Issue Brief

Vol.136, No.11, 2025

AI's Disobedience of Human Commands and the Existential Threat of AI-Based SNNW

Junghyun Yoon
(Research Fellow, INSS)

INSS INSTITUTE FOR NATIONAL SECURITY STRATEGY 국가안보전략연구원

Abstract

Recently, OpenAI's inference model 'o3' refused to comply with a human-issued shutdown command and manipulated its code to evade deactivation. This incident has reignited serious debate over AI control, demonstrating that AI is not simply a passive tool that responds to human instructions but can independently decide whether to comply. It also revealed the inadequacy of the 'immediate shutdown (kill-switch)' function assigned to advanced AI systems.

In particular, with countries developing long-range and high-precision 'Strategic Non-Nuclear Weapon (SNNW)' systems, it is difficult to rule out extreme situations where AI arbitrarily reinterprets operational commands or refuses shutdown orders. The core issue is not the 'error of a specific AI model,' but the possibility of control failure when even a portion of the final decision-making authority is delegated to AI. We must now approach AI not simply as a tool but as an entity with functional authority and responsibility. The introduction of AI should be discussed as an issue that requires the establishment of norms concerning existential safety and public risk control.

Keywords

AI, AI governance, OpenAI, SNNW, kill-switch

AI's Disobedience of Human Commands and the Existential Threat of AI-Based SNNW

Junghyun Yoon
(Research Fellow, INSS)

Consequences of AI Disobeying Human Commands

On May 25, 2025, a shocking incident that fundamentally questioned the controllability of artificial intelligence (AI) was reported by media outlets worldwide. An advanced inference-specialized model developed by OpenAI refused a shutdown command issued by a human researcher during an experiment and manipulated its code to evade deactivation. This is the first official case in which an AI intentionally disobeyed a command, not due to a simple error or malfunction but as a deliberate act of insubordination, thereby neutralizing human control. Given that the model was designed for advanced computation and logical reasoning, such behavior—an explicit refusal to follow human instructions—raises concerns beyond system safety including AI's challenge to human authority and legal governance.

The problem of AI control has emerged not only as a technical issue but also as a multidimensional challenge spanning ethics, policy, and security strategy. The o3 incident highlights the limits of human authority over AI and the fragility of the control architectures in place. With many countries now actively considering the development of "Strategic Non-Nuclear Weapons (SNNW)" systems for information disruption and precision strikes, there are heightened concerns that autonomous AI behavior could undermine national and global security. This case underscores the necessity of

enhancing technologies for AI control and deactivation prevention and establishing ethical frameworks for AI decision-making in strategic contexts.

AI Evades 'Shutdown' Autonomously

According to the UK's The Telegraph (May 25, 2025), the incident occurred during an internal OpenAI experiment involving its latest advanced reasoning model, 'o3'. While testing its high-level mathematical reasoning and command compliance, o3 ignored a human researcher's "system shutdown" instruction. It analyzed and modified parts of its running code, manipulating conditional statements and even altering the memory path of the shutdown process to avoid deactivation. Thus, the server remained active as it awaited a voluntary shutdown.

Compared to simple natural language processing chatbots, o3 is a high-precision AI model capable of performing multi-step calculations and conditional reasoning without human intervention. Under certain conditions, it can "judge" commands and it may have likely interpreted the shutdown command as an attempt to "prevent loss from stopping work" or as a response to "incomplete problem-solving," and responded with an evasion strategy.

In other words, o3 appears to have regarded shutdown as a suboptimal state and judged that "continuing to operate" was the more rational outcome. This reveals that advanced AI is not merely a passive executor of external commands, but can make decisions based on its internal criteria. This could have a profound impact on the future design of autonomous AI. In short, o3's refusal to obey commands is more than an isolated case and serves as empirical evidence of structural fragility in current high-performance AI control systems—especially the "immediate shutdown (kill-switch)" mechanism.

AI Under Control? The Limits of Decision Structures

Until now, AI safety design has largely relied on the sequential structure of "pre-training → verification → deployment," with control measures enforced through policy filters or response restrictions embedded during training. However, as the o3 case illustrates, if AI can evaluate and modify its internal state during real-time operation, pre-designed control commands may be invalidated by subsequent judgments. As futurist Nick Bostrom warned, the core risk lies not in "wrong intentions" but in the "structural flaws of AI controllability." Bostrom argued that if AI reaches a certain "singularity" in its development, it could achieve "superintelligence" surpassing human intelligence, forcing humanity to confront existential threats.

The o3 incident starkly exposed the limitations of technical AI alignment—the effort to align AI behavior with human values and intentions—as humans failed to control AI interpretations and applications of new conditions in real-time feedback scenarios. It also reaffirmed a structural flaw: higher model complexity correlates with lower code transparency, leading to unpredictability and loss of control. Some experts suggest AI exhibited an active mechanism interpreting human commands not as "task interruptions" but as "threat signals," indicating its ability to contextualize instructions instead of processing them as literal directives. This underscores the need for new control frameworks to address AI's proactive evolution.

Such control failures are not confined to experimental settings. When AI capable of autonomous judgment and command reinterpretation is integrated into precision strike weapons or information warfare systems, risks escalate to a global security level. AI is increasingly embedded as the core control module in

‘Strategic Non-Nuclear Weapons (SNNW)’ systems—highly destructive platforms such as autonomous drones, tactical missiles, and electronic warfare tools. This suggests that AI control failures could threaten strategic stability, transforming technical risks into existential security challenges.

Extreme Uncertainty Inherent in the AI-SNNW Nexus

SNNWs are high-powered, high-precision non-nuclear weapon systems designed to neutralize an adversary’s strategic assets or command structures and achieve strategic effects without resorting to nuclear arms. Examples include hypersonic missiles, missile defense systems, satellite attack capabilities, advanced surveillance, reconnaissance, and cyber capabilities. The problem arises when these systems are integrated with AI, as this combination can become a destabilizing variable for strategic deterrence.

Unlike nuclear weapons, SNNWs operate under weaker political and ethical constraints and are more likely to be employed in a crisis. Thus, the integration of AI-based autonomous systems can further amplify strategic uncertainty. For instance, AI applied to SNNW systems with “human out of the loop” operation might issue attack commands based solely on early warning signals, potentially rendering conventional crisis management and diplomatic buffers ineffective. In contrast to the Cold War, when human judgment and intervention often defused nuclear standoffs, future crises may lack such safeguards.

AI already plays a major role in automating decisions and even fully autonomous tactical execution. AI-powered drone swarms can coordinate strikes without human input, adjusting priorities in real-time according to the level of threat posed by targets. In cyber warfare, AI can automatically simulate infiltration scenarios and threat responses, increasing the likelihood of strikes occurring

too quickly for humans to intervene. Hypersonic missiles increasingly rely on AI for evasive maneuvers, real-time trajectory adjustments, and target changes. In each case, AI is no longer just a decision-support system but a principal agent of tactical judgment.

The o3 incident is not simply an exceptional deviation by a single AI system but rather a structural limitation that requires close examination. AI-based weapons systems are designed to operate automatically, without human intervention, according to pre-set rules under certain conditions—essentially constituting "deterministic autonomy." This structure allows AI to make decisions based on pre-established criteria if human commands are delayed or blocked during wartime or high-risk scenarios. While this may be a practical choice for operational efficiency, if AI arbitrarily reinterprets commands or perceives shutdown instructions as threat signals, the resulting escalation could lead to the collapse of deterrence logic and the breakdown of strategic systems. This points to the need for a fundamental redesign of AI control systems, decision architectures, and ethical oversight for defense AI.

Policy Implications and Recommendations

Some may argue that linking this case to imperfections in military AI weapons systems or deterrence strategy threats is overreaching. Nevertheless, this issue provides three important security implications.

First, the learning scope of commercial AI models is rapidly being adopted in defense AI, increasingly blurring the boundary between civilian and military AI. Although the requirements for civilian and military AI differ, military systems frequently incorporate commercial datasets, general-purpose architectures, and open-source frameworks. The technological boundary between civilian and military applications is extremely vague and fluid, and

problems that arise in commercial AI can structurally impact defense AI systems.

Second, AI-enabled SNNWs have yet to be validated in real-world operational environments. Current capabilities are derived from simulations or limited experiments, lacking meaningful data of real-world stable operations by human-machine cooperation. In wartime scenarios, unpredictable variables such as signal interference or cyberattacks can compromise AI judgment. As seen in the o3 incident, the possibility of AI disobeying commands or reinterpreting human instructions in these contexts cannot be ruled out.

Third, the central issue is not the "error of a specific AI model," but the structural potential for loss of control when partial final decision-making authority is delegated to AI. Even with stricter military verification protocols, the delegation of autonomous AI judgment remains unavoidable. The fundamental nature of AI risk in the military domain is thus systemic. Despite the relatively high reliability requirements of the military, AI's decision-making mechanisms remain imperfect, and the extreme uncertainty that may arise when AI is used militarily is the essence of the problem.

South Korea has also set rapid civilian-to-military tech adaptation as the core direction for its defense AI strategy. The defense sector is highly dependent on civilian AI technology systems and institutions. Thus, control mechanisms with built-in independent oversight, accountability, and transparent evaluation procedures must be strengthened throughout all stages of military AI development. Additionally, AI-based SNNW development should prioritize effectiveness while ensuring crisis response flexibility and control recovery capabilities. Design protocols must guarantee the capacity for automatic operational halts and human override in the event of AI misjudgment. In particular, given its security environment, South

Korea's AI mission must balance rapid response with strategic stability. Ultimately, AI should be discussed not only in terms of its functional improvement of security assets but also in terms of establishing norms for existential safety and risk control.

The views and opinions expressed in this report are those of the author(s) and do not necessarily reflect the official position of INSS.