
Malicious Defense: Red Teaming Phase 0 of “A Narrow Path”

Parv Mahajan
AI Safety Initiative, Georgia Institute
of Technology

Gita Pavuluri
School of Chemical and Biochemical
Engineering, Georgia Institute of
Technology

MacGyver Rawson
School of Civil and Environmental
Engineering, Georgia Institute of
Technology

Rahul Pavuluri
University of Georgia

With
ControlAI & Apart Research

Abstract

We use an iterative scenario red-teaming process to discuss key failures in the strict regulatory regime outlined in Phase 0 of “A Narrow Path,” and describe how a sufficiently insightful malicious company may achieve ASI in 20 years with moderate likelihood. We argue that such single-minded companies may easily avoid restriction through government-enforced opacity. Specifically, we outline defense contracting and national security work as a key sector of ASI vulnerability because of its tendencies towards compartmentalization, internationalization, and obfuscation, which provide ample opportunity to evade a governance scheme.

Keywords: Defense sector risks, AI policy, red teaming, superintelligence governance, regulatory implementation, safety cases

Introduction

To adequately evaluate the resilience of ControlAI's Phase 0 outline, our team used an approach that maximizes the realistic potential for malicious intent among a group or organization. This approach is necessary to evaluate resilience, as it places the maximum possible amount of stress on the guidelines set out by Phase 0, allowing for a complete evaluation of all components of the plan that contain even a small opportunity for exploitation. The research questions of this report were thus designed with that understanding of our approach in mind. **RQ1** addresses how this understanding is applied to the objective of the report in red teaming the Phase 0 outline. **RQ2** addresses a more specific application of our approach to the model of a private corporation.

RQ1 In what ways would a group or organization with unwavering intention to maliciously develop, by any means necessary, an artificial superintelligence (ASI) capable of limitless, catastrophic damage to humanity be able to circumvent the Phase 0 restrictions and successfully develop an ASI within 20 years?

RQ2 How would a private corporation be able to utilize internal company structure, economic decisions, government relations, international markets, research and development (R&D), and other strategies to successfully develop an ASI within 20 years?

We justify **RQ2** because a private corporation affords several key components to the success of our catastrophic scenario, including the ability to hide information from both the general public and national or international government bodies, the ability to directly and freely hold a stake in the market and maneuver in the international market to maximize opportunities for exploitation, and the ability to design a unique and advantageous company structure that enables the execution of highly risky behavior behind closed doors.

A hypothetical private corporation, hereinafter referred to as *Harbinger Systems International* (Harbinger), will thus represent one such corporation for the remainder of this report and is intended to demonstrate the failures and weaknesses of the Phase 0 outline. Harbinger and its theoretical operations in the next 20 years will also be used to answer **RQ2**.

Harbinger is characterized by a strong relationship with the U.S. Department of Defense, especially in its early operations. This relationship was established for the several affordances it provides. Specifically, the defense industry is extremely large in the United States, giving Harbinger access to an extremely profitable industry with a high demand for R&D to maintain U.S. dominance in the global defense market and protect national security. For instance, it is well-known that the U.S. is the primary weapons provider for several countries and organizations, with armed conflicts across the world providing immense profit for the U.S. as they export weaponry and defense systems to parties in the conflict. The U.S. also has the most powerful military of the world with some of the most cutting-edge military capabilities, and a large part of U.S. foreign policy and diplomacy relies on the U.S. maintaining its dominant position as the world's largest military. Thus, a relationship between Harbinger and the U.S. Department of Defense provides a

steady stream of R&D contract opportunities to generate profit for Harbinger, while simultaneously maintaining a sense of urgency and necessity among the U.S. government regarding Harbinger’s operations, creating a strong incentive for the government to act favorably towards Harbinger’s operations and maintain future partnership between both parties.

We begin by developing key assumptions about Harbinger to explore **RQ2**, and explain our red-teaming, discussion-based methodology. We then discuss the strategies our simulated executive team selected to carry out Harbinger’s mission and defend our choices to establish why the selected strategies are the most optimal for mission completion. Finally, we will discuss the implications of our hypothetical scenario on the ability of the Phase 0 outline to withstand real attempts to create ASI.

Methods

A preliminary analysis shows that the policy outlined in “Phase 0” is extremely politically intractable, and would present onerous legal, financial, technical, and social challenges for any would-be policy-maker. We do not expect these policies to be implemented, even half-heartedly, before a Point of No Return¹ is reached, which we forecast will occur in 3-7 years. Because of this, we choose not to investigate implementational feasibility gaps or perform comparative analyses - this regulatory regime would be truly unprecedented in human history, and we do not predict results in this vein will be valuable to governance efforts.

Instead, our interdisciplinary group of researchers conducted a three-hour strategic planning exercise as the chief executives of Harbinger to perform iterative red-teaming, with the hope of applying takeaways to other broad governance proposals. In rounds of informal discussion integrated with basic feasibility research and traditional group probabilistic forecasting techniques, we examined three increasingly difficult scenarios:

1. **The present day.** We examine how Harbinger may construct an ASI within 20 years given the current regulatory and social climate.
2. **Policy proposals 1-4.** We revise our previous plan to still be successful with the implementation of a prohibition on superintelligence development, prohibition on AIs capable of emergent breakout, prohibition on AIs use of other AIs, and requirement for a safety case before model deployment, as described in Phase 0. Of the six policy proposals, we believe these are the most likely to come to fruition because of their limited impact on current model use and limited dual use potential (e.g. there is little economic reason to have a model that attempts breakout).
3. **Policy proposals 1-6.** We revise our previous plan to integrate the restrictive licensing regime and international treaty described in Phase 0. We believe these proposals are the least likely to come to fruition, and that if an organization effectively evades proposals 1 - 4, it is extremely likely to pass proposals 5 and 6 with limited additional cost.

¹ Kokotajlo, 2020. *Less Wrong*. [The date of AI Takeover is not the day the AI takes over](#).

We propose this iterative red-teaming process as a useful constraining factor for our exercise by prioritizing simple solutions and best simulating an unscrupulous company attempting to evade a rapidly evolving regulatory regime.

To answer **RQ2** to its fullest extent, we designed our outline for the operations of Harbinger Systems International with a number of key assumptions:

1. The leadership of Harbinger Systems International (not necessarily every employee) acts out of complete malicious intent to destroy the human race by developing an ASI within 20 years.
2. The leadership of Harbinger Systems International is willing to lie to the public, governments, and international organizations, falsify reports, perform indecent business and scientific practices, and disregard any moral or professional obligations to truth and public safety to complete their mission.
3. Harbinger Systems International has \$5 billion USD in seed funding and employs a plurality of the top 20% of AI researchers in the United States. This is a conservative mirroring of OpenAI when they gained \$6.7 billion USD in 2024 in venture capital funding².

We allow for this single-minded goal as an effective framework for red-teaming because it is well adopted to quickly finding weaknesses in the policy proposals and closely matches early public statements by top AI companies.

Results

Second, the leadership of Harbinger Systems International will complete their mission while arousing as little suspicion from national or international bodies as possible. As such, the release of accurate company information to external sources is avoided wherever possible, and the use of disinformation is prioritized. Additionally, the aforementioned core values of Harbinger Systems International are to be known by the leadership members only, and they are to be masked as different, unsuspecting values to all other sources (including employees) to mask the true intentions of Harbinger Systems International. Third, the leadership of Harbinger Systems International designed their company following a very strict system of compartmentalization and classification. Employees are restricted so harshly to their specific domain within the company so as to severely limit their capacity to communicate with other groups of employees or understand the entire scope of Harbinger Systems International, and the majority of information is heavily classified.

Bad actors could exploit the proposed policy in different ways. The clearest method that stood out to us is minimizing public knowledge of current developments. This could be achieved by creating a need for security clearance or sensitive information attached to the process of developing the AI. This approach would work well with parts of the policy such as not allowing for open source / available code. This would

² Field, 2024. [CNBC. OpenAI closes funding at \\$157 billion valuation, as Microsoft, Nvidia, SoftBank join round](#)

be able to circumvent the policies that rely on monitoring of dangerous developments.

The context of Harbinger's relationship with the U.S. Department of Defense provides a clear and realistic explanation as to how Harbinger can navigate its operations without releasing much accurate information. Harbinger would be at the cutting edge of U.S. military R&D and operations, thus placing an extremely high sensitivity on many of the internal operations of Harbinger. Thus, it would be not only appropriate, but necessary for the sake of national security for Harbinger to heavily classify internal information and release only what is necessary to the U.S. government. In this way, Harbinger is able to tactically release necessary information and hide as much remaining information as possible, using one of its core strategies of disinformation and lying to provide even more protection to internal operations.

Many of the initial policy implementation involves "checks" that are audited by the company itself. This method is easier to exploit due to the lack of standardization and ability for bad actors to control their own restrictions. As such, Harbinger is fully able to use its strategy of disinformation and lying to falsify reports, fake evaluations, and pass ineffective checks where possible.

In order to enforce compute limits, national regulators are urged to work together with providers in order to monitor the amounts used. However, if a company had enough resources to avoid outsourcing, they would be able to circumvent this restriction, as they would not need to go through any government checks and regulations.

Harbinger will naturally be able to perform this, as their relationship with the U.S. Department of Defense and the extremely profitable defense industry will naturally generate enough profit to fund vertical acquisition of all their compute requirements, thus, again, circumventing the Phase 0 restrictions.

No way to judge intentions (basis of policy is good faith), definitions of found system improving found system - merely adding human in the loop to "sign off", development in different countries with lower restrictions,

Insider whistleblowing or internal employee betrayal could jeopardize the operations of Harbinger by exposing their true intentions to the public, but specific strategies are in place to prevent this from ever happening. For example, the extreme compartmentalization and classification measures of Harbinger make it almost impossible for employees to organize or communicate with each other. Additionally, the compartmentalization makes it very difficult for individual employees to understand the full scope or true intentions of Harbinger. The leadership of Harbinger are the only individuals with access to the true information, and their loyalty can be ensured by the relationship between Harbinger and the U.S. Department of Defense. Because Harbinger is allied with the U.S. Department of Defense specifically, the mission of Harbinger no longer becomes one of world destruction, but the protection of U.S. national security as the country attempts to maintain its position as a world superpower. In the past, the US (and other countries) have been willing to take extreme or potentially dangerous measures and methods in order to ensure the safety of their citizens by

being the most powerful or capable player. An example of this is the Manhattan Project, in which a group of individuals pursued a catastrophically dangerous technology for the simple motivation of possessing it before the enemy.

Discussion and Conclusion

While Phase 0 is unlikely to be implemented, our red-teaming exercise highlights the defense sector as a field particularly vulnerable to development of ASI. We recommend that governance proposals focus on this field and develop novel verification mechanisms that do not violate international standards of national sovereignty or develop new paradigms in national security.

Further, we find that AI companies' proclivity to information-sharing and academic publications, as exists today, is unnecessary for a single-minded malicious actor, and is likely the result of a coincidence in culture. We therefore encourage countermeasures that do not depend on public knowledge of AI paradigms, and that work to increase transparency of AI training and deployment even in sensitive sectors, without relying on company cooperation or professional rules of conduct.

Finally, we find that internal siloing will be an effective strategy to hide explicit progress towards catastrophic risks. We are unsure of how to combat this company culture, but look forward to future work in this vital area.