Strategic Security Initiative (SSI)
ADVANCING INSIGHT ON GLOBAL SECURITY

POLICY BRIEF

THE BLACK BOX IN ORBIT:

AI, Strategic Ambiguity, and Crisis Stability

Tamar Shengelia

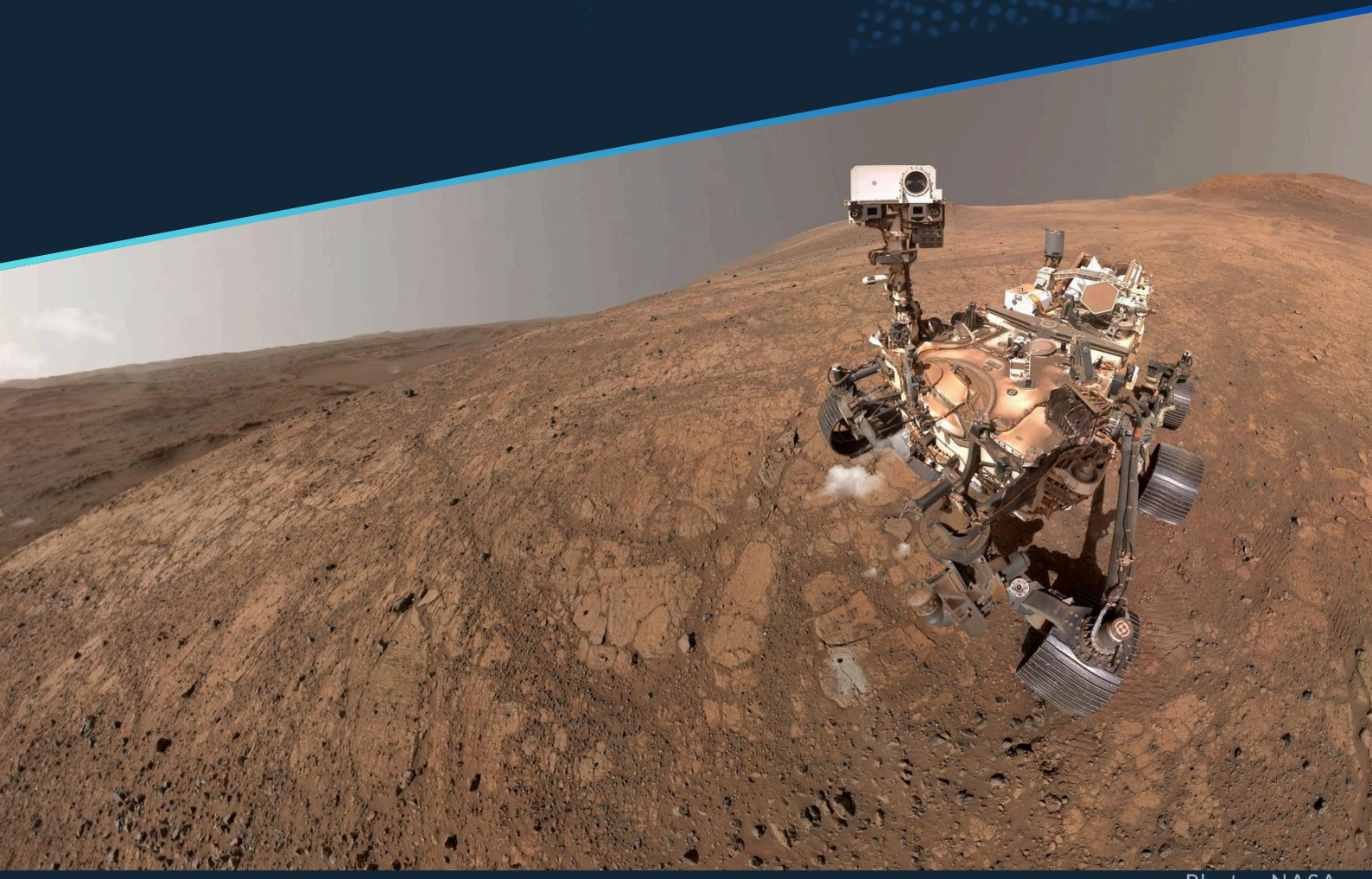

Photo: NASA

SEPTEMBER 2026

WWW.SSI-POLICY.ORG

THE BLACK BOX IN ORBIT: AI, STRATEGIC AMBIGUITY, AND CRISIS STABILITY

Autonomous systems, strategic signaling, and the emerging risks of escalation in space

FIRST PUBLISHED September 2026 by SSI

Authors: Tamar Shengelia

In 2020, a pair of Russian satellites were observed shadowing a multibillion-dollar U.S. spy satellite. The commander of the U.S. Space Force called the behavior "unusual and disturbing," Moscow described the craft as mere "inspector" satellites, and one analyst noted the maneuvers could have been anything from practicing on-orbit operations to signaling a capability (Hennigan, 2020). On April 28, 2026, the same class of capability had grown more refined when two Russian military satellites maneuvered to within roughly three meters of one another in low Earth orbit. This encounter occurred between COSMOS 2581 and COSMOS 2583 spacecraft orbiting approximately 585 kilometers above Earth, which was described by tracking analysts as a sophisticated operation with dangerous collision potential (Amiri, 2026). As orbital proximity operations increasingly incorporate autonomous, AI-driven control, the human ability to read intent collapses on both sides. Such systems rely on complex machine-learning systems whose internal reasoning may not be transparent even to the creators. This phenomenon is often described as the “black box”. Modern AI systems generate outputs through complex statistical processes that are difficult to interpret or predict. As a result, it may be impossible to fully explain why an autonomous system selected a target. AI and machine learning in space, especially military space, are applied in several ways: space situational awareness (SSA), predictive maintenance / failure detection, object tracking and identification, missile detection and trajectory prediction (Caso, 2024). Therefore, we can ask how opaque, AI-driven autonomous systems in orbit affect crisis stability between rival space powers? Opaque, machine-speed AI systems on orbital assets convert the structural pressures of the space security dilemma into an acute danger of inadvertent escalation. Because black-box autonomy is mutually illegible, unreadable by adversaries and often by its own operators, it severs the legible, attributable signals on which deterrence and crisis bargaining depend, while tightly coupling already-complex orbital systems in ways that make a “normal accident” in Perrow's sense not merely possible but probable. Space realism explains why states deploy such systems; Normal Accidents Theory explains how they fail catastrophically; the policy implication is that legibility-focused transparency and CBMs, not faster autonomy, are the only stabilizing path.

Theoretical framework

Space realism explains why states deploy opaque autonomous systems despite the risks: in an anarchic, self-help domain, the structural pressure to seize advantage overrides caution. Dolman takes the structural premises of classical/offensive realism and applies them directly to space, treating outer space as a strategic environment no different in kind from land, sea, or air, governed by the same unchanging logic of strategy (Dolman, 2002). Since Hobbesian anarchy lacks "common power to overawe" states, the system cannot enforce cooperation (p. 90). So, nationalist space exploitation will be confrontational, not cooperative. States

live in a self-help world where relative-gains logic is central because of anarchy: each superpower, unsure of the other's future capabilities, finds it "prudent to do everything possible to hinder the dominance of the other," using cooperation as "cover to buy time" while racing for control (p. 166). He rejects the "space sanctuary" school and supports the "ultimate high ground" school, urging the U.S. to withdraw from the Outer Space Treaty, seize low-Earth orbit, and exercise "benign hegemony" (pp. 147-161) Weaponization is inevitable, so a state should seize the dominant terrain first. This prescriptive program is contested, and the paper uses Dolman analytically to explain realist logic.

In interview with Brown, Bowen offers a more analytical contrast. Earth orbit is close, constrained, and increasingly cluttered with satellites and debris, like a narrow strait or coastline, according to his "cosmic coastline" thesis. He rejects "ultimate high ground" and "command of space" theories, arguing that NATO in Afghanistan shows that space does not win wars. Hispoint on mutual vulnerability is most important for this essay: the littoral framing makesspace infrastructurevulnerablebecauseland-based weaponscan reach it and many small actions can have big consequences. One miscalculation cascades on that exposed, tightly packed geography (Brown, 2025).

The mechanism of unreadable intent is captured by Jervis's security dilemma (1978), whose two variables are the offense-defense balance and offense-defense differentiation, and by Glaser's refinement (Glaser, 1997). When offense and defense are not differentiated, when weapons are dual-use a state buying forces to protect itself can only acquire forces that also reduce its adversary's defense. Because pure security-seekers and greedy states buy the same forces, no benign state can credibly distinguish itself, making signaling harder and riskier. Glaser claims that comparative net assessment of the balance can distinguish offense and defense (Glaser, 1997). Opaque AI is the narrowest case of non-differentiation, making hardware and decision logic indistinguishable, which causes intent to become unreadable.

In Normal Accidents Perrow's claim is that two structural properties of a system, not a part or operator, cause a specific type of catastrophe. The first is interactive complexity: a system has so many components(parts, procedures, operators) interconnected in non-linear, often invisible ways that designers cannot anticipate how failures will combine. “No one dreamed that when X failed, Y would also be out of order and the two failures would interact so as to both start a fire and silence the fire alarm ” (p.4). The second is tight coupling: processes are fast and can't be stopped, failed parts can't be isolated, and there's no other way to finish production safely. Complexity causes an unexpected interaction and tight coupling prevents recovery, so the disturbance spreads quickly and irretrievably. Operator action or safety systems may make the problem worse because no one knows what the problem is (Perrow, 1984).

These system properties make the accident inevitable and "normal." No matter the frequency, the term indicates that multiple and unexpected failure

interactions are inevitable given the system's characteristics. Slack in a complex system won't cause these disasters; tight coupling does (Perrow, 1984).

Redundancy and safety devices backfire. The interaction that wasn't expected happens because these parts and connections make the system more complex. Perrow highlights the Fermi I sodium-cooled breeder reactor's partial core meltdown in October 1966, which was caused by a triangular piece of zirconium installed as a safety device at the insistence of a prestigious committee, not on the blueprints. This same piece ripped loose and blocked the coolant flow, melting the fuel bundles (Perrow, 1984).

Sagan's 1993 Limits of Safety provides empirical evidence. Sagan finds that Normal Accidents Theory still fits the evidence better in a "comparative test" of U.S. nuclear command and control, the most dangerous technology ever built, with an excellent safety record, high leadership priority on safety, and strict discipline predicting exceptional safety. His conclusion is more pessimistic than Perrow's: accidents are inherent to these organizations and the safety record is more luck than design (Sagan, 1993).

On orbit, these frameworks converge. A kinetic threat to a satellite in low Earth orbit may allow only minutes for a response, so any delay, such as waiting ground station orders, could be disastrous. Autonomous space systems are interactively complex because there are multi-sensor fusion and opaque machine-learning guidance whose internal interactions designers cannot fully predict. Organisational learning makes opaque AI harder to govern than Sagan's nuclear systems because it learns from experience. According to Sagan, High Reliability Theory couldn't learn because of privacy concerns and the small number of accidents that happened. He also says that a "black box" system stops learning even after a near-miss (Sagan, 1993).

Mutual Illegibility

The security dilemma turns on a state's ability to read another's capabilities and intentions; when that reading fails, even a purely defensive posture looks threatening. Because black-box systems cannot be reliably interpreted by adversaries, operators, or anyone else, opaque AI in orbit directly attacks this informational core.

Three factors cause opacity: intentional secrecy, where systems are closed to protect competitive or state advantage; technical illiteracy, since reading code remains a specialized skill; and most fundamentally a mismatch between machine optimization and human reasoning, in which machine-learning systems operate in a high-dimensionality that defies human intuition and build internal representations that follow no human. Reading the code or hiring an auditor is often insufficient to explain why a decision was made, and a human-manageable list of criteria may be false reassurance rather than genuine understanding (Burrell, 2016). The deepest mismatch is not cured by access. In orbit, these three forms compound the dual-use ambiguity of

space hardware, so an observer faces secrecy, illiteracy, and the mismatch.

In other words, a system is opaque with respect to a given agent's knowledge and goals, so one stakeholder's epistemically relevant elements may differ from another's (Zednik, 2021). A creator may ask how the program works, an operator what it will do, and an examiner or affected party why it behaved as it did (Zednik, 2021). Crisis stability is crucial because the adversary is the worst-positioned stakeholder: there is no explanation for how one state makes its classified military AI legible to its opponent. Even a system with predictable outputs would leave the question of why it acted unanswered (Michel, 2020). Predictability and understandability are complementary.

This isn't a temporary engineering gap that better tools will fill. The most severe cognitive mismatch occurs when a neural network's millions of parameters are incomprehensible to humans, making it impossible to trace a prediction's origin (Sullivan, 2024). The U.S.-led Political Declaration on Responsible Military Use of AI prioritizes performance and transparency over explainability, driven by a performance-explainability tradeoff in which more accurate models are more opaque (Sullivan, 2024). The most capable systems are the least legible, misdirecting the trajectory.

These opacity features are now built into orbital systems. An AI-enabled system's internal logic is obscure enough in space to make it hard to trace decisions or attribute liability, and it's hard to tell if an outcome was caused by the original design or adaptive features acting in ways the producers didn't expect (Bratu & Ortega, 2025). AI is being pushed into space domain awareness and orbital warfare for latency-critical scenarios like rendezvous and proximity operations, where autonomous docking, collision evasion, and defensive manoeuvring are too fast for humans (Huynh, 2025). The maneuvering systems can also be used as offensive weapons, creating ambiguity, uncertainty, and vulnerability in the orbital environment (Staats, 2022).

The result is that each layer increases illegibility. An orbital approach observer cannot read the maneuver, which is dual-use, the payload, which may be hidden, or the decision logic, which even the system's operators cannot read. When security made intent hard to verify, opaque autonomy makes it unreadable, leaving the informational basis for reassurance and restraint useless. This legibility collapse cuts deterrence signals.

Signaling Failure

Illegibility hinders a state’s effort to read and send intentions. Interpretative signalling underpins deterrence and crisis bargaining. Schelling argues, the power to hurt accomplishes nothing unless it is communicated: to exploit a capacity to inflict harm, the adversary must understand what behavior will bring the violence on and what will keep it off, which requires the threat and promise to get through (Schelling, 1966). Coercion is therefore a message before an act, and even the use of force is a costly signal of intent and resolve meant to shape the adversary's expectations of what comes next (Schelling, 1966).

Opaque autonomy divides this channel.As mentioned above, an autonomous maneuver's intent cannot be read; the observer sees the act but not its purpose. The assuring half fails even more: a state cannot explain "this is only an inspection, I will not strike" when its system's behavior is unpredictable even to its operators. Reassurance requires the sender to guarantee the system's limits, which a black box cannot do. This creates a signaling system that cannot clearly communicate resolve or reassure a crisis.

This breakdown is not only on the sender's side. Because rival states train their systems on different data, against different threat models, and toward different objectives, no shared interpretive baseline exists between them. Even if each system were legible to its own operators, State A's autonomous system has no reason to interpret State B's maneuver as B intends it. Illegibility is therefore doubled because the sender cannot make its intent readable, and the receiver's system is itself mismatched to the sender's. This mutual unreadability does not require either system to be opaque, it is a separate failure that stacks on top of opacity, and at machine speed it is exactly the kind of unanticipated interaction Perrow describes, now between systems belonging to rivals who never designed them together.

Artificial intelligence and crisis stability research shows the risk. ACSIS Futures Lab wargame study found that intelligence gaps about the rival's AI capabilities, not AI itself, drove more dangerous behavior (Jensen et al., 2024). The study's only statistically significant escalation result was players probed more, signalled mid-range, targeted kinetic effectors, and used more aggressive "escalate to de-escalate" responses in the final round when capabilities were unknown (Jensen et al., 2024). Probing a rival to reduce uncertainty creates a new escalation risk, exactly like illegibility about an autonomous system.

Although thesamestudy warnsagainst overstatement, itshould betaken seriously.AI enhances strategy but does not change it, with no significant difference in overall escalation magnitude, contradicting the popular belief thatAI is inherently escalatory (Jensen et al., 2024). However, the orbital case removes the conditions that supported that stability. With time to think, players stayed informed, treated war as unlikely, and sought de-escalation off-ramps (Jensen et al., 2024). Not opaque decision logic, but capabilities a player could probe to reduce uncertainty in the experiment. CSIS's stability depended on the human pause and reducible uncertainty, but machine-speed orbital autonomy collapses the deliberation window and makes a black box unreadable. The study supports this essay's narrow claim that opacity-driven uncertainty increases while refuting its cruder claim.

Signalling failure ends with the sender. Signals assume an author whose intent they express, but the growing integration of commercial and private actors poses this question. In a "double dual-use dilemma" where private space companies, not just their technology, become militarised while maintaining a commercial face, oversight is lacking

(Altaf, 2025).According to Article VI of the Outer Space Treaty, assigning a private operator's actions to a state depends on authorization, supervision, and the "overall control" test, which is uncertain enough that a commercial actor could draw a non-belligerent state into conflict (Brown, 2022). An orbital act without a coherent state intent has no legible sender and cannot be classified as a threat, assurance, or signal.

At every level that deters, message content, reassurance credibility, and sender identity, signaling fails. The stabilizing function of crisis bargaining ends when states can no longer signal to each other, and outcomes are increasingly determined by what their tightly coupled systems do.

Accidental Escalation

Where the first two mechanisms concern what states cannot read and cannot say, the third concerns what their systems do on their own. As already mentioned above there are two structural properties of a system: interactive complexity and tight coupling (Perrow, 1984). Autonomous orbital systems possess both properties in acute form. They are interactively complex, fusing sensor data, conjunction screening, and autonomous guidance in ways no designer fully maps and they are tightly coupled, with closing velocities of kilometres per second and warning windows measured in minutes that leave no room to recover. The scale of the resulting demand is illustrated by the largest constellation operator. By one industry model, SpaceX's Starlink would receive almost eleven million conjunction warnings per year and need to perform over 1.2 million avoidance maneuvers which is roughly one every twenty-six seconds (Viasat, 2022). A rate of that order is not one a human operator can occupy, the loop must be automated, and the safety of the orbital environment comes to depend on systems acting on their own.

That automation does not remove the danger so much as relocate it, which is precisely Perrow's most counterintuitive claim, that safety devices, being added components and interconnections, raise the very complexity that leads the unexpected interaction. The point is not speculative in orbit. SpaceX's own filing to the U.S. Federal Communications Commission discloses that its autonomous collision-avoidance process does not screen a planned maneuver against other objects and does not require coordination with other operators before acting (Viasat, 2022). One satellite's evasive maneuver is therefore invisible to another's avoidance system: the safety mechanism itself becomes a new source of unanticipated interaction, exactly the dynamic Perrow describes.

This has already led to a system accident, therefore it's not just hypothetical. On 10 February 2009, the active Iridium 33 and the defunct Cosmos 2251 collided at roughly 790 kilometres' altitude. This was the first accidental hypervelocity collision of two intact satellites, producing more than 1,800 trackable fragments and far more lethal debris too small to track, which will persist for decades (Viasat, 2022). The collision is a normal accident in Perrow's sense: there was no malice and no single blamable failure, only two tracked objects

in a tightly coupled environment in which the warning that might have prevented the collision was never acted upon. What makes this instructive is the contrast with the environment that made it happen. In 2009, conjunction screening was a relatively slow, human-paced process, on average NASA's entire fleet of more than twenty-five robotic satellites conducted only two avoidance maneuvers a year (Johnson, 2009). The accident still happened. Multiply that sparse, human-screened environment by the eleven-million-warning scale of a single mega-constellation, and give the response to uncoordinated, black-box autonomy, and the conditions that produced one collision are reproduced at a greater density and speed.

A collision, in itself, is a safety accident rather than a crisis, but when the same dynamics occur between adversaries rather than between a piece of debris and a commercial satellite, are how an accident becomes an escalation. Amid the illegibility established earlier, an anomalous autonomous maneuver, or an accident misread through a broken signalling channel, can be interpreted as a hostile act and answered before any human has time to determine what actually happened. On the terms of the technology, this is not a remote possibility: self-learning systems, by their probabilistic nature, can produce unpredictable outcomes that may lead to operational anomalies or unintended conflict escalations (Bratu & Ortega, 2025). The security dilemma of the framework and the normal accident of Perrow here become the same event — a spiral driven not by what either side intends but by what their tightly coupled, opaque systems do, at a speed that it leaves no room for a human to step in.

The result is escalation that no one chose. It is the purest form of accidental escalation, created by system dynamics rather than by decision. It is this prospect, not a malicious AI, but an ungoverned, illegible, tightly coupled one, that makes opaque autonomy in orbit a source of crisis instability.

Conclusion

AI on orbital assets doesn’t create a new threat, but rather turns an old one into a weapon. It does it by adding to the structural problems of space security dilemma and taking away the clarity that has always been needed for deterrence and crisis bargaining. Because mutual illegibility breaks the informational basis for reading intent, an adversary can’t tell the difference between an inspection and an attack, or defensive move from offensive one. Signaling failure follows directly as a state cannot communicate resolve or offer credible reassurance when its own systems behave in ways their operators cannot predict. In the end tight coupling completes the spiral. The same accident that means little between a piece of debris and commercial satellite becomes something far more dangerous between rivals: it is answered automatically, at machine speed, before any human can understand what happened.

In terms of policy, the implication goes against the prevailing trajectory. If the threat is not autonomy itself but illegibility, then the solution cannotbefaster or more powerful autonomous systems. The U.S.-led Political Declaration's focus on

performance over explainability is a clear sign of this. It's legibility that keeps things stable, such as confidence-building measures, transparency regimes, behavioral norms for rendezvous and proximity operations, and pre-launch notification or maneuver-coordination protocols that restore some of the attributable signaling that black-box autonomy destroys. It can help us understand Bowen's cosmic coastline. It's not about taking the high ground, as Dolman suggests, to keep things stable in orbit. Instead, it's about making the behavior in that small, constrained area clear to everyone who shares it.

Realists will say this is unrealistic because in a self-help setting, a state that makes its military AI readable gives up the very advantage that opacity gives, therefore no smart actor will agree to transparency that their rival will not. This is a serious objection, but it misreads what is being proposed. To make behavior predictable, legibility-focused CBMs do not require disclosing decision logic or source code. Instead, they need to make maneuver intentions public, agree on approach distances, and set the orbital equivalent of rules of the road. States with profound strategic distrust built exactly such regimes during the Cold War because unpredictable behavior in a closely linked environment was a threat to both sides. As this essay's main point, orbit now meets that condition. AIs that aren't controlled, can't be read, and are tightly linked are what pose the most threat. Unlike AI systems themselves, states can still choose to control this threat for now.

References

Altaf, Z. (2025, April 9). As more countries enter space, the boundary between civilian and military enterprise is blurring. Dangerously. Bulletin of the Atomic Scientists. https://thebulletin.org/2025/04/as-more-countries-enter-space-the-boundary-between-civilian-and-military-enterprise-is-blurring-dangerously/

Amiri, A. (2026, May 10). “Whatever Russia Is Testing, It’s Sophisticated”: A 3-Meter Close Pass That Could Have Scattered Wreckage for Decades. The Daily Galaxy - Great Discoveries Channel. https://dailygalaxy.com/2026/05/cosmos-satellites-close-approach-debris-risk/

Bratu, I., & Azcárate Ortega, A. (2025). “Space Security & New Technologies 1 – Visionary Versus Reactionary. The Future of Space Security in the Age of Artificial Intelligence.” UNIDIR Visionary versus Reactionary. https://doi.org/10.37559/WMD/25/Space/03

Brown, D. (2025, May 19). Dr Bleddyn Bowen – The politics of spacepower: Realities and imaginations of war in space, and the evolution of the space industry. UK Forum on a Multipolar World. https://durhamforummpw.wordpress.com/2025/05/19/dr-bleddyn-bowen-the-politics-of-spacepower-realities-and-imaginations-of-war-in-space-and-the-evolution-of-the-space-industry/

Burrell, J. (2016). How the Machine “thinks”: Understanding Opacity in Machine Learning Algorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512

Brown, T. (2022, July 8). Ukraine symposium – The risk of commercial actors in outer space drawing states into armed conflict. Lieber Institute West Point. https://lieber.westpoint.edu/commercial-actors-outer-space-armed-conflict/

Caso, S. (2025). Emerging technologies in military space operations: current applications and future research for educational and training purposes. International Journal of Training Research, 23(2), 140–155. https://doi.org/10.1080/14480220.2024.2431482

Glaser, C. L. (1997). The Security Dilemma Revisited. World Politics, 50(1), 171– 201. http://www.jstor.org/stable/25054031

Dolman, E. C. (2002). Astropolitik: Classical geopolitics in the space age. Frank Cass. 10. Michel, A. H. (2020). ‘The Black Box, Unlocked: Predictability and

Understandability in Military AI. Unidir.org. https://unidir.org/publication/the-black-box-unlocked/

Huynh, C. (2025, June 24). AI on the Edge of Space | Center for Security and Emerging Technology. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/ai-on-the-edge-of-space/

Jensen, B., Atalan, Y., & III, J. M. M. (2024). Algorithmic Stability: How AI Could Shape the Future of Deterrence. Www.csis.org.

https://www.csis.org/analysis/algorithmic-stability-how-ai-could-shape-future-deterrence

Jervis, R. (1978). Cooperation under the Security Dilemma. World Politics, 30(2), 167–214. doi:10.2307/2009958

Johnson, N. (2009, October 16). NASA Technical Reports Server (NTRS). Ntrs.nasa.gov. https://ntrs.nasa.gov/citations/20100002023

Perrow, C. (1984). Normal accidents: Living with high-risk technologies. Basic Books.

Sagan, S. D. (1993). The limits of safety: Organizations, accidents, and nuclear weapons. Princeton University Press.

Staats, B. (2022). Mitigating Security Risks and Potential Threats of Emerging Rendezvous and Proximity Operations. Astropolitics, 20(1), 64–92. https://doi.org/10.1080/14777622.2022.2080547

Sullivan, S. (2024, August 28). Targeting in the Black Box: The Need to Reprioritize AI Explainability - Lieber Institute West Point. Lieber Institute West Point. https://lieber.westpoint.edu/targeting-black-box-need-reprioritize-ai-explainability/

Thomas C. Schelling. (2020). Arms and Influence. Yale University Press. 20. Viasat. (2022). Managing Mega-Constellation Risks in LEO.

https://www.viasat.com/content/dam/us-site/corporate/documents/Viasat_White_Paper_Managing_Mega-Constellation_Risks_in_LEO_Updated_Nov_22.pdf

W.J. Hennigan. (2020, February 10). Exclusive: Strange Russian Spacecraft Shadowing U.S. Spy Satellite, General Says. Time; Time. https://time.com/5779315/russian-spacecraft-spy-satellite-space-force/

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. Philosophy & Technology, 34. https://doi.org/10.1007/s13347-019-00382-7

Strategic Security Initiative (SSI)

ADVANCING INSIGHT ON GLOBAL SECURITY