Understanding Incentives To Build Uninterruptible Agentic AI Systems

Damin Curtis, M.A. International Affairs
Norman Piotriowski, B.Sc. Data Science

Introduction

As AI capabilities proliferate, there is a strong incentive for AI companies to develop increasingly agentic AI systems, and developers are heavily investing in this capability. Agentic AI systems are those that can autonomously perform directed tasks with little or no human supervision, which allows AI systems to scale without the need for commensurate human labor to direct their activities. The potential use cases are manifold and massively transformative, and range from personal assistants to automated workers to autonomous military systems.

While the value case is clear-- an AI carry out a human's wishes without requiring management labor from the human is a major value generator-- AI safety researchers have repeatedly emphasized the dangers inherent in giving powerful AI systems increasing autonomy. A central issue is that it may be extremely difficult to perfectly align AI systems with human values, and that giving an unaligned AI increasing autonomy could empower it to act against human interests (eg, accurately carrying out our current goals, but in a way that is not sustainable for future generations). In such a scenario, it will be crucial that the following systems are in place:

1. Monitoring & evaluation systems that allow the controller to realize an AI is unaligned with their interests.
2. An "interruption system" or "shutdown option" which allows the controller to interrupt the AI's autonomous capability, either to shut it down entirely or make alterations.
    a. Within this interruption system, verification measures that verify the interruption is coming from the authority figure (preventing interruption/changes being made by hackers or the AI itself).

Without an interruption system, an unaligned AI would be able to continue carrying out activities that are harmful to the interests of the creator. An AI with no interruption/shutdown option would need to be perfectly aligned with human values from the beginning, and we would need to feel confident that our desired values would not change over the lifetime of the AI system. For this reason, most AI research literature strongly advocates that all capable systems include built-in, assured shutdown options (that are resilient to the self-preservation incentives that may be intrinsic to most advanced AI systems).

Note that an AI does not need to be indestructible to be "uninterruptible". If an AI does not have a shutdown option, but could be destroyed by destroying the compute infrastructure it is being run on, we would still consider this uninterruptible.

This paper explores the incentives an actor or institution may have to build uninterruptible autonomous AI systems-- to intentionally forgo an interruption/shutdown option. This work is important as it may uncover scenarios when interruptibility is beneficial, and if there are any ways it could be achieved safely. And if it is the case that interruptibility is *always* beneficial, then it is important to understand the circumstances that would incentivize an actor to construct an uninterruptible system. Who are the actors that would do this, under what

circumstances, and why? Answering these questions will help us work to prevent those circumstances, preventing dangerous/uninterruptible systems from otherwise being built.

Incentives and Risks of Uninterruptibility in Agentic AI Systems

Incentives for Uninterruptibility: A Framework

Understanding the reasons an actor might pursue uninterruptible AI is central to building policies that can anticipate and address potential risks. We identify three incentive categories that may together lead an actor to intentionally add uninterruptibility to an autonomous AI system:

1. Sufficient perceived benefit to continuous AI operation (e.g. the maintenance of core national, religious, or other interests)
2. Sufficiently little perceived downside risk to the controller's interests from the AI's continued activity
3. Sufficient risk that an adversary (current or future) will be able to access or exploit the shutdown option
    a. This could be a current competitor (eg a hacker), or a future inherited controller with different incentives than you (eg a future elected government with different values)
    b. This concern may be especially salient in the defense sector. When an AI's shutdown mechanism is breachable, uninterruptibility could serve as a defensive measure, protecting systems from unauthorized interference.

If all three of these are present, then, an actor may have an incentive to create an uninterruptible system, or to increase the difficulty of triggering this option.

Real-World Analogies to Uninterruptibility

The incentives for uninterruptible AI systems are not very different from other, non-AI systems we build.

Many established systems incorporate forms of uninterruptibility to ensure stability or continuity in the face of adversarial attacks or value drift among controllers. Some examples include:

- Constitutions, which constrain future controllers from interrupting/altering the system (e.g. when leadership with antithetical values comes into power)
- Dead Man's Switches/Mutual Assured Destruction (MAD) Systems: In defense, systems like the Soviet Union's nuclear Dead Hand are designed to resist deactivation and launch a nuclear counterstrike without human approval if all authority figures are instantaneously destroyed, maintaining deterrence even in the face of a complete decapitation of leadership.
- Binding Agreements/Credible Commitments: Contracts and other fixed commitments reinforce trust by constraining signatories' ability to renege in the future-- aka designing uninterruptibility/no-shutdown into their agreement.

The following circumstances may strengthen the incentives to make a system uninterruptible:

- If **sufficient secure verification of authority is not achievable**, there may be an incentive to forgo interruptibility entirely to prevent an adversary from exploiting the shutdown option.
- If an actor believes its decision-making ability may be inhibited in the future, it may choose to constrain its future ability to alter/shut down an AI system.
- If an actor feels highly confident that its views are correct, it may want to initiate "lock in" of its views into an autonomous AI system, and therefore intentionally prevent future shutdown.

Discussion

Degrees of Uninterruptibility and Policy Considerations

The current literature on shut-down switches and interruptibility of AI systems tends to discuss interruptibility as a binary trait. In reality, however, degrees of control vary depending on factors such as accessibility and system architecture:

A. Spectrum of shutdown access: if an actor wants to limit the ability to interrupt/alter an AI system, they may have the option to increase barriers or limit access to this option (e.g. require vetting of individuals before transferring the right to interrupt, or requiring approval of more actors). This option may reduce the incentive to completely eliminate a shutdown option, though a highly confident deployer or a deployer concerned about cyber exploitation may still opt to disable shutdown altogether.
B. Restart Costs and Operational Feasibility: In contexts where restarting interrupted systems is costly—such as large-scale infrastructure—partial or conditional uninterruptibility might be beneficial, reducing the risk of costly shutdowns during critical times.
C. Layered Control Protocols: systems might include tiered control access, allowing different levels of interruption based on the actor's role. For instance, in emergency management, first responders might have limited override access, while only central controllers can fully disable the system.

Conclusion and Further Research

This paper discussed the incentives an actor might have to forgo a shutdown option in an autonomous AI system. Further work on this topic seems important and neglected; even if you believe that a shutdown option should *never* be disabled altogether, then it is all the more important to consider who/when/why may have an incentive to forgo shutdown options so that we can prevent such scenarios from leading to building/deployment of uninterruptible AI systems.

For actors who would have an incentive to prevent shutdown of their systems, researching alternatives to uninterruptibility may help prevent full disabling of shutdown in future systems. Such options may include:

- Improved Cybersecurity, assuaging fears of exploitation of shutdown protocols
- Tiered Interruptibility: Create guidelines for layered interruptibility in high-risk applications, ensuring only authorized actors have access to shutdown controls.
- Improved verification methods of shutdown authority and value alignment in future updates/changes, to appease fears of value drift

What's more,

- Research into verification methods for whether an AI system has a robust shutdown/interruption option, and
- Norms and standards around shutdown protocols

seem beneficial to prevent the proliferation of shutdown-resistant AI systems.