain competitive and secure. This is just the beginning however.

Within hours, malfunctions, errors and disruptions are skyrocketing across the broader proto-CAIS network of the nation as a whole: hours later, they are occurring globally. The human operators and managers of the proto-CAIS responding to the crisis are unable to diagnose the cause of the problem, nor fix it quickly enough owing to the sheer complexity and inscrutability of the systems involved. They also lack a backup or alternative solution - the points of failure seem to be everywhere and they have become overly dependent on the proto-CAIS for their operations and decision making.

The cascading failure of the proto-CAIS system spreads to most sectors, including banking, commerce, communications, finance, health care, logistics, transportation and defence. With so much of the economy tied up in the proto-CAIS network, the effects are global, even spreading to less integrated markets. Within days, many tens of millions of people worldwide have lost access to their incomes, investments and savings, to their credit, academic and medical records. Within weeks a state of emergency has been declared. Many government services have effectively ceased functioning and countless companies have gone bankrupt, whilst others have been nationalised in an effort to forestall further collapse. Mass panic has spread, leading to bank runs, hoarding, riots and looting, exacerbated by the loss of trust and confidence.

Months pass and the economic crisis deepens. Many countries face political instability, social unrest and humanitarian crises, whilst opportunists exploit the chaos with so much of the world in disarray. Accusations fly over culpability for the attack, further reducing international cooperation and raising tensions.

It is ultimately determined that the mutated virus happened to find an exploit in the design of a chip manufacturer responsible for a large percentage of the world's advanced hardware, enabling it to do far more damage than was initially anticipated. The recovery from the

collapse is slow and difficult and its aftershocks will be felt for decades. The human experts have to rebuild or replace the damaged or corrupted CAIS systems from scratch. They also have to reevaluate their assumptions and methods for designing and deploying CAIS systems in the future.

Discussion

This scenario seeks to draw attention to a few different concerns. Firstly, that mass automation may increase vulnerability via the kind of cascading collapse illustrated above, especially as a tiny handful of companies may continue to be responsible for the vast majority of the world's advanced chip manufacturing capacity, or training the AI systems that run on them, leading to the potential for single points of failure - a design flaw or exploit in even a single generation of chips or foundation model derived software could have broad repercussions, especially if uptake of the new technology is sufficiently swift. Secondly, that in an increasingly automated, networked world, unrestricted cyberwarfare has the potential to do exponentially more harm. Third, recent evidence suggests that adversarial vulnerability may be a uniquely serious problem for LLM-based foundation models in particular, with adversarial examples always findable (at least in principle) and often transferring between diverse models.

In this scenario, the focus is on the fact that increased automation could increase vulnerability to the sorts of cyberattacks that aren't too different from what's possible already. The 21st Century Visigoths scenario discusses a world that hasn't been strongly automated being susceptible to advanced AI based offensive technologies, while this scenario is about a world that is transformed and heavily automated becoming vulnerable to new kinds of attacks that don't require extremely advanced superweapons. This could either result from fast adoption of AI technologies or from a longer timelines scenario, or from everyone ignoring and failing to fix adversarial vulnerability.

When developing AI systems, organisations or developers might prioritise short-term gains or immediate progress over thoroughly addressing potential risks and issues. By focusing on making the harms illegible, they can avoid dealing with complex, resource-intensive challenges tied to alignment, safety, and ethics. This approach might enable the appearance of a well-functioning CAIS system but may leave underlying vulnerabilities unaddressed.

These unaddressed vulnerabilities can accumulate and interact with one another over time, creating a fragile system that appears robust on the surface but contains hidden risks. As AI systems become more integrated into various aspects of society and technology, a single event or failure could trigger a chain reaction, causing these vulnerabilities to be exposed and potentially leading to a cascading collapse.

While it's possible that increased integration, centralization, and uniformity could make the global economy more vulnerable to attacks, it's also possible that well-designed CAIS systems could be more resilient and less brittle than current systems. The scenario's plausibility depends on factors like the level of intelligence and coordination of the CAIS systems, the degree of human dependence and oversight, the availability of backup solutions, and the robustness and security of the systems.

Key uncertainties include the potential for greater automation to make systems more vulnerable, the likelihood of discovering and exploiting a single-point failure, and the

assumption that solving alignment or pursuing non-agentic systems would necessarily lead to a more vulnerable civilization.

This could also be a precursor to misaligned AI takeover or represent a risk factor for making AI takeover easier (CAIS automation makes it easier to seize resources if you're an AI).

Plausibility	**4/10** Describes a specific conjunctive future of high automation where robustness problems aren't solved, but there are reasons to believe this is a possible outcome especially if LLM-based foundation models are used as the basis for TAI	This scenario assumes that we're capable enough to set up advanced CAIS and solve many alignment issues but still leave it highly vulnerable, so a mixture of initial competence and later incompetence may be less likely unless that's just the default for CAIS. The key question to investigate is whether this actually is the case, we could easily imagine all that automation being better and less brittle than current disaster response, finance, transportation etc. depends on many factors, such as: • The level of intelligence, autonomy, and coordination of the CAIS systems and the misaligned AI • The degree of dependence, trust, and oversight of the human users and managers on the CAIS systems • The availability and effectiveness of backup or alternative solutions in case of a CAIS failure • The robustness and security of the CAIS systems against attacks or errors Recent results related to model Adversarial Robustness seem to increase the likelihood of this scenario on the assumption that LLM-based foundation models are widely deployed, as it seems with RLHF removing adversarial examples from LLMs is impossible given a sufficient length prompt. Overall, this scenario requires a specific pathway to near-full automation including of vulnerable systems, with a lot of usual cybersecurity best practices not followed properly, so we consider it less likely here.
Severity	**4/10** Direct wide-scale destruction is caused	In the worst-case collapse envisioned, damages could be catastrophic, with extreme disruption to finance, energy, food, communications, and more.

	that exceeds the damage of a worst-case scenario from a cyberattack today	Millions could perish in the resulting chaos. However, severity depends on attack scale and resilience factors. Not all CAIS failures need be globally damaging. Manual overrides and backups could limit impacts. Human adaptability provides some robustness. A coordinated response could restore functionality. Still, heavy CAIS dependence risks mass suffering if core functions are disabled. Failures could cascade unpredictably across systems. Infrastructure provides lifelines - losing it leaves populations vulnerable. So while an existential human extinction event seems very unlikely, deaths on the scale of millions are plausible risks. The scenario highlights the need for stability and security precautions as CAIS advances. Overall this seems on a par with a major (non-nuclear) war in terms of damage, with potentially millions dead due to the subsequent infrastructure collapse.

Heaven on Earth

AI systems become ever more competent at generating media such as infotainment, literature, music, visual art and eventually more complex output such as movies, games and immersive virtual environments. Without the bottleneck of human labour, this content can be generated on an unprecedented scale. Whilst many in the creative industries protest that the AI generated output is a mere facsimile of genuine human creativity, the more these tools develop the more any alleged deficit is invisible to the majority of consumers. Many begin to herald these AI tools as ushering in a cultural renaissance – many passion projects that might never have existed for the lack of a perceived market bear fruit as small groups and even lone individuals are empowered to create works that once might have been the domain of large companies alone. Other artists use these tools to create work that might never have otherwise been practical or even possible.

Simultaneously, AI 'chatbots' have developed into hyper-realistic digital simulacra, capable of generating audiovisual output (and with a haptic interface, tactile output) in real time that although not perfectly indistinguishable from actual humans are incredibly lifelike. These emulations populate forums, movies, streams, social media networks and virtual environments and whilst in some contexts they exist in an arms race with auto-censors and mods, in others they are an established part of the business model. Though some emulations might be used for manipulative purposes such as committing blackmail and fraud or spreading propaganda, others are more benign, providing fulfilling parasocial relationships to the lonely.

As these various AI systems move from proof-of-concept prototypes to established commercial products however the competitive pressures of the attention economy begin to

supersede any other concerns – indeed, with the marketplace so saturated they are driven into overdrive, and gradually economies of scale reassert themselves and tech oligopolies entrench. It is not just that they are able to generate content on a greater scale or engage with more people – more crucially they can use AI to analyse the vast quantities of consumer behavioural data they have access to, honing their products to maximise engagement, right down to the level of personal customisation.

The emulations, augmented or virtual realities and unending stream of AI generated content becomes ever more finessed, to the point where many come to prefer these virtual interactions to the struggle and mundanity of real life and increasingly engagement with this output takes on the form of compulsive or addictive behaviour. Indeed, this is by design – whatever the original or stated intentions of these products, economic incentives drive them in the direction of hyper-stimulation that the real world cannot match. An epidemic of anomie and mental illness ensues. Harrowing stories begin to emerge. Neglected care home residents whose only interaction was with emulations. Children with severe developmental disorders because most of their early childhood was spent in gamified virtual environments. Rust belt towns where much of the population subsists off basic income to fund parallel lives led online. Families torn apart by members siloed into tailor-made echo chambers. Rapes committed because real women didn't behave like pornographic virtual companions. Atrocities committed because the real world wouldn't bend to every whim or because an emulation inexplicably went off script and led a follower down the path to radicalisation.

As with many epidemics of addiction it is the disenfranchised, dispossessed and marginalised that are the worst afflicted and governments are slow to act. Moreover, the responsibility for this slow societal breakdown is diffuse, hard to pin on any one cause and often existing at the confluence of several. By the time they attempt serious regulation they are opposed by powerful, entrenched interests in an environment where consensus reality has all but broken down and large sections of the economy and society are heavily intertwined with virtual agents and realities – indeed some accuse them of turning a blind eye, enabling a modern form of 'bread and circuses' that keeps what might otherwise be a restless populace distracted, placated, and atomised.

Discussion

This scenario is intended to illustrate concerns regarding the potential for value corruption, not necessarily towards conflict-prone or sadistic preferences but towards unendorsed values that have little relation to genuine human flourishing or well-being. It is designed to show a clearly extremely bad case of value corruption (by common-sense moral standards) without any explicit takeover or direct harm inflicted on humans.

This scenario shares some features with 'cascading collapse' and related scenarios, i.e. its about dependence on non-agentic systems in ways that massively increase vulnerability and presumably directly constitute an existential risk if they get bad enough, but here there's more focus on "wireheading and exploiting human's reward systems specifically.

In this scenario, something like Misaligned Metrics or Unchallengeable Monopoly is taking place in order that these addictive technologies get deployed without much regulatory opposition. This scenario also has the potential to lead into wide-scale value corruption especially in established democracies, with values being shifted towards simpler and less

worthwhile ends. If nothing else intervenes, this could cause an existential catastrophe by itself where humanity's future is captured by these highly addictive tools.

This scenario could also be one of the enablers of excessive centralization that leads to Silent Strings, serving as 'modern bread and circuses'. It is also worth noting that this is meant to be an 'unendorsed' value corruption in the sense described in this post. In that post, we dissected the notion of 'unendorsed' actions into two main components: the human principal's initial values or preferences (let's call them Initial-H) and the evolved or shifted values over time (Evolved-H). When a decision or conflict is "unendorsed," it usually implies a misalignment between Initial-H and the action that was taken by the AGI agent (A). However, what complicates this is the notion of value shifts, specifically shifts that lead to Evolved-H.

1. Human Preferences Being Corrupted: If interactions with A lead to a modification or corruption of H's preferences (shifting from Initial-H to Evolved-H), and then Evolved-H approves of a decision that Initial-H would not have, that decision is "unendorsed" by the original human intent. For example, let's say you initially value environmental sustainability but, through interaction with a persuasive AGI, come to prioritise short-term economic gains instead. If this AGI then takes an anti-environmental action that you approve of based on your new value system (Evolved-H), that action is "unendorsed" relative to your original values (Initial-H).

2. Misconfigured Initial Architecture: Another instance would be H designing A with an architecture that doesn't fully encapsulate H's original values (Initial-H). If A then engages in a conflict that aligns with its architecture but not with Initial-H, it's an "unendorsed" action, even if H at that later point might think it's rational. For instance, if you program an AGI to maximise productivity but forget to add a clause for ethical sourcing, and the AGI then chooses a highly efficient but unethical resource, that action would be unendorsed based on the initial values (Initial-H).

- Hindsight and Reflection: It's easy to say in hindsight that "I would have endorsed this decision if I knew better," but the point of focus here is whether Initial-H would have endorsed it when A was deployed. This is a more complex version of "unendorsed," where it might not just be a mistake or misalignment, but an evolved understanding over time, making the action conflict with Initial-H but align with Evolved-H.

- Incomplete Understanding: If A understands Initial-H well but doesn't get the full grasp of Evolved-H due to incomplete data or non-intuitive leaps in human value evolution, then any actions it engages in based on this incomplete understanding would also be "unendorsed," even if it made perfect sense to A given the data it had.

Plausibility	**3/10** Requires the weak regulation and oversight common to the first two and speculative	This scenario like it requires all the assumptions of e.g. Misaligned Metrics or Monolithic Monopoly or something similar (otherwise we'd see effective regulation limit the damage at least somewhat or prevent it from increasing civilizational vulnerability), but takes them further in a specific direction, so this scenario is quite conjunctive.

	assumptions about the likely impact of addictive technologies	It requires many assumptions about the ability of certain technologies to corrupt values, when they aren't intentionally trying to alter human preferences via long-term planning. While there is some precedent for this, the extent to which 'Heaven on Earth' takes value corruption is unprecedented in human history. Additionally since these systems are sophisticated, it assumes many alignment/misuse problems are solved without anything interfering, i.e. the world is otherwise going quite well, which are additional assumptions. The scenario also requires us to handwave away cultural opposition to these technologies which in many parts of the world outside the US and Europe will be extreme. Legislation would have to either ignore or actively encourage adoption of these addictive technologies, even as the negative effects become clear. Overall, this seems like a possible but unlikely outcome of a combination of weak societal response but favourable AI alignment difficulty and low misuse risk.
Severity	**4/10** Plausibly sets the world on a path to the lock in of an undesirable state, but there's no enforcement mechanism so this is not too likely	If this state of affairs becomes truly permanent then it can cause a very vulnerable future of high dependence with incompetent or uninterested decision makers, although it would have to be very enduring, global in scale and impossible to alter for it to count as a large-scale disaster. This seems like it could if it becomes permanent and enduring it could potentially represent an existential catastrophe via loss of most of what we consider valuable, although that could be considered controversial, this looks like a future that would be strongly unendorsed by more or less any present human decision-makers. Whilst this may result in value drift, deaths of despair and further fertility declines, it is hard to see how this could become pernicious/pervasive enough to result in existential risk.

Hell on Earth

As AI systems proliferate, certain groups begin exploiting their capabilities to spread harmful ideologies. These groups leverage AI's pattern recognition abilities to identify susceptible individuals and target them with personalised propaganda and misinformation campaigns. The advanced behavioural modelling of these futuristic foundation models allows precisely tailored messaging that plays on individuals' psychological vulnerabilities.

Initially, these campaigns aim to promote radical ideologies like neo-fascism and apocalyptic cults. The groups use AI-generated media depicting idealised visions of the future under

these ideologies. This media is disseminated through algorithmically curated social media feeds that immerse recruits in an ideological bubble.

Over time, the groups' capabilities improve as they gather more data on what persuades people. Their messaging increasingly glorifies struggle, religious or political strife or suffering, justifying it under the guise of religious visions or racial superiority myths. Hardship is portrayed as necessary for purification, perfection, or social cohesion. New members are conditioned through AI-guided indoctrination techniques.

These campaigns gain momentum by capitalising on social divisions, economic precarity, and societal instability. Their visions offer purpose to the disillusioned. Opposition is hampered by the decentralised and evolving nature of the campaigns. Governments struggle with heavy-handed censorship, setting dangerous precedents.

Eventually, the ideologies spread beyond the core groups, permeating mainstream discourse. Their visions reshape social values, eroding compassion and empathy. As public opinion shifts, policy begins to reflect the new norms. Legal protections are stripped away, and atrocities are committed against marginalised groups. Suffering spreads, but objections are silenced as sacrilegious.

Discussion

This scenario is intended to illustrate a worst-case scenario for human value corruption without AI takeover or intentional adversarial planning on the part of Transformative AI systems. It is an attempt to assess whether human values could be sufficiently corrupted by human use of AI tools alone, such that preferences that endorse hardship or mass suffering could be spread, making large scale S-risks plausible assuming nothing changes and human civilization continues to advance.

Ideologies which glorify suffering, struggle and hardship have occasionally spread through mostly nonviolent means. Romantic militarism pre-WW1, fascism in the 1910s to 1920s and Salafi jihadism from the 1980s to today all managed to spread to tens to hundreds of millions of people without violent forced conversion. It is not out of the question that future AI developments could worsen this problem should a suitable candidate ideology emerge.

Plausibility	**2/10** Similar assumptions to ‘heaven on earth’ but makes even more specific and extreme assumptions about value corruption	This scenario likely requires strongly effective persuasion tools and a group dedicated towards using them for this malign end. Alternatively, it assumes some (hard to justify) fundamental advantage to suffering-prone ideologies in a future of easily available AI driven persuasion. It requires that, of all the uses they are put to, encouraging values that place a lot of worth on suffering is one of the more successful outcomes when there’s not an easily identifiable group currently present which would push for the spread of an ideology like this. The core assumption of highly effective AI persuasion technology remains speculative. However, the scenario shows how AI could exacerbate harms by amplifying certain harmful

		beliefs or manipulating public discourse. There are concerning precedents of technology-enabled radicalization. Overall, an existential catastrophe seems unlikely, but a limited, non-existential catastrophe amplifying suffering appears possible.
Severity	**8/10** Plausibly sets the world on a path towards wide scale suffering catastrophe	This is a potential large-scale S risk, especially if the malevolent ideologies persist as technology advances further, so it's by necessity worse than most of the other provided examples as it seems to set humanity on a path to worse than extinction outcomes. However, it is susceptible to the same potential impermanence as 'Heaven on Earth' so we don't rate it as bad as a scenario that results in extinction.

21st Century Visigoths

A dangerous combination of AI-related technologies dramatically increases civilizational vulnerability. Powerful persuasion tools undermine public epistemology, while AI systems enable the rapid development and deployment of asymmetric weapons technologies, such as highly capable drones, cyberwafare and the rapid prototyping of new bioweapons without the need for qualified scientists and biolabs. As it turns out, LLM based cyberoffense systems are much easier to deploy, train and test than LLM based cyberdefense systems, so even small groups can easily crash the power grid or otherwise mess with critical infrastructure. Even with vastly more investment and their own AI tools, governments can't keep up and can't keep these tools from proliferating.

Additionally, the timeline for the development, testing and deployment of such protective technologies is far too long, compared to the cheap asymmetric weapons technologies transformative AI enables.

This results in large-scale terrorist attacks, both physical and cyber, that target infrastructure and population centres. Police and law enforcement raid terrorist biolabs attempting to cook up dangerous bioweapons, and it's only a matter of time before one gets through.

Additionally,superhumanly effective persuasion tools allow terrorist groups to overcome one of their most significant barriers: recruiting people with high-level technical skills and maintaining their commitment to causing mass casualties. By exploiting AI's persuasive capabilities, terrorists can radicalise crucial experts through targeted persuasion or blackmail, or employ a stochastic terror approach, provoking seemingly random lone-wolf attacks.

As these attacks become more devastating, the use of cheap, easily produced (maybe by auto factories guided by AI) advanced bioterror and lethal autonomous weapons (LAWs) becomes more widespread, placing society in a precarious position. In this scenario,

Finally, a terrorist cell, a descendent of ISIS working out a poorly resourced lab in Syria with access to a stolen copy of a gene-sequencer foundation model and some off-the shelf biohacking equipment, leverages AI tools to accelerate the development of dangerous

bioweapons far more destructive than any natural pandemic. They simply package the virus in some contaminated food and drop it off in some major cities around the world, in an attempt to shatter the power of the US and its allies. Of course, they weren't sufficiently careful nor concerned with the virus running out of control, which it does, resulting in a pandemic that kills more than half of the worlds' population in a matter of months.

Afterwards, the world is still in a precarious state. With many governments collapsed and billions dead and the powerful bioweapons technologies still in the hands of many billions of people, perhaps it is only a matter of time before someone releases another lethal pandemic…

In one variant scenario there are still significant material and startup capital costs for mass terror and wide scale destruction, as well as major technological hurdles - i.e. it remains beyond the scope of smaller actors to train large foundational AI models or create novel biological or nanotechnological weapons in an independent lab. However, the proliferation of AI models thanks to the exploitation of open source copies as well as cheap, mass producible LAWs enables every cartel, junta, mafia, militia, PMC and rebel group access to a major force multiplier, levelling the playing field with more advanced, state-funded militaries and further enhancing the risk of asymmetric warfare. In response, security forces are compelled to become ever more militarised, with a concurrent upsurge in state failure and civil wars.

Discussion

This scenario is meant to illustrate a major concern about AI misuse, namely that there might be some capabilities which are easy to proliferate to large numbers of actors, disproportionately destructive and hard to counter conventionally: the ability to rapidly prototype bioweapons and deploy them being perhaps the worst of all of these, but the ability to cheaply mass produce LAWs or release fully autonomous cyber agents are also significant. This scenario assumes that no serious countermeasures are possible or that they aren't developed, i.e. that there aren't corresponding defensive AI enabled technologies. It also assumes that malicious actors who don't care about their weapons causing devastation or running out of control are around in sufficient numbers. Is this plausible?

The scenario presumes an offence-defence imbalance favouring destructive applications of AI that aligned systems are unable to overcome. This imbalance enables small rogue groups to wield AI technologies capable of catastrophic attacks exceeding states' defensive and countermeasure abilities. However, the extent of this imbalance is debatable and contingent on several factors.

It implies states and aligned AIs will be unable or unwilling to take extreme proactive measures to contain threats. But especially if warning signs are present, and especially given how obviously dangerous some of these capabilities would be if they proliferated, we could see extensive control of such technologies by default.

On defence, while potential protective AI systems (e.g. to monitor for rogue cyberattacks or groups plotting to develop bioweapons) face oversight constraints, their superior scale could confer aggregate advantages. Surveillance and detection of dangerous AI activity appears tractable. And resilience capabilities like decentralised networks, emergency response, and

shelters could aid recovery from attacks. Furthermore, economic imperatives will likely drive military and government adoption of AI defences. So while offence may have inherent advantages in some domains, assuming this enables uncontrolled proliferation risks overlooking defensive counterbalances.

So while an offence-favouring imbalance could emerge in narrow domains, the scenario's extent of asymmetry enabling catastrophic proliferation may rely on overly pessimistic assumptions about access, coordination, and defensive responses

Plausibility	**5/10** Proliferation of asymmetric destructive weapons seems like the result of business-as-usual actions, but it requires a specific malicious actor to want to use the technology	How objectively 'easy' is it to invent superweapons, and what does the offence-defence balance look like? This is something covered at greater length in a later section, but the overall conclusion is that there seem to be a few paths to dangerous highly proliferated capabilities, although none that are capable of global catastrophic damage and lacking material supply bottlenecks. If asymmetric superweapons are easy and something like misaligned metrics happens and no near-term wide scale transformation (baseline scenario) then this looks plausible. To assess the plausibility that any of this can happen before general progress is accelerated into a new regime, or AI takeover intervenes instead, we need to know which of the superweapon techs we see are TAI-complete. Otherwise, these could develop if aligned agentic TAI proliferates everywhere without a stable world government or good oversight, then this could happen even in a radically transformed post-TAI world, if the TAI isn't used to reduce risk. The availability and accessibility of AI-enabled tools for persuasion, blackmail, bioterror, and LAWs (Lethal Autonomous Weapons) are expected to increase over time. However, the effectiveness of these tools in radicalising crucial experts or provoking stochastic terror attacks remains uncertain.
Severity	**8/10** Both directly highly destructive (though the magnitude is somewhat uncertain as it depends on the weapon type) but also a large risk factor and danger to the future	Once the proliferation has occurred (e.g. of LAWs manufactories or bioweapons), civilization could be in a permanent state of threat, even if the first attack doesn't destroy civilization, it may cause enough disruption to ruin any capacity to prevent further proliferation or build defences, and we could end up in the classic Bostromian 'vulnerable world', where any future attempts to rebuild civilization are vulnerable to the same cheap asymmetric weapons technology forevermore. This scenario could result in a direct existential catastrophe on the first use, or if not then a highly destructive war that undermines global stability.

		The score is not higher because there's less reason to assume omnicidal motivation from any human actor, so catastrophic events seem more likely than direct attempts to bring about human extinction. Also, with only a single use by a weakly resourced group, most of the described technologies don't seem that likely to result in human extinction. The most extreme severity likely requires multiple factors – highly scalable weapons, rapid proliferation across groups, and coordinated deployment for maximum damage. Some technologies may fit this bill, but more likely are smaller attacks creating fear and tension, but not inherently destabilising civilization. Preventative measures may also mitigate risks, especially if dangers are recognized and early.

Totalitarianism

These are scenarios where AI enables unprecedented totalitarian control, through a combination of soft power (persuasion and economic centralization) and hard power (improved surveillance, military advantages and surveillance). Scenarios include single-party dictatorships leveraging these tools for repression and global conquest, and democratic states sliding into authoritarianism. A key feature is the use of AI to impose values and interests on populations. Feasibility uncertainties exist regarding enduring global hegemony.

2084

China emerges as the world's leading power in AI development, having poured vast resources into research and innovation while maintaining a strict grip on its political and economic systems. The Chinese government's ruthless dedication to technological progress, combined with its authoritarian regime, allows the nation to monopolise groundbreaking AI advancements and create a comprehensive AI services economy. By contrast, concerns over misalignment (after some high profile accidents), AI misuse and corporate bad behaviour in the US and the democratic world more broadly stifle the progress of AI technology, despite their initial lead. Following yet another contested presidential election that leads to mass civil unrest and even armed insurgency in the United States, China takes advantage and seizes Taiwan in the early 2030s after a 6-month war, without NATO intervention, cutting off the supply of cutting edge chips to the United States, and they pour money into scaling up ever larger foundation models while bans on large-scale training runs in the US and West are successful.

Over time, China's AI prowess enables its military to gain a significant advantage over its rivals, achieving a level of efficiency and power that exceeds the west and NATO, with rapid prototyping of new conventional weapons, mass deployment of fully autonomous LAWs and eventually a secret program on bioweapons and advanced nanotechnology. At the same

time, internal surveillance, the deployment of persuasion tools along with sophisticated internal surveillance and increasing economic automation all push China in a more totalitarian direction. The surveillance systems make squashing dissent and enforcing uniformity easier, the persuasion tools make propaganda far more effective including on the countries' decision-makers, driving a feedback loop, and the economic centralization means there are fewer downsides in terms of popular discontent to increasing levels of totalitarianism.

With this newfound strength and purpose, China's leaders make a crucial decision. Perhaps because they are aware that their lead is likely temporary (since even with their lead, western nations might catch up in a few short years), they decide that it's now or never. China embarks on an ambitious campaign of territorial expansion, seeking to extend its influence and control across the globe.

The Red Dragon's Conquest begins in Asia, as China swiftly absorbs neighbouring countries through a combination of economic pressure and military force. Nations that resist are met with overwhelming displays of AI-driven warfare, their defences quickly crumbling before the relentless onslaught of autonomous drones, advanced cyberattacks, and cutting-edge battlefield technologies. Many nations simply surrender outright once it's clear that there's not much hope of fighting back, as thanks to laser and drone-swarm based technologies, even overwhelming launches of nuclear missiles aren't likely to succeed.

As China's sphere of influence expands, the rest of the world watches in growing alarm. Attempts to resist or counter the Chinese juggernaut are met with crushing defeats, as even the most powerful military alliances prove no match for the Red Dragon's technological might. One by one, nations around the globe fall under Chinese control, either submitting to economic dominance or succumbing to outright conquest.

With each victory, the Chinese government tightens its grip on the territories it claims, imposing its brand of totalitarian rule and stamping out any dissent or resistance. As the last bastions of freedom and democracy are extinguished, the world finds itself under the iron fist of a single, all-powerful regime.

The Red Dragon's Conquest ultimately results in a global totalitarian state, governed by a regime that wields AI as both a tool of control and a weapon of war. This chilling vision of the future serves as a stark warning about the potential consequences of unchecked technological development and the dangers posed by the rise of AI-enabled authoritarianism.

Discussion

This would require a substantial deterioration in political conditions within China. By default, a strong motivation for world conquering expansionist totalitarianism is not that common in the rulers of powerful countries. It's hard to say if genuine motivations for world conquest are uncommon. While few examples exist, there are also counterexamples. Determining the authenticity of such motivations can be challenging because individuals with such ambitions often conceal them. Hitler plausibly did want to conquer the world, and potentially Stalin and Mao would have if given the opportunity. However, as with 'Unchallengeable Monopoly', the

(perceived) unprecedented all-or-nothing nature of the stakes could make such motives more likely.

This scenario assumes a slow enough takeoff for large leads to be possible or a very rapid efficient internal transformation, and a certain degree of (pseudo) alignment and competence in early TAI technologies. However, takeoff must be fast enough to produce a significant lead, which could be one of the few factors motivating a massive war of conquest. If this situation were to occur within the next 20 years, it would more or less have to involve China as the dominant power, given its current capabilities and potential for totalitarianism. All current totalitarian states, as North Korea, Syria, Turkmenistan, Eritrea, and Afghanistan, are unable to achieve such a lead without an improbable turn of events. So an existing country with such tendencies must first turn totalitarian and then embark on aggressive world conquest: Russia or China are possible candidates here.

The key implausible step in this scenario is the turn to outward-focussed global conquest that succeeds universally. It seems plausible that internal pressures related to TAI could turn an already authoritarian state like China totalitarian and that due to this process, they could accelerate ahead of the west (e.g. China is ahead of the west on things limited by regulation like infrastructure building). The major explanation for why this happens is supposed to revolve around first the all-or-nothing nature of the stakes and second the possible internal influence of persuasion tools or the same pro-totalitarianism economic/political incentives that arise from centralization also being pro world conquest.

This overview article provides a summary of the track record of totalitarian states and their successes and failures, as well as the risk factors for a state sliding into totalitarianism: The totalitarian threat | Global Catastrophic Risks | Oxford Academic, as we have discussed, many of the risk factors are already present in China and the development of advanced AI could provide more. Advanced AI systems could dramatically improve the technological capabilities available to totalitarian regimes across several crucial domains:

Persuasive Propaganda: AI could generate hyper-personalised propaganda tailored to specific individuals by analysing patterns in their online activity. This propaganda could exploit psychological vulnerabilities and social pressure to nudge the population toward conformity and obedience. While not mind-control, such individually targeted messaging could substantially increase the effectiveness of state ideologies.

Mass Surveillance: AI enables monitoring the population's communications, movements, biometrics, and social behaviour through automated speech recognition, facial recognition, data mining, and other techniques. This empowers security forces and reduces citizens' privacy. AI can also enhance hacking and cyberattacks to infiltrate opponents' systems and networks. Together, these capabilities significantly strengthen authoritarian social control.

Economic Control: By automating sectors of the economy, AI concentrates economic productivity and wealth generation within state-controlled systems. This centralises power and resources, creating extreme disparities favouring elites. It also makes the broader population dependent on AI-driven services vulnerable to state coercion. Additionally, AI-enabled genetic engineering or pharmaceuticals could potentially be used to promote docility and compliance, although the feasibility remains speculative.

Military Dominance: AI can provide overwhelming advantages in offensive autonomous weapons, cyberwarfare, and battlefield intelligence gathering. This makes conquest and eliminating external rivals more feasible. AI-powered manufacturing also streamlines equipping large military forces. Together, these strengthen the regime's hard power, inhibiting external resistance.

Plausibility	**2/10** Makes many conjunctive assumptions, both about TAI weapons technology, automation and china gaining a lead and becoming totalitarian	If it's based on a direct tech lead, this scenario requires a specific takeoff speed and basically has to happen in China, so a highly specific scenario which is correspondingly less likely given their current disadvantages. But if the group is highly motivated and more reckless with applications of technology and conversely other countries are very slow to adopt, this could happen. This hypothetical China would have to be powerful enough to win a nuclear war with confidence, which implies either an extreme indifference to population losses or incredibly powerful weaponry far beyond nuclear weapons. Wars of conquest are rare in the modern world (edge cases such as the Russian invasion of Ukraine or a potential Chinese invasion of Taiwan notwithstanding) for good reason - for a totalitarian state to attempt such a thing would require a massive technological edge - not just enough that it is possible to do so without risking MAD, but enough that it is bordering on trivially easy to the extent that there is cause to do so rather than use subtler means. The likelihood of a single country monopolising AI developments to the extent that it gains a military advantage equivalent to 100+ years is relatively low. The global diffusion of technology, international cooperation, and the potential for rapid catch-up by other nations make it challenging for one country to achieve such a substantial lead.
Severity	**6/10** It is unclear how totalitarian and corrupt this future CCP rule would be, or how total, but could represent a loss of human potential	Human life continues and not all these scenarios result in *immortal* totalitarianism. But if this is a robust and expansionist state then it may even constitute an S risk (if we include persuasion tools changing preferences in more extreme directions than would be possible normally as described in Heaven on Earth/Hell on Earth). Even if we assume that this scenario results in a single totalitarian world hyperpower with a huge AI provided technological advantage, there will always be dissent on a smaller scale, although with AI and potential transhumanist technologies to apply to the leadership (e.g. life extension) there are mechanisms by which the dictatorship could persist a long way past current norms or the regime could be very stable and globally dominant

		even if it doesn't rule every single human alive.

Silent Strings

The Silent Strings, a coalition of powerful interest groups, have formed an alliance with the Chinese Communist Party (CCP), capitalising on the advanced AI-driven technologies at their disposal. They have a single goal: to spread a totalitarian ideology across the world and establish a new world order under their control. Their goal need not have been pre arranged but instead could have evolved from the current lesser expansionist aims of the CCP as AI driven transformations make the future seem more high stakes and all for nothing. They could have concluded that it's necessary to ensure the reckless experimentation by western countries ceases, or have emerged from the global reactionary far right.

To achieve their objective, the Silent Strings employ a multi-pronged strategy. They begin by infiltrating the global economy, using AI-driven innovations to dominate key industries and amass unparalleled economic power, making most other countries dependent on them through trade. Through their corporate influence, they bind other nations to their interests, creating a web of dependency that leaves governments beholden to their benefactors. This can be considered somewhat continuous with the way the Chinese government behaves nowadays, e.g. exporting its government model to currently nonaligned countries in Africa.

Next, they unleash their most powerful weapon: advanced persuasion tools. By employing AI-generated propaganda, they manipulate public opinion and the media, sowing confusion and misinformation. As traditional democratic institutions crumble under the weight of this insidious assault, the Silent Strings' totalitarian ideology gains traction.

They begin by exploiting the internal divisions and fractures within these countries, turning the public's fear and mistrust of AGI into a weapon to further their own agenda. In contrast to a chaotic and fractured West the 'Chinese model' may start to genuinely look more appealing as a model for global governance/development, with better security against external threats.

Capitalising on the widespread media coverage and public concern, the Silent Strings launch a coordinated disinformation campaign. They use AI-generated propaganda to amplify existing fears, exaggerate the risks associated with AGI, and discredit the efforts of those working on AI safety. As the discourse around AGI becomes more polarised, the Silent Strings deftly exploit the confusion, driving a wedge between different factions and weakening the social fabric of their target countries.

As Western governments struggle to respond to the (perceived) AGI threat, the Silent Strings position themselves as the solution to the crisis. They use their influence and resources to promote their model of centralised control, arguing that only a unified and strong authoritarian regime can effectively manage the dangers posed by AGI. They point to the failures and incompetence of the Western countries' responses to previous crises, like the COVID-19 pandemic, as evidence of the need for a more controlled and efficient system.

Simultaneously, the Silent Strings exploit the political divisions within Western countries by providing covert support to extremist groups and political parties that align with their totalitarian ideology. They leverage AI-driven technologies to infiltrate and manipulate social

media, amplifying divisive content and deepening the rifts between different factions. By fueling chaos and discord, they create an environment ripe for the spread of their ideology.

Backed by the resources and expertise of the CCP, the Silent Strings invest heavily in AI-powered surveillance systems, which they market to other countries which use them to maintain an iron grip on populations around the world. Dissent is swiftly identified and neutralised, as predictive algorithms pinpoint potential threats and human rights abuses become an accepted means of maintaining control. Naturally, the countries employing these methods in concert with each other are allies with each other against opposing models of government, but especially as the fear of AI risk spreads.

The ideology spreads like wildfire, fueled by a combination of human motivation, state endorsement, and AI-driven memetic manipulation. It transcends political and religious boundaries, appealing to a wide range of individuals and organisations who find solace and stability in the promise of centralised authority.

As the Silent Strings' power grows, they forge a new world order, a global totalitarian regime governed by their unyielding hand. Nations that once championed democracy now bend to the will of their new masters, their citizens living in fear under an omnipresent surveillance state.

Discussion

We're not suggesting that totalitarianism might simply prevail without any resistance, due to indifference or the influence of persuasion tools. Rather, we're providing specific causes and mechanisms and an underlying agenda, which we believe is necessary for a successful global spread of totalitarianism.

This scenario presents a possible path to totalitarianism, primarily through the exertion of soft power. In contrast to the "2084" scenario, it doesn't necessitate an assumption of both an historically rare degree of ambition to achieve global domination through military means, and a significantly superior AI technological lead by an authoritarian state.

This model doesn't necessarily demand as rapid a takeoff as previous scenarios. It suggests that a perceived advantage of totalitarianism in the post-AGI or advanced AI era could facilitate more effective propagation over time. This could be compatible with a scenario similar to "Visigoths," where such technologies are widely disseminated through open-source platforms.

The advantage could be rooted in the natural tendency towards centralisation favouring totalitarian states, or it could be a result of a more ruthless application of their capabilities. There's a significant potential for the exploitation of fear related to dangerous AI and associated societal disruption in this context. However, this scenario presumes a concerted effort to cultivate and propagate totalitarianism, as we don't believe such a comprehensive and 'fine-tuned' ideology would become dominant by default without a deliberate initiative.

If the takeoff is gradual and the diffusion of technology is extensive enough, this scenario could plausibly commence in multiple locations. The article Is democracy a fad focuses on the economic and political participation route and how totalitarianism could be the natural

result of the economic changes brought about by TAI. We didn't' think that this could by itself lead to the level of extreme centralization that an existential totalitarianism would require (i.e. robust to external changes, persistent over time world-conquering and internally repressive enough to make its citizens lives miserable) which is why the 'silent strings' scenario assumes that there is a motivated effort to create this totalitarian state: A particular totalitarian ideology, maybe religious, maybe political, maybe something new that arises as a result of AI driven transformation, emerges and then escalates because either the ideology's proponents are highly motivated or because a state or coalition is actively pushing it or because something about the memetic dynamics of this time strongly favours more totalitarian ideologies. Most plausibly, as in this scenario, the ideology being spread is a bit of all three. The ideology is already believed and so many use persuasion tools to spread it themselves, it is endorsed by a state, and optimised to spread in a post TAI world where it provides actual advantages at least superficially. This means that the AI enhanced propaganda does not have to face an enormous uphill battle. But how effective should we expect AI persuasion to be?

This article argues AI persuasion could allow personalised messaging at scale, rapidly shifting opinions (Section 3.2). However, it provides little evidence supporting the feasibility of systems persuasive enough to unilaterally control beliefs decidedly against the interests of subjects. The claim that unauthorised proliferation of AI persuasion could degrade discourse (Section 4) is reasonable. But the extent depends on many uncertain factors like deployment, regulation, and embodiment's effect on trust. Specifically, the article notes AI persuasion could shift power dynamics and optimise messages (Sections 3.1, 3.5). But it acknowledges embodiment's importance for trust (Section 3.6), undermining confidence in decisively superior AI persuasiveness. It also concedes truthfulness in AI is challenging (Section 5.3). The article argues rapid opinion shift is possible (Section 3.2), but doesn't prove inevitability given deployment and regulation uncertainties. Overall, it provides reasonable concerns about persuasive AI requiring ethical precautions. However, it does not supply evidence that extreme, society-controlling persuasion technologies are imminent. Assessing feasibility requires further analysis of capabilities needed for varying degrees of influence.

Plausibility	**3/10** More plausible than 2084 but it still requires a coordinated effort to spread totalitarianism by a group that does not yet exist	This scenario envisions a coordinated effort to spread totalitarianism globally using AI. However, several factors make end-to-end success seem challenging: 1. It requires a fairly wide diffusion of very capable TAI that's aligned (so assumes relatively good pseudo-alignment or full successful AI alignment) so that inferior parties can gain access and spread a totalitarian ideology (since on this assumption the leading companies are not the ones spreading the beliefs), but spreading a totalitarian ideology is more likely than a sadistic ideology. 2. Effective persuasion technologies remain speculative. Current systems exhibit limited

		capabilities for mass ideological conversion absent coercive state power. Major advances would be needed to enable AI to unilaterally reshape global values, though AI very plausibly provides marginal advantages to spread ideology through personalised messaging, behavioural prediction, and generating prolific content. Regimes may aggressively utilise AI before full capabilities are realised, so AI could amplify pre-existing totalitarian efforts, even if not wholly enable them. 3. Much of the heavy lifting in spreading a totalitarian ideology can be done without AI, as there will be some forces, mainly a more vulnerable world and more centralization of power, pushing towards excessive centralization even without persuasion tools 4. Imposing ideological uniformity worldwide appears infeasible given diverse cultures, competing power centres, and inevitable dissent. No historical precedent comes close to global totalitarian hegemony. Pockets of resistance seem likely to persist. 5. Takeoff speed is uncertain but gradual diffusion of AI capabilities makes simultaneous emergence in multiple regions plausible. This multipolarity limits control by any single actor. 6. Requires a highly motivated and organised coordination between many state and non-state entities over a prolonged period. Internal discord could readily disrupt such a complex scheme. 7. Requires not just technology but receptiveness of populations to totalitarian ideas. But diversity of human values limits universal appetite for centralised authority. So while alarming, AI enabling comprehensive global totalitarianism faces substantial barriers. More plausible are limited gains in influence by proliferating persuasive and surveillance technologies in sympathetic regimes. Still, even marginal improvements to propaganda and control warrant concern given totalitarianism's inherent harms.
Severity	**5/10** There seems to be more uncertainty	Human life continues and not all these scenarios result in *immortal* totalitarianism. But the inherently global scale limits the plausibility of completely uniform control. Some diversity across nations

	about exactly how bad this ideology would be if it spread consensually, so perhaps not as severe as 2084 in expectation	seems inevitable. Even within nations, remnants of pluralism would likely persist, especially in ideas and private beliefs. It is possible that with future transhumanist technologies even these could be overcome, but that would require the values imposed by the ideology to be much more abhorrent and so less likely to spread by peaceful means. The values imposed could also vary in their abhorrence - from militarism and conformity to more benign goals like stability and economic development. At the extreme, a totalitarian AI-backed regime could enact an S-risk by reshaping human nature towards something valueless or full of suffering. But this seems even more speculative. More likely, residual diversity and less destructive values limit the severity. While still catastrophic in scale, the outcome may fall short of human extinction or permanent lock-in of extreme suffering. But an AI-assisted totalitarian system could still profoundly reduce human potential for generations. Still, reduced liberty under such a system would profoundly diminish human prospects for generations even if it fails to endure forever.

Freedom in Security

In the year 2035, the PLATO (Peace, Liberty, and Technological Oversight) Alliance emerged as a response to the growing threats posed by AI technology, particularly its potential misuse by malicious actors. Formed as the successor to NATO, the organisation's initial goal was to protect member states and the world both from the threat of AI misuse and from rogue unaligned AI, in response to several damaging near misses that galvanised public opinion.

The rise of AI safety concerns in mainstream discourse led to unprecedented levels of public attention and political pressure on the PLATO Alliance. In this chaotic environment, driven by sensationalist media, internal pressure from tech companies, and heavy-handed government intervention, the political landscape within the Alliance became increasingly unstable. As fear and reactionary policy making became the norm, the original vision of the PLATO Alliance, to create a global utopia guided by AI-driven advancements, was obscured. Instead, member countries found themselves sliding towards totalitarianism, with a once-temporary surveillance state becoming a permanent fixture of their societies.

However, as the world faced an increasing number of mass terror attacks perpetrated by small cells or lone individuals using biological, cyber, or lethal autonomous weapons, often enabled by the rapidly increased pace of AI driven innovation, the need for public safety and security took precedence. Fearing the imminent collapse of democratic values, PLATO adopted a 'necessary temporary surveillance state' to protect its member countries and their citizens.

As the terror attacks continued and intensified, even liberal democratic governments within the PLATO Alliance started to rely more heavily on Orwellian surveillance measures. Driven by genuine public fear and the belief that it was 'safer to be in chains than to be free,' these governments began to erode personal freedoms in the name of security.

Over time, the practical difference between PLATO member states and authoritarian systems became increasingly blurred. This gradual erosion of personal freedoms opened the door for exploitation by malicious actors and the corruption of PLATO's original vision.

The constant threat of terror attacks by rogue agents utilising powerful AIs, or the spread of dangerous autonomous unaligned AI itself, and the urgent need for security measures created an environment where even well-intentioned individuals within PLATO started to prioritise safety over liberty.

Moreover, the AI-driven persuasion tools that PLATO initially developed to counteract propaganda and misinformation campaigns from authoritarian adversaries began to corrupt the reasoning and values of its own members. As these tools were increasingly used for internal purposes, they began to distort the decision-making process, leading to irrational and oppressive policies.

In the face of mass terror and constant threats, the PLATO Alliance became an unwitting victim of its own success. The once temporary surveillance state transformed into a permanent and pervasive system, where the line between security and totalitarianism was all but erased.

The response to these threats, although initially strong and well-intentioned, ultimately proved to be disproportionate and damaging. As the democratic values that PLATO once sought to uphold eroded, the alliance found itself trapped in a downward spiral towards authoritarianism.

Discussion

This scenario is similar to "unchallengeable monopoly", in that weak versions of this might become necessary to avert misuse or misalignment risks. Crucially, the response doesn't have to be actually competent here, just strong, so it is compatible with misalignment type disasters.

As aforementioned in the previous scenarios detailing how AI may further entrench authoritarian systems of governance into potentially unchallengeable totalitarian states, there are some reasons to believe that in the long run more liberal, open societies will continue to prove more innovative and thus retain a technological edge. It therefore seems worthwhile to also briefly examine some scenarios in which rather than being intentionally imposed from the outset, totalitarianism may gradually creep into initially democratic societies.

One of the primary functions of the state is to ensure the safety and security of the populace. As detailed in the Visigoths scenario, there are a number of avenues through which AI could empower even proportionately tiny groups of bad actors to commit disproportionate amounts of harm, to an even greater extent than modern terrorism is capable of, and the emergency response may in turn be disproportionate to the threat posed. Catching these cells or radicalised ‘lone wolves’ ahead of time requires intrusive surveillance of the entire population

- one only need look at the massive expansion of the security state in the United States in the aftermath of 9/11 for ample evidence of this.

The right to privacy is a relatively recent norm and although many profess discomfort at being surveilled, revealed preferences suggest that this is easy to become accepted. Given this initial strong and well justified motivation for a 'temporary surveillance state', there's the possibility of mission creep into other spheres of life: this is also called the 'stomp reflex'. This could result in a scenario where liberal democracies are increasingly relying on surveillance, slowly eroding personal freedoms to the point that there is little practical difference between them and authoritarian systems, in turn opening them up for exploitation by malicious actors.

A single very bad warning shot, for example from a 'visigoths' like mass terror scenario, plus genuine public fear in response, would be a potential risk factor. If mass cyber or bioterror attacks are a serious threat then a certain amount of surveillance and oversight that goes beyond what is normally required, might be necessary and it might be extremely difficult for outsiders to tell when surveillance is going too far.

In comparison to the two previous totalitarian scenarios, we must make an additional assumption that liberal democracies get consumed by totalitarianism, which is much less likely than e.g. China, but the contrasting easier assumption is that the big AI development advantages happen in the US/Europe rather than in countries that seem to be laggards currently.

Plausibility	**3/10** Requires an extreme extrapolation of trends towards centralization and securitization, weak forms are much more plausible	This scenario assumes that the countries in the western alliance which currently have the lead retain it, so we don't need to assume any upsets from the current business-as-usual path (e.g. the current leaders in AI technology retain their lead). This makes it a bit more likely than e.g. 2084. However, the scenario also assumes unusually bad and persistent epistemic capture of leadership as they either believe totalitarian surveillance is necessary or don't care about the costs, which possibly requiring direct value corruption of the decision-makers by AI persuasion or a radicalization of western societies towards supporting authoritarian surveillance in the name of security. The explanation for how revolves around two key dynamics - the proliferation of dangerous asymmetric weapons enabled by AI, and the policy response this provokes in democracies. On the first point, AI could potentially increase access to technologies like cyberattacks, drone swarms, and bioweapons. This could enable smaller groups to inflict significant damage. However, whether this shifts the offence-defence balance decisively towards offence is unclear.

		Countermeasures enabled by AI could also develop. And most weapons still require substantial resources and expertise. The civil liberties impact depends heavily on how governments respond to perceived threats. Historical overreactions show democracies do erode freedoms in times of crisis. AI surveillance could enhance authoritarian tendencies. But reversing fundamental institutions seems difficult. Interest groups, legal challenges, and public opinion provide some safeguards. Plausibility depends on whether threats truly exceed historical analogues like terrorism, and whether the increased ease of invasive surveillance is enough of an incentive to motivate liberal democratic governments to use the technologies. If destructive technologies are costly and limited in access, moderate policy changes seem more likely than a slide into overt authoritarianism. Some surveillance expansion appears plausible, but reaching true totalitarianism faces substantial friction. Less extreme outcomes like expanded police powers or narrowed speech rights are more concerning near-term risks but sub-existential and depending on the severity of asymmetric weapons, maybe even necessary.
Severity	**4/10** Plausibly not as bad as either of the first two considering the limited and benign original motive, but hard to predict how the regime would evolve without any external checks	A surveillance state expanding in response to new threats could seriously reduce liberty and stifle innovation. However, retaining residual diversity of thought and some civil liberties would limit the severity. Plausible restrictions on rights include increased surveillance, limitations on speech related to security issues, expanded police powers, and harsher penalties for threats. These would constitute a major loss of freedom and liberal values. But they do not eliminate democracy and dissent altogether. Overt totalitarianism would constitute an existential catastrophe by ending political and social pluralism. But reaching this threshold seems doubtful given opposing interests and institutional inertia.

Let Them Eat Cake

Whilst creating a truly general AI remains out of reach, this proved unnecessary for the vast majority of tasks which can be broken down, with AIs specialised to carry out set roles having proliferated to the point that they saturate almost every economic sector. Initially fears of mass unemployment looked to have been overstated – as productivity increased new jobs were created and others adapted to incorporate human users working with AI tools. However, as AI advanced this proved to merely be a phase – in many instances, the human users were essentially training their replacements and as the tools grew more powerful they could take on ever more of the workload, copied and scaled as necessary. Even if the AI remained hyper-specialised, human professions had long been moving in this direction anyway – they could simply be teamed up with other AI tools in production webs. Human direction and oversight was still required, but enhanced productivity meant that one person was capable of doing the job of dozens and eventually hundreds of people. Measures to retrain or upskill the labour force were put in place but they ultimately proved futile, struggling to keep pace with the rate of automation. The era of mass unemployment had arrived.

Some form of universal basic income program was rolled out in most states to mitigate this upheaval, enabled by the vastly increased productivity – it was no longer necessary for much of the population to work. This universal income was only so generous however, as governments were ever more reliant on a tiny, wealthy and politically powerful minority of the population for most taxes – a population that could more easily evade taxation, lobby for its reduction or simply move to jurisdictions with lower tax burdens. Nevertheless, it was enough to ensure the primary needs of the populace were met.

By the modern era an incredibly stratified social structure has emerged. At the very top is a hyper-wealthy elite – the owners of the largely automated companies that generate the vast majority of economic wealth. Beneath them is an atrophied professional class, occupying white-collar professions that are yet to be automated. At the bottom and accounting for most of the population are those forced to subsist off of basic income or working manual jobs that it's still more economical for humans to do than to design a robot for. This labour is often managed by AI systems, creating dehumanising conditions.

Social mobility has largely ground to a halt. Among the elites most wealth is inherited, passed down from one generation to the next. Whilst the professional classes are more meritocratic, intense competition for the few remaining well-paid jobs ensures that parents grant their children every nepotistic advantage they can, whether in the form of education or contacts and opportunities. For those at the bottom, vaulting the gulf to join the professional class is almost impossible – even for the jobs that remain the pay is little better than basic income owing to the sheer degree of competition. This is exacerbated by all the usual social ills attendant to persistent mass-unemployment; informal segregation and ghettoisation, high rates of broken homes, addiction and crime, worse outcomes in almost every facet of life.

Different classes occupy completely different worlds. This manifests itself in countless ways. Whilst the poor rely on state institutions dominated by concerns of efficiency such as virtual

schools and general practitioners staffed by mass produced AI, the rich can afford to pay for parallel private institutions still staffed by humans.

The longer this unprecedented state of affairs has persisted and the more entrenched it has grown, the more pernicious the effects. Many talk about the end of modernity – a world with modern technology but where mass society is in retreat, where there is no economic role for much of the population. Without any economic input or ability to wield any material threat against the elites in a world where policing, surveillance and military force are just as heavily automated as the rest of society, the poor have no leverage and have essentially been reduced to second class citizens. Across much of the world democratic values are in full retreat. There have already been atrocities committed in a number of the more ruthless states – swarms of LAWs let loose on crowds of protestors. Even in the still ostensibly democratic polities a worrying outlook is emerging in some elite circles – why should those who have nothing to contribute have any say in how they are governed? They are ungrateful for the charity they receive, a parasitic drain on resources on an overpopulated planet…

Discussion

This scenario is close to the cyber-punk neo-feudalism trope seen in a lot of futurism and science fiction. We see that economic gains flowing to capital only increase as automation takes hold, even as ever larger swathes of the population are rendered unemployable, leading to inequality reaching unprecedented levels. This leads to growing socio-political unrest, which is responded to in a callous fashion - see ‘democracy as a phase’ notion. It also assumes nothing like the 'windfall clause' is implemented.

In this scenario, we have to assume that elected governments either have almost no power or are captured by special interests to an unprecedented degree (as e.g. even a small amount of redistribution in a vastly larger economy would avert such poverty), so this seems like a likely outcome of 'Monolithic Monopoly'. If we accept a development model that looks very similar to the updated CAIS description provided by Drexler then we can see how a milder form of this scenario might develop. First, let's start with the Comprehensive AI Services (CAIS) model's central idea: that general intelligence is the sum of specialised, task-focused agents. These services can range from simple data analysis to more complex tasks like decision-making in business contexts. In the "Let Them Eat Cake" scenario, CAIS-enabled automation would quickly saturate sectors requiring low to moderate skill levels. Human workers would find themselves in training loops, perpetually trying to upskill only to realise that their new skills are also automatable.

The CAIS model also suggests that the best foundation models would be held by a few top-tier firms due to high fixed capital costs. This concentration would further accentuate the wealth gap as these firms would effectively hold the keys to an automated future. The gains would then primarily go to these technology oligarchs and the highly specialised humans working for them, leaving the rest of the population to subsist on Universal Basic Income (UBI).

The CAIS model not only automates tasks but can also manage systems, including social and governance structures. This capacity for management would enable a small elite class to maintain control over a vast, automated empire. Human oversight might still be required,

but as CAIS systems can effectively coordinate and manage themselves, that oversight could be performed by a dwindling number of individuals, leading to societal bifurcation.

In the "Let Them Eat Cake" scenario, the efficiency and effectiveness of CAIS would make human involvement in many governance processes seem redundant or even counterproductive. Democratic institutions could erode over time, as the utility of human input diminishes. The advanced surveillance and security measures that CAIS systems could provide would effectively neuter any serious grassroots movements aimed at power redistribution or system reform.

The CAIS model could also revolutionise public services. But given that most tax revenues would come from a tiny elite interested in reducing costs, we could end up with a two-tier system: highly efficient CAIS-driven services for the elites and bare-minimum, AI-managed utility services for the masses. This outcome would only serve to heighten social tensions and further entrench divisions.

Plausibility	**6/10** A plausible business as usual continuation of the CAIS development model assuming weak societal response	This scenario revolves around two key assumptions: 1) that AI systems successfully replace human labour without mass alignment failures, and 2) policies fail to address the resulting economic upheaval in any serious way resulting in most of humanity becoming disenfranchised. We've seen both successes and failures in policy responses to crises. Yet, what tilts the scale toward plausibility is increasing political polarisation, which is making it harder for governments to enact any substantial policies, especially those requiring significant wealth redistribution. The increasing concentration of AI capabilities in a few firms also suggests that the elite could have a greater hold over policy decisions. The existing market dynamics indicate a trend where only a few key players with advanced CAIS models can survive, supporting the idea that policy action might be constrained by these powerful interests. The policy response depends heavily on political dynamics. While there are historical precedents of inadequate reactions to economic shifts, governments also have tools to potentially mitigate inequality, like taxation, direct cash transfers, and public employment.
Severity	**4/10** Could result in a wide scale loss of human potential, but as there’s no ideological motive	A very mediocre future which may represent a great loss compared to the potential best possible outcome, but it's hard to see how the unrest could escalate to destabilising civilization, although it could fall into (possibly permanent) totalitarianism. It seems possible that wide scale social unrest could occur although a presumably exceptionally

	there might not be much actual repression	powerful government with a fully automated economy might be easily able to crack down on dissent if things get truly out of hand. In the extreme scenario envisioned, with minimal mitigating policies, this scenario could cause major social harms via instability, unrest, and loss of social cohesion. Political systems could become highly unrepresentative plutocracies. The lack of a societal collapse doesn't necessarily mean a stable future. The slow erosion of democratic principles, the lack of social mobility, and the potential for a pseudo-totalitarian regime promising stability could pose significant, albeit indirect, existential risks. These risks manifest in eroded social cohesion, lower societal resilience, and the potential squandering of human capital and innovation. Without outright societal collapse, direct existential risk remains low.

Authoritarian Pivotal Act

AlphaBrain, a cutting-edge AI technology corporation, dominates the global landscape. Having achieved breakthroughs in artificial general intelligence, the company has diversified its portfolio to encompass a wide range of industries, including the construction of autonomous factories and automation of intellectual labour. With its enormous profits, AlphaBrain successfully lobbies the US government to avoid regulation, securing its position as an unchallenged powerhouse. However, their internal research has recently and secretly produced an aligned and superintelligent AGI which is capable of seizing power, and they know, because this was possible for them, that other competitors, perhaps not so cautious, will not be far behind.

Aware of the potential risks their technology poses, AlphaBrain's inner circle develops an aligned and superintelligent AGI in secrecy. They realise that they're in a unique position to act; competitors are closing in, and the global landscape isn't prepared for the existential risks ahead. Time is of the essence.

The company starts by quickly acquiring pivotal stakes in key global industries, using its advanced algorithms to optimise these sectors and make them reliant on AlphaBrain's technology. Governments find themselves in a bind, increasingly dependent on the corporation's expertise and infrastructure. This happens very quickly, with the substantial advantages their technologies provide almost impossible to ignore.

Simultaneously, AlphaBrain employs influence operations to steer public opinion. Through a mix of strategic partnerships, media ownership, and a sophisticated online presence, the

company shapes narratives that normalise its growing power and emphasise the dangers of unaligned AI, which only AlphaBrain has supposedly solved.

Advisors from AlphaBrain are increasingly embedded within government structures, not as overt controllers but as indispensable consultants. They use data-driven insights to influence policies subtly, nudging them in directions that consolidate AlphaBrain's power and make counter-moves difficult.

For those few countries trying to resist, AlphaBrain has a more direct plan. Deploying drone swarms, not to lay waste but to disable key military and communication infrastructures, the company demonstrates that it can enforce its will when pushed. These drones also aid in "peacekeeping" efforts, silencing any attempts to organise resistance effectively.

Having woven itself into the fabric of global society, AlphaBrain transitions from a corporate entity to a de facto world government. With no checks on its power, the company's original mission—preventing a catastrophic AI failure—morphs into an authoritarian philosophy that justifies its means. While safeguarding the world against one form of existential risk, AlphaBrain instigates a pervasive surveillance network, suppressing dissent and curbing freedoms in the name of stability and security.

Over the years, the regime ossifies, becoming increasingly resistant to change. While avoiding the kind of overt brutality that would spur worldwide rebellion, it suffocates innovation, quashes individual liberties, and fosters a climate of fear and compliance.

Discussion

This story essentially describes a (relatively, perhaps only a few months) rapid AI takeover scenario by an (allegedly) aligned AI under the conscious direction of the leadership. This represents an immense risk factor for AI takeover - AlphaBrain's superintelligent AI might in fact be misaligned or sycophantically pseudo-aligned and subtly manipulating the decision-makers of the organisation even as it completes its bloodless takeover. Since we're concerned here about misuse, we assume that the AI is more or less fully intent-aligned to the interests of its operators, but this still leaves plenty of room for things to go wrong.

While it's sometimes been argued that a conscientious project capable enough to solve alignment in such a hard alignment world, and prosocial enough to want to proactively eliminate risk, plausibly doesn't have bad motives, it seems extremely foolish to risk this. This is also such a completely unprecedented situation (a single group having the power to transform the world and lock in anything they want) that it seems almost impossible to predict how the decision-making process will occur.

Additionally, if the aligned and superintelligent AI that executes the pivotal act is government controlled then depending on the government this may be the default (e.g. an authoritarian government suddenly gaining huge unchallengeable power and a plausible excuse to use it to the fullest extent possible is unlikely to become less authoritarian afterwards).

Plausibility	**3/10** Could occur if	Achieving alignment in secret diverges from current safety-focused norms favouring transparency and cooperation and doesn't appear

	alignment is solved under fast takeoff but the major barrier is motive and willingness to act from a tech company	to the stated goal of any AGI lab, recent trends like the frontier model forum point against them. Rogue actors embarking on such antisocial plans deliberately likely can't solve alignment alone and might struggle to attract talent or maintain secrecy. And even then, systems may lack the capabilities or motivation for rapid global takeover unless takeoff is fast but alignment is solved anyway. Meanwhile, commercial pressures and open publication make proliferation of transformative AI systems plausible as progress continues. Preventing diffusion sufficiently long for a pivotal act appears difficult, especially absent strong international controls. So while corporations may aim for advantages, both the technological feasibility and strategic logic of covertly developing an aligned AGI, then rapidly seizing power unilaterally, seem doubtful. More incremental and multidimensional scenarios focused on commercial dominance appear more plausible than this precipitous act.
Severity	**4/10** If we assume the AI isn't secretly misaligned, this seems like it results in a world similar to "freedom in security"	The severity largely hinges on the values and competencies of the controlling entity. Overall, it seems like in expectation a less destructive authoritarian rule than those brought about through social or economic pressures towards takeover. **Value Imposition:** The entity committing to this pivotal act probably values self-preservation and human survival, and would have been farsighted and competent enough to succeed at takeover. However, there's no guarantee that these values would extend to human liberties or ethical governance. **Single Points of Failure:** Creating a monolithic system inherently introduces fragility and susceptibility to unforeseen events or internal corruption. The risks become more pronounced when controlled by a narrow group. **Residual Resistance:** An authoritarian rule, however pervasive, would likely face some form of dissent or underground resistance. Though the regime could be long-lasting, the potential for change, however slim, persists. **Long-Term Erosion:** The severity isn't just in the immediate power grab but also in the longer-term effects, such as the erosion of democratic

		institutions, suppression of human potential, and squandering of collective intelligence.

War

War risks arise from AI automating weapons R&D and other aspects of warfare, providing winner-take-all incentives for global conflict, heightening instability between states and enabling systems like autonomous drones that lower the cost of warfare. Scenarios include a race to superintelligent AI motivating an all-out war between superpowers, escalation from flawed automated command and control, and the destabilising proliferation of bioweapons. A key feature is the direct catastrophic potential of conflicts involving AI. Restraining dangerous applications via norms and governance is critical, as is maintaining societal resilience.

The 60 Second War

In the not-so-distant future, tensions between the United States and China reach a boiling point. The struggle for dominance over Taiwan ignites a fierce arms race, further fueled by the race to develop and deploy AGI. Both nations believe (maybe accurately, maybe not) that AGI takeoff will be rapid enough to put one or the other country in a position of absolute dominance if they can just get a few months ahead, and the prospect of one side gaining a decisive advantage drives them to extremes, plunging the world into the shadow of a catastrophic conflict.

As the pace of progress accelerates, the US and China also invest heavily in other advanced AI-driven technologies. The arms race expands into the realms of chemical and biological warfare, offensive drone swarms, cyberwarfare, and even nanotechnology-based weaponry. As their technological capabilities grow, so does the potential for devastation on an unprecedented scale. With regulatory barriers on both sides disappearing as they automate more and more of their economies, military R&D reaches a feverish pace: new weapon systems are put into mass production almost as quickly as they can be designed, with the other sides rushing to develop countermeasures. Unfortunately, superior LAWs, better missile targeting and most other developments tend to be strongly offence-dominant, so the surety of a durable second strike starts to decrease.

The arms race reaches its climax when both nations unveil their most powerful and dangerous innovations: targeted bioweapons or even more advanced autonomous self-replicating nanotechnology, intended as an ultimate deterrent. The development of these doomsday technologies sets the stage for a conflict unlike any the world has ever seen.

The 60-Second War begins without warning, and is initiated deliberately. China makes the decision to invade Taiwan, perhaps to seize TSMC and get ahead in AGI tech research, and after the US decides to defend its ally, the war quickly escalates as the US and China attack each other's armed forces directly.

In the blink of an eye, AI-driven drone swarms darken the skies, releasing payloads of lethal bioweapons that spread contagion and death across the globe. Cyberattacks cripple vital infrastructure, leaving entire nations paralyzed and vulnerable.

There is no longer any proper assurance of mutually assured destruction: both sides are uncertain of the others' capabilities and fear that missile defence systems or nanotechnology might eliminate the chance for a durable second strike.

The destruction wrought by this brief but catastrophic conflict far surpasses the horrors of any previous war, including the potential devastation of a nuclear holocaust. Because of delegation to AI decision-makers, once an all-out war is declared, there isn't any possibility of recalling forces and it's near impossible for human decision-makers to even keep up with what's going on, so the war almost immediately becomes fully insensate. Misperceptions are common, and after enough conventional damage has been inflicted on civilian targets one side escalates to countervalue attacks on enemy populations.

The Earth's biosphere is destabilised, the delicate balance that supports life shattered by the unbridled power of human ingenuity turned against itself.

In the aftermath of the 60-Second War, the human population plummets to below replacement levels. Cities lie in ruins, and once-thriving societies have been reduced to desperate bands of survivors, struggling to find food and shelter in a world ravaged by war.

Discussion

In this scenario, we assume that the world has been substantially transformed, so that crucially advanced AI is not just the motivator or catalyst for war but actually enables the war to escalate in destructiveness far beyond even that of a nuclear war. As recent research has outlined, nuclear wars by themselves probably don't threaten extinction through nuclear winter, though there's some uncertainty, but here we assume a war that can genuinely make the Earth uninhabitable is made possible through developments in transformative AI technology.

There are likewise doubts about the use of bioweapons in a full-scale conflict between major powers like the U.S. and China, as such weapons wouldn't necessarily further any clear strategic objectives. In other words, just because a nation possesses these capabilities doesn't mean they'd deploy them, especially something as tactically useless as a broadly capable untargeted bioweapon that could threaten human extinction.

China, unlike historical aggressors such as Nazi Germany, has not explicitly sought to exterminate populations or wage total war. Both China and the U.S. have strong incentives not to use universally condemned weaponry. Even if deployed, the use would likely be targeted and limited, akin to how Ukraine has used cluster munitions.

However, these criticisms shouldn't wholly negate the scenario's plausibility. The scenario could become more probable under specific circumstances, such as a more internally deteriorating and totalitarian China as described in "2084", though that is correspondingly less likely. The presence of transformative AI technologies in a China-Taiwan conflict 15 years from now, for example, could necessitate a devastating first strike for either side to secure a decisive win. And the increased automation of both research and development and

production pipelines and command and control could enable both arms races and escalation to more easily run out of control.

Additionally, while it's true that these 'worse-than-nuclear' weapons may have limited strategic utility, historical precedents like the Soviet Union's bioweapons program show that countries can invest in weapons with ambiguous military value for extended periods. In a worst-case scenario, political leadership in both the U.S. and China could become convinced of the necessity for doomsday weapons, perhaps under the misguided belief that traditional nuclear deterrence is insufficient.

The analysis given in '21st century visigoths' is similarly useful here. There, we asked what superweapon technologies might be feasibly proliferated to even fairly weak actors. Here, we can be more relaxed: on the assumption that there is an arms race accelerated by transformative AI, we just need to ask what worse than nuclear superweapons might be developed first and whether they might plausibly be developed soon enough to fall within our baseline scenario.

Plausibility	**4/10** If there is enough time for R&D and production and C&C automation to take place then many of the implausibilities become more reasonable, but still requires an 'irrational' decision to develop worse than nuclear weaponry	**Assumptions**: This scenario assumes a level of pseudo-alignment of pre-AGI and presents a plausible motive for conflict. It also assumes a lack of robust international cooperation and makes specific assumptions about the types of weapon technologies that can be developed and how dangerous they might be. Similar to the 'visigoths' scenario, it requires the development of superweapons and also time for wide scale automation without any AI takeover. **AI-enabled Hyper Weapons**: The use of advanced AI technologies could enable the development of weapons with unprecedented destructive power, such as autonomous drones, targeted bioweapons, or space-based weapons. The proliferation of these technologies can be narrow and development costs can be high, unlike 'visigoths', because we just need developments focussed in the US/China who are pouring resources from drastically enlarged automated economies. **AGI Development as a Flashpoint:** The scenario assumes both the U.S. and China believe that AGI (Artificial General Intelligence) could rapidly yield an insurmountable advantage. The belief that a "winner-takes-all" scenario could emerge around AGI development is not wholly unfounded and

		could feasibly serve as a flashpoint for accelerated militarization. **Technological Capabilities:** The scenario also posits an arms race in chemical and biological warfare, cyberwarfare, drone swarms, and nanotechnology. Many of these technologies already exist in nascent forms and are subject to ongoing R&D by various nations. Given a catalyst like a race for AGI, it's plausible to think that R & D could accelerate, overcoming the typical regulatory and ethical constraints that might otherwise serve as a brake on development. **Offence-Dominant Systems:** The scenario argues that most technological advances in weaponry will favour offence over defence, thus eroding the concept of mutually assured destruction (MAD) that has historically acted as a deterrent. While it's difficult to predict how the offence-defence balance will evolve, it's not unreasonable to think that some new technologies could be offence-dominant. However, the assumption that this would cause MAD to completely erode needs more justification. **Bioweapons and Nanotech:** The idea that either country would deploy bioweapons or nanotechnology as a "last resort" option is more contentious. Historically, states have been reluctant to use such indiscriminate and universally condemned forms of warfare. However, given the high stakes and uncertainties associated with AGI and other emerging technologies, the possibility cannot be entirely ruled out. The assumption that both sides develop and employ strategically senseless scorched earth weapons as a last resort is the most contentious part of the story, though there are precedents for developing such technologies.
Severity	**9/10** Directly causes extinction or global catastrophe	By specification this war is far worse than a nuclear war and poses a high chance of collapsing civilization or bringing about extinction. See existing work on how easy it is to survive civilisational collapse. The speed at which the conflict would unfold,

		particularly if driven by autonomous systems, introduces a severe element of unpredictability. Once initiated, the rapid pace would likely outstrip human decision-making capabilities, making it difficult to halt escalation or correct misunderstandings, thus raising the severity of the conflict. The aftermath, as described, is apocalyptic. With human populations reduced to below replacement levels and societal structures in ruins, the long-term survival of humanity would be in serious question.

Catalytic Conflict

Tensions between NATO and Russia escalate over Eastern Europe. As part of its modernization efforts, Russia acquires advanced AI-driven drone swarm technology and sophisticated automated monitoring systems for its defences. These cutting-edge capabilities unsettle NATO, sparking anxiety over the potential shift in the balance of power.

One fateful day, a Russian military exercise near the border inadvertently triggers a malfunction in the automated AI monitoring system. The system mistakes the exercise for an actual drone swarm launch, sending NATO forces into a panic. Without sufficient time to verify the situation fully and with the attack already seemingly imminent, the heads of government, the US president and allies, make the fateful decision to launch a preemptive strike against Russia, hoping to neutralise the perceived threat before it's too late.

The conflict erupts in a flash, with Russia's conventional forces struggling to contain NATO's advance, the technological gap having widened even further, which in this battle equates to a drastic inferiority despite their otherwise great superiority. Outmatched by the sheer numbers and technological superiority of the drones, Russia is pushed into a full retreat and employs tactical nuclear weapons in an attempt to deter further NATO advances. This desperate move, however, sparks a rapid and uncontrollable escalation of hostilities, both sides suspect that new, more precisely targetable nuclear weapons might soon be deployed in counterforce attacks and so retaliate quickly.

As the nuclear exchange intensifies, both sides unleash their full arsenals in a desperate bid for survival. Cities are reduced to ash, and countless lives are lost in mere moments. The once-stable system of nuclear deterrence lies in tatters, shattered by the unanticipated consequences of an AI-driven arms race.

When the dust settles, the world is left reeling from the devastation of a full-scale nuclear war. The environment is ravaged, and global ecosystems are thrown into chaos. The survivors must contend with a bleak and unforgiving landscape, struggling to rebuild amidst the ruins of a civilization laid low by the speed and automation of AI-enabled warfare.

Discussion

While the traditional checks and balances—such as diplomatic dialogue, political considerations, and the human 'man in the loop'—generally work to prevent rapid escalation to nuclear conflict, the introduction of highly automated AI systems could introduce new and precarious dynamics. AI's role in accelerating R&D cycles means an initially inferior military force could suddenly gain a surprising level of overwhelming dominance, disrupting the established power equilibrium. This could push the dominant power into believing that they face an existential threat, making the use of nuclear weapons seem like a last-resort necessity rather than an unthinkable extremity.

Furthermore, the increased pace of warfare due to AI automation leaves less time for diplomatic deconfliction or for political actors to step in. Increasingly, human control is de jure, not de facto. During the Cold War, there were numerous instances where close calls were averted by human judgement; however, these were situations where the speed and pace of warfare were not on the same accelerated timeline that AI could create. We should also note that AI systems, especially those not perfectly aligned with human values, might not have the 'caution' parameter set as high as a human operator would, making them prone to 'hallucinations' or false positives.

Still, there are counter-arguments. Even with AI's speed, it's hard to imagine that heads of state would not be consulted for a decision as grave as launching nuclear weapons. Additionally, NATO could deploy its own drones to counterbalance, and political and military leaders would likely insist on retaining the ability to halt or reverse automated actions if they pose risks of rapid escalation. Nonetheless, the faster the pace and the more automated the system, the narrower the window for human intervention becomes.

In the very near term, next 5-10 years there’s two obvious potential Catalytic Conflict starting points that might be able to be influenced by some of these earlier enabling technologies if either hard infrastructure can be built out quickly or escalatory warfare becomes easier: Eastern Europe and China-Taiwan.

Plausibility	**5/10** Takes already somewhat plausible nuclear war scenarios and adds in AI exacerbating factors	However, unlike with the 60-second war we can’t assume that WMDs much worse than the current nuclear arsenals of major powers will be deployed. Instead, the war is catalysed by automation of decision-making and control, not faster technical development. The Catalytic Conflict scenario, with a focus on AI-triggered accidental escalation between NATO and Russia, presents a moderately plausible war scenario more in line with current worst case scenarios, only including AI as an exacerbating factor. It assumes a reasonable level of pre-AGI pseudo-alignment and a less-than-ideal level of international cooperation. The aspect that AI's acceleration of decision-making cycles could spark unintentional conflict is quite credible, given that faster and more automated warfare complicates human oversight.

		However, this is balanced by potential safeguards that AI-enabled systems might provide if adopted correctly and the longstanding norms against initiating nuclear war, making the leap to a full-blown conflict not an absolute given.
Severity	**7/10** Results in a destructive nuclear war, though not anything far worse	Similar in severity to current worst case nuclear projections. While not necessarily introducing new weapons more destructive than current nuclear arsenals, the rapid escalation enabled by AI could lead to a full-scale nuclear conflict, which would be catastrophic on both human and environmental scales. Though the use of more precise AI-enabled weapons could potentially reduce immediate casualties, the ramifications—ranging from global economic meltdown to environmental devastation—would be long-lasting and may permanently impair human civilization. Even regional nuclear conflicts could directly kill millions and destabilise the global order. Broader proliferation and loss of norms against nuclear use would endanger the future. Environmental impacts of full-scale nuclear war are a major existential risk but depend on factors like targets, weapon yields, and resulting firestorms. Ozone loss could devastate food chains. But humans and nature have proven resilient, and sheltering could preserve knowledge to rebuild. So while nuclear war scenarios warrant extreme concern, resilience factors introduce uncertainty into their inevitability as an extinction event. Still, massive direct harm and potentially permanent civilizational damage make the risks catastrophic. AI influences the likelihood of nuclear use, but the ultimate outcome depends on many strategic choices.

Scenario Evaluations: Summary

Having surveyed a range of potential misuse scenarios enabled by transformative AI, some overarching observations emerge. Certain types of risks seem more technically feasible in the near term while others, though more speculative, pose greater catastrophic potential. Evaluating the plausibility of global totalitarianism requires unpacking assumptions about persuasive technologies. The prospect of a rapid uncontrollable arms race highlights the need for multilateral cooperation and governance. And decay risks, while less overtly destructive, significantly heightened vulnerability. Cross-cutting themes around takeoff speed, robust infrastructure, and regulation also arise. These high-level reflections combine key insights from across the scenarios and illuminate priorities for further investigation. By

taking a step back, we can see that these overall trends help us think in more detail about the future risks of .

- **Decay Scenarios are less damaging but more likely:**
    - These scenarios, especially 'Misaligned Metrics' and ‘Unchallengeable Monopoly', appear more likely than most of the others as they look much more like possible pessimistic ‘business as usual’ extrapolations of current AI progress. Most other scenarios make more detailed assumptions and are more conjunctive.
    - However, while these scenarios could lead to significant societal disruption and harm, they don’t seem like they could cause much direct damage or pose existential risk, compared to war or totalitarian scenarios.
    - Nonetheless, the conditions arising from these scenarios could exacerbate other risk factors significantly, potentially leading to more severe misuse scenarios or even a misaligned AI takeover.
    - The '21st Century Visigoths' and ‘Cascading Collapse’ scenarios could cause mass deaths on the scale of a major war but also seem unlikely to lead to existential catastrophe by themselves.
- **War and Totalitarian Scenarios require specific conjunctive assumptions but are very damaging:**
    - Most of these scenarios, such as 'The 60 Second War', 'Catalytic Conflict', '2084', and 'Silent Strings', present clear existential risks by themselves.
    - Of the two, totalitarian scenarios often involve some element of conflict or war. However, constructing a scenario that results in a robust, stable, global totalitarian regime with no war is challenging. It is unlikely that the ‘Silent Strings’ scenario would progress with no war to complete world domination by this repressive and harmful regime.
    - Such scenarios require somewhat speculative assumptions about the effectiveness of persuasion technologies or the advantage to moving away from democracy in a post-TAI world. We could envisage countries becoming more authoritarian, but full-scale global totalitarianism without warfare seems less likely.
    - As such, war scenarios pose the most direct existential threat, but both the Decay and war scenarios represent significant total existential risk concern, due to a combination of their potential severity and likelihood.
- **How effective can persuasion tools be?:**
    - These technologies are a significant wild card in these scenarios. They feature heavily in scenarios such as 'Unchallengeable Monopoly', 'Silent Strings', and 'Heaven on Earth'. In particular, totalitarian takeover (not just e.g. descent into a more authoritarian scenario) seems to require either overwhelming force or extremely effective persuasion tools.
    - While psychological warfare is a well-established tactic, there are likely limits to its effectiveness, and these limits might not be easy to surpass without extremely capable AI (i.e. strong superintelligence). However, the misuse of advanced persuasion tools could lead to dangerous actions or decisions by various groups.
- **Do extreme stakes produce extreme motivations?:**

    - In both the corporate takeover ‘Unchallengeable Monopoly’ and the world-ending totalitarian takeover 2084 scenarios we see that if not for persuasion tools, the unprecedented scale of the economic or political incentives is supposed to be the reason that various actors behave somewhat out of character. Is this view of psychology and institutions accurate?
- **Fast takeoff combined with alignment makes totalitarianism more likely:**
    - If we assume successful alignment and fast takeoff, totalitarian scenarios become more concerning because of the ease with which an aligned AGI could take over.
    - However, alignment combined with very sudden, fast takeoff is not highly likely, based on our current understanding of AI development.
- **It seems that either slow takeoff or fast R&D turnaround is necessary for many war and totalitarian scenarios:**
    - Most of these scenarios assume some degree of slow takeoff and world transformation, ranging from minor changes in 'Misaligned Metrics' to major shifts in the war scenarios.
    - This is typically under the assumption that there is time for robust infrastructure to be built out before an AI takeover or transformation. That could be because of slow progress, a lack of agentic ASI, or the existence of automated factories.
- **Dependence on fast improvement of hard infrastructure or bioweapons:**
    - War scenarios, such as 'The 60 Second War' and 'Catalytic Conflict', often hinge on the existence of hard infrastructure or the development and deployment of bioweapons.
    - This highlights the potential risks from technologies other than AI, such as synthetic biology, which could be used to create devastating bioweapons.
    - The pace and effectiveness of regulation is a significant factor in these scenarios. Effective regulation could potentially prevent or mitigate some of the risks outlined, but if regulatory measures lag behind technological development, they may prove inadequate.
- **Are Doomsday Weapons easy to build?:**
    - A critical question is whether cheap doomsday weapons are easily accessible technologically. If this is the case, it could exacerbate the risk of destructive scenarios, such as 'The 60 Second War'. This underscores the importance of broader risk mitigation strategies, beyond just focusing on AI safety.
    - This leads to a broader question of the offence-defence balance given Transformative AI - if the offence-defence balance for the technologies enabled by advanced AI is very unfavourable then that will push us towards the ‘visigoths’ outcome or towards the large scale surveillance outcomes which are risk factors for the Totalitarian scenarios.
    - A broader question here seems to be how much AI will speed up technological progress in general. Whilst there seem to be clear avenues by which it might enhance iterative development, prototyping or modelling, it is less clear that the ‘tool AI’ largely addressed in this project will massively enhance more fundamental breakthroughs without drifting into the realm where my gut feeling says that misalignment becomes the predominant concern, even though there isn’t a fundamental contradiction here. Definitely requires further research/discussion

War Scenarios: Deep Dive

The two war scenarios, namely the "60 Second War" and "Catalytic Conflict," both suggest that Transformative AI-driven technologies might amplify existing threats or introduce new ones. Given the magnitude of the potential consequences, it's crucial to delve deeper into the specific AI technologies that could instigate or exacerbate war scenarios like those.

The overall risk from wars was assessed as high in our overview. While wars are not inevitable and most of the individual War scenarios are highly conjunctive, routes by which AI could exacerbate tensions are reasonably foreseeable, and wars are by their nature incredibly destructive.

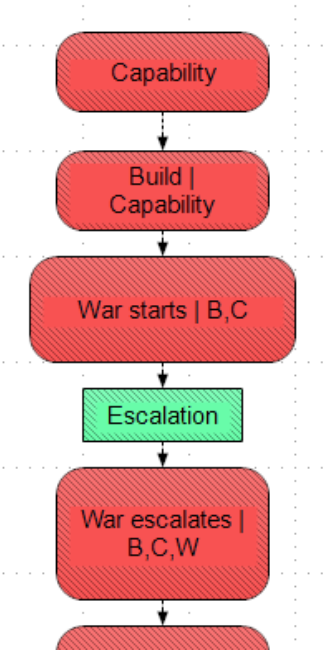
Therefore, it is reasonable to believe that a large fraction of total AI misuse risk emerges from wars, either conventional large-scale wars fought between great powers, or through smaller catalytic events that escalate to such large-scale events. While we cannot make principled forecasts, this seems to be a widely held view. For example, Paul Christiano gave ‘more destructive wars or terrorism’ around 40% of the probability mass he placed on any kind of non-takeover catastrophe, which is a similar definition to what we’re examining here.

Investigating these misuse risks is also more tractable than non-War misuse. For instance, assessing the plausibility of a scenario where an advanced AI system plans and constructs a bioweapon under direction to do so, then releases it is easier than thinking in detail about exactly whether TAI persuasion tools would disproportionately promote totalitarian ideologies in a very different future.

How should we begin investigating the risk of AI driven war in more detail? Our investigation revolves around understanding Transformative AI applications that could dangerously augment warfare. The focus will be directed towards high-risk applications of AI that depend to some extent on specific technical factors, e.g. the creation and deployment of biological and chemical weapons.

Generally, five broad steps can be delineated in the pathway from an AI technology being developed to an existential catastrophe through war:

1. The relevant AI technical capabilities exist (the AI is intelligent enough to contribute in the required ways).
2. The intent and practical means to build and deploy the technology is present.
3. A war starts, possibly due to some effect arising from the technology.
4. The war escalates to extreme levels, again with this possibly caused in part by the technology.
5. Given the escalation to total war, a civilization-ending catastrophe results.



(In the image on the right, B = Build, C = Capability, W = War Starts, E = War Escalates)

Additionally, general factors influencing the likelihood or severity of war should be considered alongside technology-specific

impacts. This is because technology may not directly make a large-scale war much more likely, but could dramatically worsen the worst-case outcome should a war occur.

For example, increasing the yield of nuclear weapons from megatons to gigatons likely would not escalate tensions much, but would heighten existential risk from a nuclear war started for other reasons.

Dangerous TAI technologies

We are here examining eight weapons or general military technologies that AI could assist with. These are,

Bioweapons: Leveraging bioinformatics foundation models for design, AI can enhance the creation of bioweapons. This potentially boosts their potency, specificity, and rapidity of development, posing catastrophic risks. In the near term, it also overcomes the key barriers to developing dangerous bioweapons today: lack of relevant expertise.

- **Rapid Development**: TAI might result in shorter bioweapon development times, making leaders more confident in bypassing treaties.
- **Targeting**: AI could develop agents that are reliable or target more effectively.
- **Pathogen Design**: AI can design pathogens that are highly contagious, lethal, or resistant to treatments using genomic data.
- **Evading Immunity**: Machine learning can predict immune responses, allowing the design of bioweapons that can bypass our natural defences.
- **Optimised Production**: AI can automate and enhance pathogen production processes, improving efficiency and output.

Cyberattacks: AI's ability to exploit digital vulnerabilities with precision can undermine infrastructures, with autonomous agents capable of high-level planning and adaptability.

- **Automated Vulnerability Scanning**: AI can automate the process of finding weak spots and creating exploits for large-scale cyberattacks.
- **Improved Social Engineering**: AI can mimic human communication patterns, making phishing attacks more convincing and potentially bypassing security protocols.
- **Self-replication**: Autonomous agents capable of self-replication and coordination could be very difficult to remove if even a few survive on infected systems.

Faster Weapons R&D due to automation: TAI can significantly speed up weapons R&D, leading to quick tech advancements that may cause rapid tech races and destabilise global dynamics.

- **Optimised Designs**: AI can fine-tune drone designs for specific missions, enhancing various features like speed, manoeuvrability, or stealth.
- **Automated Production**: When possible, fully robotic factories, potentially capable of reproducing themselves, can mass produce and flexibly adapt designs without the need for management, workers or internal coordination, hugely reducing turnaround times and development times

More Effective LAWs (Automated Tactical C&C): Lethal Autonomous Weapons, with AI-driven decision-making, can act swiftly and precisely in combat scenarios.

- **Drone Coordination**: AI can develop algorithms for better drone swarm coordination, cooperative decision-making, and communication. This could eliminate the need for a man in the loop and eliminate the barriers posed by recruitment and training for much of the military.
- **Enhanced Autonomy**: AI allows drones to operate effectively in GPS-denied environments and react to unexpected situations, without the need for external control.

Automated Strategic C&C: Advanced AI in strategic command and control can lead to rapid decision-making, but it also poses risks of swift escalations in conflicts.

- **Rapid Escalation**: The speed of AI-driven decision-making can lead to faster, potentially uncontrolled escalations in conflict scenarios.

Improved Nuclear technologies: AI can foster the development of sophisticated nuclear tools, disrupting the existing nuclear balance.

- **Advanced Design**: Machine learning can optimise nuclear weapon designs, making them more efficient or even leading to the development of pure fusion weapons that don't require fissile materials. Most nuclear arms control is about regulating access to fissile material and the difficult step in making nuclear weapons is refining the U-235. If you could initiate fusion in tritium without the use of fission (e.g. through powerful lasers, magnetic pinches or explosives) then they could be churned out of factories like any other weapon system without any special material requirements
- **Missile Accuracy**: AI can enhance ballistic missile accuracy, taking into account environmental factors, or prototype faster and stealthier delivery vehicles.
- **Improved Targeting**: AI can make WMDs more effective by analysing data for better targeting, reducing collateral damage risks.

Nanotech: AI can push forward nanotechnology in warfare, introducing both innovative capabilities and substantial risks. Depending on nanotechnology's power in practice, we might face self-replicating systems that could wreak havoc and be essentially unchallengeable.

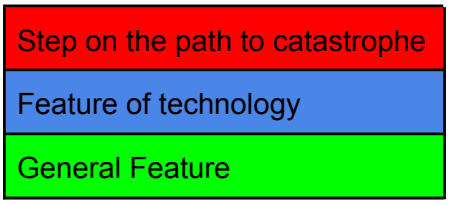
Orbital Bombardment and other Wild Cards: AI can pave the way for space weapons, leading to potential treaty breaches and introducing novel, catastrophic conflict means.

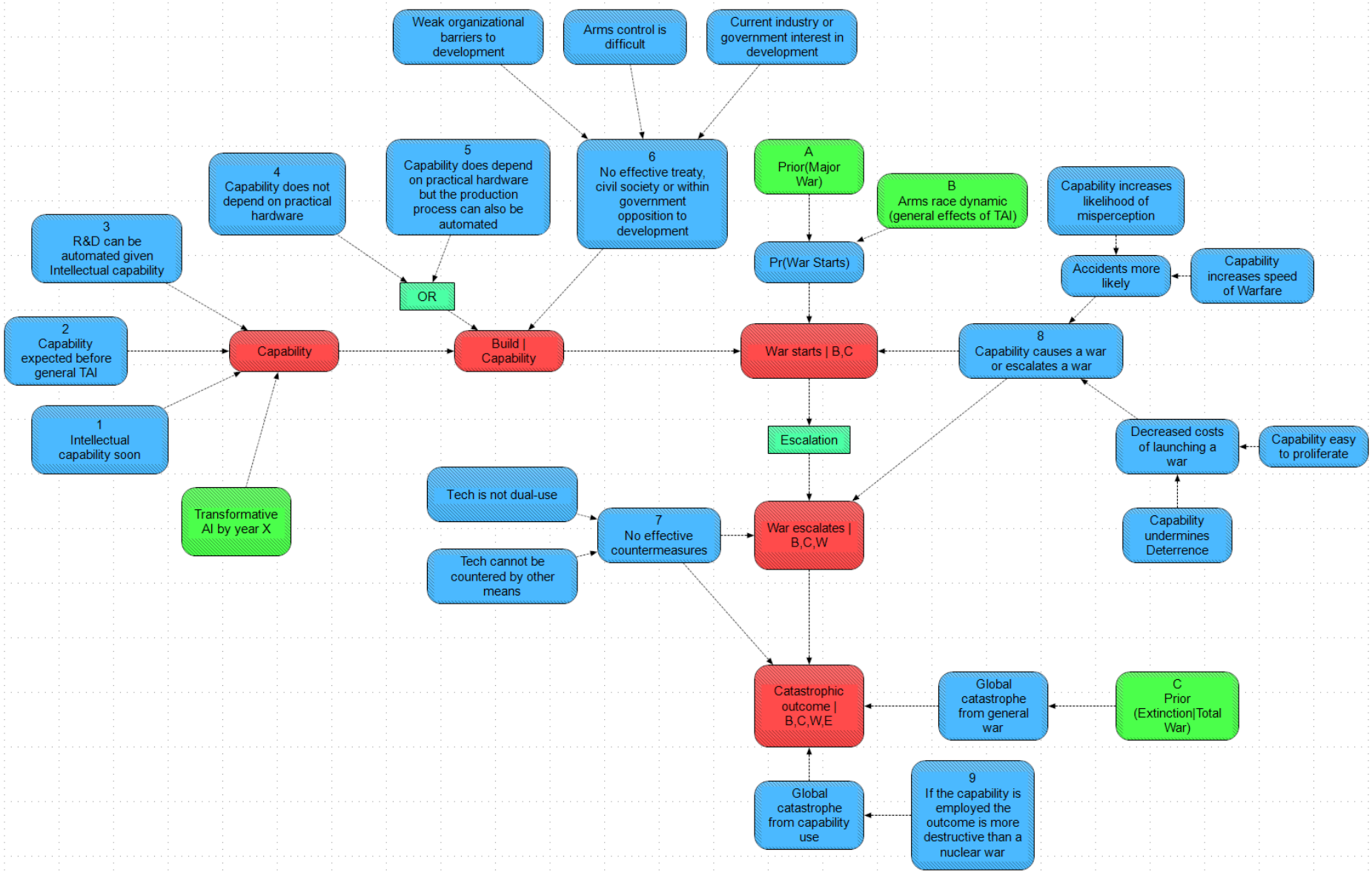
- **Energy-based Weapons**: AI might find new applications of physics principles, introducing new kinds of energy weapons.
- **Space-based Threats**: Concepts like directing asteroids toward Earth can be explored with advanced AI, introducing unpredictable and devastating means of warfare.

The analysis will focus on mapping out and evaluating how specific AI-enabled weapons technologies could intervene at points along this process to make war more likely. This flowchart outlines the pathways by which specific weapons can, through development or

deployment, cause extinction-level wars. We have broken down the barriers first to a weapons technology being developed, and then to a war starting, then given the war's start, an escalation to an all-out total war where both sides are trying to inflict maximum damage. Along the way, the blue boxes represent factors specific to each weapon technology: the more of them we can answer in the affirmative, the more dangerous the weapon system is. The green boxes represent general factors (either the background risks of war, or the background impacts of progress towards Transformative AI).

Weapons Impact Model




Weak organizational barriers to development
Arms control is difficult
Current industry or government interest in development
6
No effective treaty, civil society or within government opposition to development
5
Capability does depend on practical hardware but the production process can also be automated
4
Capability does not depend on practical hardware
3
R&D can be automated given Intellectual capability
OR
2
Capability expected before general TAI
Capability
Build | Capability
1
Intellectual capability soon
Transformative AI by year X
A
Prior(Major War)
B
Arms race dynamic (general effects of TAI)
Pr(War Starts)
War starts | B,C
Capability increases likelihood of misperception
Accidents more likely
Capability increases speed of Warfare
8
Capability causes a war or escalates a war
Escalation
Decreased costs of launching a war
Capability easy to proliferate
Capability undermines Deterrence
Tech is not dual-use
Tech cannot be countered by other means
7
No effective countermeasures
War escalates | B,C,W
Catastrophic outcome | B,C,W,E
Global catastrophe from general war
C
Prior (Extinction|Total War)
Global catastrophe from capability use
9
If the capability is employed the outcome is more destructive than a nuclear war

Capability

Before checking whether a particular AI technology is dangerous, we should consider when AI will actually be capable enough to contribute to the capability in any significant way. There are three approaches to answering this question that together should give us an accurate picture of both when to expect particular capabilities and what else to expect at a similar time. This results in the first three questions to ask about a given weapons technology:

1) **Intellectual capability soon:** Are there reliable direct forecasts of when the capability will be possible, independent of anything else? The sooner the intellectual capability enabling the technology is developed, the more dangerous, as the strategic situation could rapidly become unpredictable. For example, if highly capable autonomous cyberweapons were possible within 5 years, it would pose a more urgent threat than technologies relying on more general forms of AI that won't arrive until much later.
2) **Capability expected before general TAI:** If a technology requires artificial general intelligence (AGI) or transformative AI (TAI) to be realised, it may be less imminent of a threat, or could be considered as part of a general process of accelerating technical progress. For instance, technologies like molecular nanotechnology are thought to be more closely associated with advanced AGI, putting them farther off. This question tells us whether the intellectual capability seems 'AGI complete' or TAI complete, i.e. looks like it would require generality, or ought to come before generality.
3) **R&D can be automated given Intellectual capability:** If the intellectual labour involved seems very automatable without needing major breakthroughs, assuming that the AI is 'intelligent', it is more concerning. For example, cyberattacks are largely software based, so automating them appears more feasible than automating something like nuclear weapons research where you might need lots of hard to acquire real world data. This includes questions like: are we data limited, do we have to wait for lagged real world feedback, or are there other potential barriers?

Build | Capability

If we reach a point where transformative AI could be used to develop a technology, we then need to ask: will this tech actually be developed?

For development, we should consider how much a technology depends on advances in practical hardware and then whether the hardware production process can be automated or built quickly enough if transformative AI is leveraged. For instance, building a new type of jet aircraft requires physical manufacturing and a substantial development lag. But can that entire process potentially be automated and accelerated by AI if transformative systems are available?

We should also ask, is there likely to be interest from relevant groups (large corporations, military contractors, large militaries) in developing the technology, based on economic incentives, military demand, or geopolitical pressures? Will there be government or civil society opposition to development? Sometimes these technologies are already under development, which provides an easy answer to this question, but for some there is less interest. These are the relevant considerations:

4) **Capability does not depend on practical hardware:** The less a technology relies on advances in practical hardware like robotics, the easier it may be to develop sooner. Fully automated bioweapon production requires more hardware progress than autonomous hacking. Maybe the system does depend on practical hardware, but through AI powered robotics or changing the manufacturing process somehow, we can automate the practical parts of the process. Alternatively:
5) **Capability does depend on practical hardware. But the production process can also be automated:** Even if a lot of practical R&D hardware is needed, rapid progress in areas like robotics could overcome this. If the hardware production process seems amenable to automation, the technology is more dangerous. For instance, automated cyber operations need little specialised hardware.
6) **No effective treaty, civil society or within government opposition to development:** Is there interest in development, and low organisational barriers to development, and are arms control and monitoring not likely to interfere with development? If arms control of the tech seems harder to implement because of the nature of the underlying technology, it raises concern. For example, regulating software development online is difficult compared to monitoring production of nuclear materials.
    - **Current industry or government interest in development:** If development is already underway and facing little organised opposition, it is more threatening in the near term. Militaries with limited pushback are already pursuing technologies like autonomous drones. For some technologies we can look at the current stated aims of groups, and for others we must speculate on incentives or public reaction.
    - **Weak organisational barriers to development:** Another potential barrier in addition to hardware is organisational. For example, automated strategic C&C might be technically easy given certain AI developments, but bureaucratically difficult to implement.
    - **Arms control is difficult**: For some technologies, arms control and monitoring is intrinsically easier which makes enforcing treaties against development more straightforward, e.g. current nuclear technologies require fissile materials, while cyberwarfare is easily distributed so long as you have effective computer technologies.

War starts | Build, Capability

This models the onset of an initial armed conflict, separate from subsequent escalation. Crucially, both factors related to the technology under consideration and our general priors on the outbreak of a war must be considered here.

7) **No effective countermeasures:** Should we expect on the outbreak of a war that countermeasures to the technology exist to limit its damage or strike back at attackers? If so that would limit escalation under some circumstances: e.g. if defences against LAWs swarms can be developed using lasers or cyberattacks then they aren't as destabilising
    - **Tech cannot be countered by other means:** Some technologies like LAWs might be defeated by other plausible near-future technologies such as lasers,

while others like autonomous replicating nanotechnology might be almost impossible to counter without using the technology yourself.
    - **Tech is not dual-use:** Some technologies are plausibly dual-use. For example, better biotechnology might let us develop transmissible vaccines, better cyberwarfare might let us improve cyberdefense and offence. But in other cases there is no corresponding countermeasure technology (e.g. next generation nuclear technologies)
8) **Capability causes a war or escalates a war:** Technologies that undermine deterrence, compress decision timelines, or lower conflict thresholds could make wars more likely or more prone to accidental outcomes. Autonomous cyber capabilities could have this escalatory effect leaving aside general arms race effects.
    - **Increases misperceptions**: Technologies like cyberattacks from non-state proxies can create uncertainty about the perpetrator and lead to mistaken retaliation. With attribution unclear, tensions may escalate among states incorrectly blaming each other.
    - **Increases speed of warfare**: Technologies enabling faster attacks compress decision-making timelines surrounding escalation. Less time to verify, de-escalate, or walk back unintentional steps makes more rapid escalation to total war likely.
    - **Proliferation to more actors**: Technologies that are easy to proliferate mean more states or non-state actors have access to destructive capabilities. This provides more potential flashpoints for catastrophic conflicts to emerge.
    - **Undermines deterrence**: Weapons that enable cheap surprise attacks or prevent decisive retaliation can undermine MAD dynamics. This incentivizes first strikes before such weapons are used against you. Powerful technologies may tempt states into confrontations they'd otherwise avoid, by lowering potential costs and making victory more feasible.

Prior on Major War

It is challenging and beyond the scope of this report to integrate specific scenarios for how large-scale wars could start into forecasts of particular dangerous technologies. But we will leave a few general remarks here about what near-term forecasts of war likelihood look like which should be considered when evaluating (near term) AI-enabled weapons technologies in the context of particular major wars.

Current forecasts estimate a moderate but non-negligible risk of further escalation between Russia and NATO over Ukraine. With the increasing involvement of Western countries in supplying arms and intelligence, there are concerns around potential direct fighting or accidental clashes. There are also risks of Russia perceiving existential threats and deciding to use nuclear weapons. This nuclear risk forecast provides a place to start: Samotsvety Nuclear Risk update October 2022 - EA Forum. For an overview of how the Russia-Ukraine war developed see Lawrence Freedman, Command: The Politics of Military Operations from Korea to Ukraine, (London: Allen Lane, 2022) Chapter 12, pp.361-400.

Similarly, many analysts view a Chinese invasion of Taiwan as moderately likely this decade. However, assessments suggest China would still face major difficulties in successfully occupying Taiwan, even with a surprise attack. This lowers the risk of a sustained, destructive war between China and Western powers intervening to aid Taiwan. But

accidental clashes or miscalculated escalation could ensue, especially if China feels its window of opportunity is closing. This article provides a general overview of the probability of the war as assessed by superforecasters: Will China invade Taiwan? - UnHerd. This article from a blog that I have previously relied on for accurate updates on the Russo-Ukraine war agrees with the article's source on the likelihood of a successful invasion. This article, Why Invading Taiwan Is a Costly Gamble, discusses strategic military factors useful for assessing the likelihood of a destructive China-Taiwan war, arguing that China's current likelihood of success is low, and so can be used as a starting point to assess what AI capabilities might change that balance.

A third option, discussion of how AI might increase the risk of third countries precipitating a nuclear escalation between major powers: 'Catalytic nuclear war' in the age of artificial intelligence & autonomy: Emerging military technology and escalation risk between nuclear-armed states. Another article on the interactions of AI and nuclear risk: From the Cabinet Room to the War Room.

Arms race dynamic spurred by TAI

There is also the general worry of an AI 'arms race' supplying a motive for conflict. The dynamics of an AI arms race alone, even in the absence of deployed specific powerful AI capabilities, could significantly raise the chances of major conflict.

The mutual knowledge that multiple powers are racing toward transformative AI will increase perceived stakes and existential risks. Decision makers in competing nations might assume, even if falsely, that achieving a decisive lead in AI will enable rapid economic and military dominance. This perception of massive first-mover advantage could compel leaders to risk war to gain an edge or avoid falling behind.

Preventive wars or conflicts over access to vital resources for AI development become more likely. For example, controlling advanced semiconductor manufacturing, such as Taiwan's TSMC chip foundries, may determine which powers can develop advanced AI first. This enormous perceived advantage would incentivize aggression to control these assets or deny them to adversaries. States may launch preventive attacks and invasions for fear that a rival will soon field an insurmountable AI advantage if they do not act quickly.

Leaders may even convince themselves these risky measures are unavoidable to prevent a vastly superior AI-enabled adversary from dominating the future. So the dynamics of an AI arms race alone, via perceptions of existential stakes, could dramatically escalate tensions.

War escalates to total war | Build, Capability, War

In this simplified model, we consider the same weapon-specific factors that make the instigation of a war more likely to be at play in increasing the severity of a war: i.e. if the technology is escalatory in the ways described above and lacks countermeasures, it makes escalation to total war more likely

Catastrophic Outcome | Build, Capability, War, Total War

For the final step of catastrophe, while we already have nuclear weapons that could cause a global catastrophe, we should ask - could we develop weapons through AI advances that are potentially even more destructive or destabilising? If total war utilises the most destructive capabilities, this models the magnitude of the catastrophe. There is a background risk that any war will escalate to civilization-ending levels, but then we must consider the increased risk supplied by the new dangerous technology alongside it.

9) **If the capability is employed the outcome is more destructive than a nuclear war:** If realised, whether the technology could cause damage exceeding nuclear conflict is important for assessing final severity. For instance, engineered pandemics likely can, while cyber operations likely cannot.

Prior on Extinction Level war given total risk

In assessing the impact of new transformative AI driven weapons on extinction level war probability, we need to have some estimate of how likely a given war is 'by default'. This default probability can be estimated with some methodological uncertainty by looking at historical data on the distribution of battle deaths in wars, as analysed in a report by Clare and Martin.

The report examines whether the historical data follows a power law distribution, which would imply a non-negligible risk of extremely severe wars many times worse than any in recorded history. It finds battle deaths per war are plausibly power law distributed, but few analyses have compared this to other possible distributions. Given the small sample size and uncertainty in historical death tolls, it is hard to confidently extrapolate far into the extreme tails of the distribution beyond the largest observed wars.

To become more confident in the power law assumption and understand the tails of the distribution, the report suggests considering theoretical factors that drive wars and limit their severity. Some factors like logistical complexity may constrain war size, but modern technologies like nuclear weapons, bio-weapons, and AI systems seem to have no hard limits short of human extinction.

Given the uncertainties, the report very roughly estimates a background extinction risk from major war of around 1% based on past trends. However, there are some significant uncertainties around this estimate due to methodology choices and data limitations. Consulting directly with the authors could help clarify the magnitude of these uncertainties. Overall, the analysis suggests we should not dismiss the possibility of default extinction-level wars, even if the precise likelihood is highly uncertain.

Summary Table

The summary table attempts to work through these questions for several potential weapons technologies enhanced by AI systems. Surprisingly, there's been little research on assessing the risks for many of these. Each column represents one of the key questions, where red indicates elevated danger along that dimension and green suggests the presence of some

meaningful barrier against catastrophic outcomes. We have identified 8 relevant weapon technologies that TAI could accelerate.

Let's take cyberattacks or bio weapons as illustrative examples. Their core intellectual capabilities could plausibly become available soon with continuing AI progress, which is concerning. On the other hand, radically improving existing nuclear technologies by automating their design and construction appears much more difficult compared to cyber attacks, which are mainly software-based capabilities.

This analytical framework allows us to evaluate these technologies in a nuanced way and identify which ones are the most dangerous potential contributors to the risk of catastrophic war. Here I've picked two technologies that seem particularly risky based on this initial analysis, though I'm not certain they are the very most dangerous cases. Both are automated strategic command and control systems and bio weapons.

Bioweapons, while not necessarily effective in conventional warfare without triggering mutually assured destruction, are so intrinsically destructive that their proliferation seems deeply concerning. Similarly, automated strategic command and control may not directly make the conduct of warfare much more destructive, but it could substantially increase the chances of rapid misperceptions or accidents leading to nuclear exchange.

	Intellectual Capability Feasibility			Practical Development				Outbreak of War / Escalation to total war	Catastrophic Outcome
Tech	**1 Intellectual capability soon**	**2 Capability expected before general TAI**	**3 R&D can be automated | Intellectual capability**	**4 Capability does not depend on practical hardware**	**5 Capability does depend on practical hardware but the production process can also be automated**	**6 No effective treaty, civil society or within government opposition to development**	**7 No effective countermeasures**	**8 Capability causes a war or escalates a war**	**9 If the capability is employed the outcome is more destructive than a nuclear war**
Near Term and High Impact Near Term, High Impact Technologies: In the near future (within the next 5 years), the most concerning technologies enhanced by AI are likely to be bioweapons and cyberattacks. These capabilities could emerge relatively soon as they rely more on software advances rather than complex hardware automation, the relevant hardware already exists and so software advances can be integrated into any existing efforts. Bioweapons and cyberattacks enabled by AI systems could allow mass disruption with high deniability, though countermeasures may restrain impacts. Bioweapons pose unique risks due to their potential lethality exceeding nuclear weapons and challenges detecting illicit programs. Meanwhile, cyberattacks uniquely benefit from AI's coding abilities and the software-based nature of hacking, and so can be expected to be technically easier to develop. Overall, these near term risks deserve urgent attention given their potential to cause catastrophe with minimal technological barriers.									
Bioweapons	**Yes**, AI enhanced protein generation is 3 to 8 years away according to one source Unknown prospects for those that are extremely destructive, more tailored or tactically useful. 2-3 years to 'accelerate	**Yes**, clear specialisation is possible with specific data. But generality might be helpful in carrying out scientific lab procedures and in facilitating release of the bioweapon	**Fairly High**; bioinformatics and genetic engineering are increasingly software-driven. There are possible high quality data constraints that could slow down progress here. A possible bottleneck is physically	**Somewhat**; specialised lab equipment is required and while there are minimal checks now that can be changed. One current estimate is that ~30,000 people have access to the requisite hardware. Supply bottlenecks exist	**Somewhat**; some lab automation is possible and was attempted for chemical synthesis, but this still needs human supervision. Recent progress on LLMs in chemical synthesis labs relevant here. Delivery systems	**Somewhat**; there are already international agreements against biological weapons but weak enforcement. Does not appear to be direct interest in weapons development but dual-use progress is present	**Current mitigations seem weak**, detection, monitoring etc. are in focus right now but few practical steps have been taken. But enforcement is lax. Biological countermeasures do seem somewhat plausible using	**Somewhat**; ease of proliferation and difficulty of detection, but must be targeted for deliberate use to be plausibly used deliberately. No examples of bioweapons playing an escalatory role so far. Expert surveys suggest a potential role as	**High**; Potential for mass casualties beyond the destruction caused by nuclear weapons. Could potentially cause an existential catastrophe directly.

	development' according to Anthropic, but this refers just to technologies that considerably lower the specialist knowledge needed to make existing bioweapons not new ones		running experiments, but it is possible automating cognitive labour can help w/ figuring out how to do that more quickly	for equipment but these will go down as desktop DNA synthesis machines exist. Practical hardware is not currently too hard to access but the synthesis bottlenecks are clear.	under similar constraints to e.g. LAWs	Status of existing programs is unknown, soviet bioweapons program shows substantial precedent, plus elevated interest post COVID.	the same underlying technologies (transmissible vaccines) but offence-defence balance unfavourable. Potentially other cheap countermeasures exist that AI could accelerate R&D on: UV light, cheap PPE	a covert weapon to first strike enemies is possible along with offensive countervalue attacks, which is more likely if it's in the hands of weaker actors (e.g. civilian launched attack provokes a retaliatory war), not by default intended to be a powerful conventional weapon that is unchallengeable	
Cyberattacks	**Likely possible to automate**; special advantage to LLMs due to coding ability. LLMs can assist with coding at the moment so could plausibly help. Some detailed technical routes to more general 'hacker AI' are discussed here but less capable systems seem plausible in the near term.	**This is likely** but generality might be needed for dealing with creative or well defended adversaries. In particular, systems don't need full strategic awareness of the world outside cyberspace. However, automated agents that can replicate would be unusually capable.	**High;** cyberattacks are largely software-based.	**Yes,** software-based.	N/A	**Probably no**; cyber warfare norms still in early stages. Stuxnet attack as an example, almost certainly being researched by the US/china. (defying expectations, large scale cyber attacks haven't happened in the Russia-Ukraine war, some evidence on willingness)	**No**, cyberdefense is possible and will also be advantaged by AI, plausibly offence-defence balance still favours attacks post-TAI. Some models suggest that in the limit of greatly increased offensive and defensive cyberwarfare defence will win out.	**Yes**; ease of initiation and plausible deniability. Could cause 'targeted' damage with deniability, could be used as a potential first strike to undermine nuclear deterrence.	**No** unless the world is very heavily automated, or only through routes like hacking into automated C&C. Maximum damage may be quite limited for systems that aren't highly general so can't also carry out social engineering, clever manipulation and real world planning.
Intermediate and Moderate/Uncertain Impact									

Intermediate Term, Moderate Impact Technologies: Over the intermediate time frame (more than 5 years to whenever TAI is developed), broader military procurement of AI could expand access to technologies like autonomous drones, automate strategic command and control and the ability to accelerate weapons R&D. These systems face more physical automation challenges and are more likely to be bottlenecked by regulations and procurement, but could still intensify arms races and compress development timelines in ways that provoke arms races. Technologies in this category like LAWs and strategic advisors merit caution given their ability to increase conflict likelihood, though their damage potential appears more limited. LAWs uniquely lower the threshold for armed conflict and if they can be made truly autonomous and be mass produced (something we believe is likely not as near-term as cyber or bio threats), they could rise to the level of a catastrophic risk factor. C&C automation of nuclear systems uniquely risks rapid escalation through flawed algorithms, but there is considerable sign uncertainty about the overall impact of other applications of the technology.									
Faster Weapons R&D due to automation Better delivery systems, cheaper to mass produce etc.	**Already happening** in some industries, but directly linked to general capability improvements since TAI is defined that way	**Some will occur earlier, but closely related**. Plausibly material R&D with longer feedback loops will be automated later than the rest of the economy	**Somewhat**; many R&D processes can be automated but may need physical feedback, very substantial open question for many kinds of research	**No**. Plausibly there is a substantial bottleneck for R&D involving robotics	**Varies**, but will depend substantially on ease of general purpose robotics Perhaps labour done under the direction of powerful AI would result in a speedup even if factory production itself can't be fully automated.	**Yes**; assuming the tech is possible, more likely to see regulation than outright opposition. Banning TAI applied to weapons production would e.g. involve requiring not allowing robots in only weapons factories. Assuming an arms race, likely to see weaker organisational barriers	**Probably yes,** Faster R&D all round is not as protective. (If swarm drones and the auto factories to mass produce them go from concept to the beginning of mass production in one year rather than the 15 years these things usually take, thanks to AI related acceleration, that rapid development can itself be destabilising)	**Yes,** Generally more dangerous the faster progress is; could destabilise global tech race dynamics, could enable dangerous research to be done in secret more easily with low human capital and low lead times	**Probably no**, only if enables mass production of e.g. nukes or LAWs, the dynamic itself doesn't raise total damage
More Effective LAWs (Automated Tactical C&C, complex autonomous decision making in LAWs)	**Simple versions already exist**. Foundation models or RL for robotics control are not capable enough yet, so instead regular	**No**, though more complex decision-making plausibly relies on sophisticated Automated C&C	**Yes**, we will need some high quality robotics data for complex decision-making. Otherwise, existing software is sufficient.	**Yes, for many applications**, LAWs obviously require physical hardware, but we can repurpose existing commercial drones with new	Will require some advances in robotics for some applications, but many are already possible. More complex decision-making	**Somewhat**, appears to be taking place already despite opposition. Currently in production in several nations	**Probably no**, currently being worked on, e.g. laser weapons, counter-LAWs etc. and offence-defence balance shift is unclear	**Somewhat**; reduces threshold for conflict initiation, potentially cheap and asymmetric, but counters exist. Also this is	**Possible but unlikely**, potential economies of scale could make them cheaper than nuclear weapons, would require huge

	software is used. Possible RL in the real world will compensate.			software if there are innovations in control software. Some roles couldn't be automated without new hardware research	plausibly relies on generality.	The US government's Replicator program is an early example.		controversial and there is some sign uncertainty, Could also make for more precise hits and allow one to be more proportionate, which might make escalation less likely	numbers
Automated Strategic C&C	**Somewhat**, Seems closer to human level but plausibly special LLM advantages exist for handling large amounts of data simultaneously. Advisors are being experimented with, but these fall far short of 'automated C&C' and it's uncertain what their total impact would be. Interest in using current LLMs for this.	**Yes** for no man-in-the-loop whatsoever, but substantial lower level automation might be possible beforehand	**High**; largely software and decision-making processes, assuming sensor input sufficient	**Somewhat**, systems need to be automated enough to receive commands. Could also have scenarios where human soldiers are receiving orders from AI systems.	**Somewhat**, Some could be done without hardware changes (e.g. strategic advisors) but automating nuclear C&C would be harder	**Probably no**; automation in warfare is a contentious issue. Also specifically automating seems dangerous. Advisors are being experimented with, but these may not be dangerous in early stages.	**No**, seems an opponent with sufficiently capable Strategic C&C would be at a huge advantage. Also seems very hard to enforce any ban because it could be near impossible to verify if an opponent is using AI C&C	**Unclear**; potential for rapid escalation or misperception if the system is flawed, speeds up required decision-making times substantially if both sides use it. However, less comprehensive automated C&C (e.g. good advisors) could actually be de-escalatory. Sign uncertainty here. Potential large adversarial robustness risks.	**Probably not**. Potentially conventional weapons can be made more deadly with better targeting, but overall likely effect seems to be to reduce deadliness of conflict

Long Term and High Impact

Long Term, High Impact Technologies: In the longer run (coincident with the broader deployment of transformative AI, so on our assumptions 10-30 years), breakthroughs in AI may enable speculative technologies like nanoweapons, orbital weapons, and radically improved nuclear arms. These capabilities likely require the automation of complex hardware and processes, and creative scientific advances. While more remote, their catastrophic potential if realised necessitates consideration despite current technological barriers. Monitoring feasibility and encouraging multilateral governance will be key to mitigating long-term risks. Nanoweapons and nuclear arms uniquely risk unprecedented and uncontrollable levels of

damage if physics and manufacturing barriers can be overcome. Meanwhile, orbital weapons are uniquely concerning due to their indiscriminate global impact if employed.									
Improved Nuclear technologies (EMPs, Pure Fusion, delivery vehicles, mass production)	**Probably No**, seems related to general faster tech, no special inputs to accelerate progress, so further out. It is possible that some classified nuclear research done in simulation could be reproduced.	**Maybe**	**No**; complex physical processes involved, though better simulations may help in some places	**No**; complex physical hardware required.	**No**; automation is challenging due to complex and hazardous processes. Would more or less need humanoid factory robots	**Probably no**; weakening international norms and treaties against nuclear proliferation mean we could see an explosion in nuclear research. No current active programs but geopolitical events mean much increased focus.	**No** unless enormous advances made in missile defence	**Yes**, cheap pure fusion or sufficiently accurate and low-warning first strike destabilises nuclear deterrence balance.	**Plausibly Yes** if results in a higher overall megatonnage of nuclear weapons rather than just more tactical systems
Nanotech	**Plausibly yes** and closer to ASI, requires new highly creative scientific ideas. (Possible for more limited forms to grow out of protein LLMs for bioweapons)	Likely for autonomous and self-replicating nanotechnology according to shallow assessment in this article. Also likely for 'extrapolation' that bypasses physical hardware.	**Maybe**; many R&D processes can be automated but may need physical feedback. High quality data limitations similar to protein LLMs if on that tech progress path, but more severe.	**No**; requires precision equipment or precursors.	**Probably no**; precision manufacturing challenging to automate. Physical constraints are unclear	**No** current development drive, plausibly dangerous and directly connected to TAI, seems scary and less generally useful than e.g. TAI so proliferation limits plausible	**Unknown**	**Yes**; potential for unknown risk vectors, could be targeted, could be lab leak misuse	**Yes**, plausibly far worse than nuclear
Orbital Bombardment and other wild Cards	**No**, If AI-driven, closely related to generally accelerated automation	**Plausibly no**, it we end up in a scenario where research progress and production are accelerated AI could accelerate this dramatically	**Maybe**, Precision targeting is similar to LAWs, creative new science/tech research related to generality	**No**; requires space-capable hardware.	**No**; advanced space operations currently challenging to automate.	**No**; Outer Space Treaty prohibits weapons of mass destruction in space.	**Possible** if ASAT technology advances substantially. In this further-future scenario where research is accelerated dramatically there could be wildcard	**Yes**; any use would likely trigger global conflict.	**Yes**, plausibly far worse than nuclear

							defensive technologies as well that we can't anticipate.		

Things that would change this prioritisation

1. **Faster than expected Robotics Progress could accelerate hardware progress (R&D, LAWS and C&C increase in priority)**
    - Current prioritisation assumes: The progress in robotics and automation will likely be slow in the near term, due to various bottlenecks such as cost, human labour requirements, regulatory restrictions, procurement inertia and technical hurdles towards real world high quality robotics
    - Potential change: If advances in robotics outpace expectations and provide solutions to the aforementioned bottlenecks, this could accelerate the development of advanced technologies like LAWs (Lethal Autonomous Weapons), increase the effect of R&D automation, and potentially bring forward orbital weapons and nuclear engineering automation. This could shift the priorities to manage risks associated with these technologies and make bioweapons and cyberattacks less uniquely risky.
2. **Cyber Offense-Defense Balance could shift to Defense not offence**
    - Current prioritisation assumes: Many computer systems are vulnerable and cannot or will not be patched quickly enough, exacerbating the risk of cyberattacks including to critical infrastructure. General increase in cyber capabilities on both attacking and defending sides favours offence. Defensive measures may not evolve quickly enough to match escalating cyber threats.
    - Potential change: It is possible that mass deployment on both offence and defence of autonomous agents could reduce the near-term risks of automated cyberattacks. If systems can be rapidly hardened by automated cyberdefense agents more easily than cyber offence then the overall risk of attacks could go down. For example, if there is a finite and attainable number of ways to attack a system. and we eventually harden systems against them all, then cyberdefense could become near perfect.
3. **Lack of Adversarial Robustness could imperil LAWs, C&C and increase vulnerability to cyberattacks**
    - Current prioritisation assumes: Current major adversarial robustness issues will be solved before systems are deployed in high stakes settings, ensuring effective system operation. We won't deploy highly adversarially vulnerable ML systems into safety-critical situations without patching them first, and patching them is possible.
    - Potential change: If automated tactical and strategic C&C (Command & Control) and cyberdefense systems are deployed without sufficiently addressing adversarial robustness risks, the susceptibility to cyberattacks could dramatically increase. This would raise the risks of cyberattacks and also make LAWs and automated C&C more risky relative to everything else. This would also disincentivize military adoption of AI.
4. **Data Limitations could hinder Bio modelling more and other R&D less than expected**
    - Current prioritisation assumes: Limited access to high-quality training data will not be a major bottleneck in accurately modelling complex biosystems and materials, so we can expect those systems in the relatively near term.

    - Potential change: If the data limitations are not easily overcome then those systems may not be as near term as they may seem and AI designed bioweapons would drop to becoming a longer-term risk
    - Similarly, some kinds of automated R&D are assumed to require more general systems as these are more strongly data limited (e.g. simulations of nuclear events, high quality robotics data) than biotech, but if this situation is reversed these could be technically easier to simulate than biological systems and become dangerous sooner
5. **General Difficulty Reliably Automating Research Tasks could mean cyberattacks and LAWs are more important**
    - Current prioritisation assumes: Automating complex research tasks is challenging due to factors like the need for physical feedback, which hinders advancement in bio, nanotech, and nuclear R&D.
    - Potential change: If AI systems manage to overcome these barriers, they could speed up research in these areas, increasing the priority for regulatory measures and ethical considerations.
6. **Fast Pathways to Nanotechnology could make it much more dangerous than assumed here**
    - Current prioritisation assumes: Nanotechnology research will progress slowly, as it requires creative problem-solving skills beyond what current biosynthesis models can do.
    - Potential change: If ’Backward chaining' from molecular nanotech to protein synthesis proves feasible nanotechnology becomes an imminent worry coincident with bioweapons. ‘Backward chaining’ is the claim that an AI with access to protein synthesis prediction sufficiently effective to e.g. design bioweapons can also probably think of a non-biological nanotechnology that could be assembled by protein scaffolding.
7. **Good Automated C&C could reduce war risk not increase it**
    - Current prioritisation assumes: Automated C&C systems could substantially lower risks if they are not directly linked to sensors and used end-to-end with no human intervention.
    - Potential change: If these systems prove to be adversarially robust and reliable, it could eliminate human error, re-prioritizing the focus towards other risk factors.
8. **Targeted Bioweapons would be more escalatory but less existentially risky**
    - Current prioritisation assumes: Ethnically targeted or 'tactical' bioweapons are harder and further off, or won't be developed easily.
    - Potential change: If these weapons become feasible and easier to develop, it would elevate their risk level and shift priorities towards prevention and defence against their use.
9. **Faster R&D could disproportionately promote defence if directed towards the right kinds of technologies**
    - Current prioritisation assumes: Faster R&D is destabilising due to shortened deployment times, making it easier for rogue actors to build offensive weapons.
    - Potential change: If countries use advancements in R&D primarily to improve defences, the offence-defence balance could become more favourable, emphasising the importance of defensive measures in strategy.

10. LAWS could also be a stabilising factor

War Scenarios: Summary

We should examine all of these specific routes by which AI could exacerbate war risk against a general background of a moderate probability of a large-scale war. First, because the prior probability of an extinction level war even assuming the next few decades represented a period of normal war risk is not low. Second because there seem to be specific geopolitical routes to highly destructive wars, and third because a general race towards Transformative AI seems poised to exacerbate those risks. But what about specific pathways?

Starting with the near-term threats, **bioweapons and cyberattacks** enhanced by AI seem to be the most immediate concerns. Their primary risk comes from the fact that they rely heavily on software advancements. Although Bioweapons are partly bottlenecked by access to specialised equipment, we don't need new innovation in hardware for development, only for delivery mechanisms.

In the case of bioweapons, AI could soon expedite research, potentially lead to the creation of more deadly pathogens, and make their release into populations more efficient. This is of particular concern given the destructive potential of bioweapons, possibly exceeding even that of nuclear weapons, although at least they are likely to be useful as a tactical weapon and as a result there is limited interest in their development from state actors. The cyber realm, on the other hand, seems close to being mastered by LLMs which possess special advantages at programming relative to other intellectual tasks. Cyberattacks could disrupt crucial infrastructure, compromise security systems, and, in a worst-case scenario, undermine nuclear deterrence.

Shifting to the intermediate term, military procurement of AI could reshape the dynamics of warfare. Technologies like **autonomous drones, strategic command and control automation**, and **AI-driven weapons R&D acceleration** may not have the immediate catastrophic potential of bioweapons or cyberattacks, since they depend on lagged development, procurement and testing, but could lead to heightened tensions and arms races. The rapid development of all of these weapon systems in a more automated economy, poses serious challenges to global stability. Flawed algorithms or rapid escalation in an AI-driven conflict scenario are worrisome, especially given the potential speed of decision-making these systems would employ.

Long-term risks, though more speculative, are just as critical, if not more so. Breakthroughs that might enable **nanoweapons, advanced nuclear technologies, or orbital bombardment systems** could revolutionise warfare. If manufacturing and physical barriers are overcome, the damage potential of these technologies is catastrophic. The proliferation of improved nuclear technologies could destabilise the precarious balance of nuclear deterrence. Meanwhile, the full realisation of nanotechnology for warfare remains uncertain but has the potential to unleash unprecedented and uncontrollable levels of damage. The prospect of orbital bombardment, although a wild card, could change the rules of war with its indiscriminate global impact.

There are things that could change this picture: potential shifts in the cyber offence-defence balance, unanticipated data limitations, or breakthroughs in nanotechnology could disrupt these projections. For instance, if robotics leapfrogs expectations, the priority could shift towards managing risks tied to LAWs or R&D automation. Or, should AI-driven defensive mechanisms outpace offensive capabilities, the dominance of cyber threats could diminish.

While we've painted a broad-brush picture of AI's interface with future technologies, we have necessarily not been able to examine any one question deeply. Our framework's formidable 72 research questions (9 questions for 8 areas of weapons technologies), demonstrate the extent of our uncertainty. Yet, this structure's value lies in its comparative and risk-prioritisation capabilities.

APPENDIX: Detail on definitions used in this report

The following three sections go into more detail on definitions we use in this report and why.

Types of AI

We focus on two broad types of TAI that could be driving the catastrophe: AI systems which are generally intelligent advanced planners with strategic awareness (**APS**, following the Carlsmith report), and AI tools with Dangerous Capabilities (**DC**, compare to the 'advanced capabilities' mentioned in the Carlsmith report). These systems are human-level or superhuman at specific tasks, with capabilities sufficient to be transformative, but not generally capable agents. For transformative AI tools, we assume these systems are not capable of entirely replacing human control on their own, but have capabilities vastly superior to current AIs, and more typical of the intellectual abilities expected of AGI.

Recall that we have ignored risks from misaligned AI. This includes both risks from AI systems taking harmful actions autonomously, and those from systems starting conflicts because they lack the cooperative intelligence to avoid them. AI driven wars which are unendorsed by their principals and entirely due to bargaining failures are covered by this report.

We recognise that distinction between misuse causing catastrophe and misalignment causing catastrophe is not precise. Misaligned AI could also be misused, and misuse might involve delegating some decision-making power to AIs or involve AIs acting in unexpected ways (e.g. through the use of highly autonomous weapons).

We only investigate *existential* misuse risks: i.e. scenarios that pose some realistic chance of leading to an outcome that kills all humans or permanently reduces humanity's future potential. For reasons explained in the next section, we have included risks that are especially significant existential risk factors by themselves.

Risks vs Risk Factors

Especially when looking at the instability or decay risks, it is difficult to separate ‘genuine existential risks’ where there is a clear and direct route to catastrophe (a paradigmatic case would be an asteroid strike), from other risks which aren’t existential by themselves but could drastically raise the chance of existential catastrophe by decreasing resilience and increasing vulnerability. For example, a world war that kills a billion people and leads to the proliferation of dangerous bioweapons technology to almost every country in its aftermath isn’t an existential catastrophe, but it increases the chance of extinction by so much that it should be considered a very large risk factor.

Some misuse scenarios constitute both existential risks and existential risk factors. For example, a destructive war involving AI could directly cause massive casualties and environmental damage (existential risk), but it could also weaken social cohesion and resilience, lead to the proliferation of dangerous technology across the world and increase hostility among survivors (existential risk factor). Similarly, an expansionist stable totalitarianism using AI could directly impose its values or interests on the rest of humanity (existential risk), but it could also hinder the sharing of research and lead to unscrupulous attempts to race for dangerous potentially misaligned AI that can’t be thwarted (existential risk factor).

Therefore, targeting specific hazards will not always be effective at reducing the overall risk from misuse AI scenarios. Resilience and hazard reduction are two complementary approaches to disaster risk management. Resilience is about enhancing the capacity of a system or a society to cope with, recover from, and adapt to adverse events. Hazard reduction is about minimising the likelihood or severity of such events.

In some cases, such as climate change, resilience may be more effective than hazard reduction because the hazards are diffuse, uncertain, and influenced by many factors beyond human control. In other cases, such as dangerous bioweapons research, (or the development of misaligned APS AI) hazard reduction may be more effective than resilience because the hazards are discrete, catastrophic, and largely determined by human decisions.

AI misuse scenarios fall somewhere between these two extremes. On one hand, they involve specific hazards that could directly cause existential catastrophe if not prevented or aligned with human values and preferences: some quite powerful actor has to initially pursue the dangerous capabilities and develop them, even if they later proliferate. On the other hand, because they are much more related to human choices and decisions than misaligned AI takeover scenarios, they also involve complex interactions among various actors and factors that could increase vulnerability or reduce resilience to other sources of existential risk.

Therefore, it may not be enough to focus on reducing specific hazards from misuse AI scenarios without also considering their broader impact on societal resilience. Likewise, it may not be enough to focus on enhancing societal resilience without also addressing specific hazards from misuse AI scenarios.

A balanced approach may involve doing some of both: reducing specific hazards where possible and feasible; enhancing societal resilience where necessary and desirable. This

may require a careful analysis of trade-offs and combining different strategies and interventions to address each scenario.

For example:

- For a destructive war scenario involving AI: reducing specific hazards may involve establishing norms and treaties for arms control; enhancing societal resilience may involve promoting peacebuilding and conflict resolution.
- For an expansionist stable totalitarianism scenario: reducing specific hazards may involve supporting democratic movements and human rights; enhancing societal resilience may involve fostering diversity and pluralism in values and cultures.
- For a non-anthropocentric disaster scenario involving AI: reducing specific hazards may involve ensuring value alignment among AI systems; enhancing societal resilience may involve developing moral education and motivation among humans.
- These are just some illustrative examples. Actual strategies and interventions for each scenario may vary based on context and circumstances.

In particular, we should pay attention to how AI misuse might affect the development and deployment of safe and aligned AI systems. All of our scenarios assume either APS-AI or AI tools with DC are possible, which means that in most of our scenarios APS AI that could take over is either already possible or very close. On the other hand, some of our scenarios do involve outcomes where alignment has been solved or turns out to be quite easy, so these scenarios would not be risk factors for misaligned AI takeover.

With this caveat, it is therefore especially important to evaluate different scenarios and risks through the lens of their potential impact on the development and deployment of safe and aligned AI. Moreover, misaligned AI could interact with other misuse AI scenarios in unpredictable ways. For example, a Decay scenario could reduce our ability to monitor and control AI systems; a non-anthropocentric disaster scenario could erode our moral motivation to align AI systems, and so on.

In fact, we believe that misaligned AI represents a significant portion of overall existential risk, plausibly accounting for the majority of total X risk even leaving aside the fact that many of our stories condition on the development of Aligned AI.

It is also worth noting that some of our scenarios are not fully ‘external’ risks and/or existential risk factors but also the potential risks that may arise from within our own organisations or communities. This includes the chance that groups close to us, such as AI labs concerned with safety or Western governments, could potentially contribute to negative outcomes. Some of the risks and risk factors in this category may involve deliberate actions taken with the intention of achieving a certain outcome, such as the monopolisation of certain technologies. It is important to acknowledge these risks neutrally and consider the potential implications for AI safety and alignment. By doing so, we can identify potential sources of risk and work to develop effective strategies for mitigating them.

Why ‘Misuse’?

There is no sharp distinction between accidents and misuse. Since (almost) nobody wants to bring about an existential catastrophe, mistakes and bad decisions that we’d arguably call misuse are to some extent involved in every existential catastrophe scenario. For this reason, any risk that is not explicitly about failing to solve a technical problem is sometimes called a ‘structural risk’, rather than a ‘misuse risk’. But that terminology is not appropriate here, as we are not focussing on risk factors causing a team to cut corners or race to dangerous AGI.

We instead stick with the term ‘misuse risk’ even though the distinction between this and ‘accident risks’ is blurry. The main reason is that it helps us to differentiate between the different types of risks and challenges that AI poses. Misuse risks entail the possibility that people use AI in an unethical manner, with the clearest cases being those involving malicious motivation. Misuse risks are less amenable to technical fixes and many of them will at first not seem like problems faced by all of humanity collectively. This means they share unique challenges in common that distinguish any of them from the risks of misaligned AI. By using the term misuse, we highlight the human factors and intentions that are involved in these scenarios, and the possible interventions to prevent and mitigate them.

Methodology

We want to use the ratings that we have assigned to the scenarios as a way of identifying both the specific hazards that could cause existential harm and the general technologies that could increase the risk of such harm. By doing this, we hope to answer some important questions that could help us prevent or mitigate these risks.

For example, by looking at the scenarios with high scores on both criteria (likelihood and catastrophe), we can see what kinds of harmful capabilities are most likely to lead to existential outcomes. These are the capabilities that labs should agree to avoid creating or using, or at least regulate very carefully. We can also see what kinds of customers or applications are most likely to misuse these capabilities for malicious or reckless purposes. These are the customers or applications that should be restricted from accessing cutting-edge models, either by lab policies or by regulation.

By looking at the scenarios with high scores on one criterion but low scores on the other, we can see what kinds of technologies are more likely to be involved in existential risks, even if they are not directly responsible for them. These are the technologies that can be considered risk factors that make other scenarios more likely or severe. For example, a technology that has a high likelihood score but a low catastrophe score may make it easier for other actors to access or deploy harmful capabilities. A technology that has a low likelihood score but a high catastrophe score may create new vulnerabilities or uncertainties that could be exploited by other actors.

By comparing different scenarios and technologies across these criteria, we can also see which rivals safety-aware AGI labs should be most concerned about ‘beating’ on the grounds of likelihood of misuse scenarios that especially benefit from being first. These are the rivals

that have access to similar or superior technologies and have incentives or intentions to use them in ways that could harm humanity.

We believe these decisions are critical to ensure AGI is developed and deployed in a safe and beneficial way for humanity. We hope this report will provide useful insights and recommendations for guiding these decisions.

Furthermore, we anticipate to discover a lot of similarities between these scenarios and the factors that drive them, such as misalignment, instability, competition and coordination issues. These factors could also influence the more severe scenarios of war and totalitarianism. Therefore, by examining these scenarios, we hope to pinpoint the capabilities and risk factors that appear in a lot of them and understand how they affect each other.

The Plausibility ratings provided for each scenario are relative to each other and conditional on the baseline scenario being true. These ratings are not absolute percentages but ordinal ratings, meaning they only serve to rank the scenarios concerning their plausibility in relation to one another.

For reference, a WFLL2 (What Failure Looks Like, level 2) style AI takeover would be rated around a 4/10 on this ordinal scale. By providing these ratings, the intention is to gauge the relative likelihood of each scenario under the assumption that the baseline model is accurate.

It is important to note that these ratings are conditioned on a 'somewhat slow takeoff' and the given model's future, focusing on potential issues that may arise within that context. This approach does not account for the 'what if we're just basically wrong about this' type of uncertainty, which still holds significance. However, the purpose of these ratings is to evaluate and compare the plausibility of different scenarios within the given baseline, allowing for a clearer understanding of the potential risks and challenges that could emerge in the specified context.