

Chapter 8

Offence–Defence Asymmetry in the Age of Artificial Intelligence: Implications for the Cyber Threat Landscape

Apala Ghosh, PhD Research Scholar, Department of International Relations, Jadavpur University

Abstract

The swift integration of artificial intelligence (AI) into digital architecture is generally considered a transformative process in the area of cybersecurity. However, much of the current literature on the topic considers AI either as a technology that amplifies cyber capabilities or as a new form of governance challenge that demands regulation. This chapter presents a different case: that AI is best conceptualized as a structural variable that alters the balance of offence and defence in cyberspace. By using the theoretical framework of International Relations (IR) theory, specifically the literature on offence-defence asymmetry and security dilemmas, this chapter contends that the current form of AI systems aggravates the pre-existing imbalance between offence and defence in cyberspace, thereby exacerbating the instability of the global cyber threat environment. The offence-defence balance approach, which has traditionally been used to assess conventional and nuclear security, examines the prevailing technological and strategic conditions of a given environment to determine whether it is more advantageous to launch an attack or defend against one. When the costs of offence are lower than those of defence, states are likely to pursue preemptive strategies, the threat of escalation rises, and deterrence policy becomes unreliable. Cyberspace has long been considered an offence-preferred environment, where attackers need only find one vulnerability to breach the system, while defenders must protect a complex and constantly evolving target, where attribution is slow and uncertain, and where offensive capabilities can be easily copied at a low cost. The pre-AI cyber environment is already characterized by asymmetrical vulnerability. Cyber operations are different from traditional warfare in that they leverage systemic interconnectivity, civilian infrastructure, and dual-use technology. The defensive effort is high due to the expansive and constantly updated nature of digital networks, while the offensive community, whether state or non-state, enjoys the benefits of obscurity and anonymity. These characteristics establish what can be termed a persistent offence bias in cyberspace. The chapter will establish this baseline in order to evaluate the impact of AI on the balance and in which direction. Second, the chapter will explore how particular AI capabilities increase the reach of the offensive community. AI-enabled vulnerability scanning, code generation, and machine learning-driven reconnaissance decrease the technical skill set and time previously required for sophisticated cyber operations. Generative tools facilitate scalable and context-specific social engineering attacks, reducing the traditional trade-off between credibility and scalability in phishing and fraud. Adaptive malware can iterate independently in response to defensive efforts, improving the resilience and longevity of intrusions. Taken together, these trends reduce the barriers to entry for malicious actors, accelerate attack cycles, and improve the strategic reach of relatively resourceconstrained actors. Notably, the chapter will show how AI interacts with existing cyber structures to increase the flexibility of the offense. Third, the chapter will critically evaluate whether defensive AI mitigates

these trends. Advocates claim that machine learning improves anomaly detection, threat intelligence analysis, and real-time response capabilities. While these tools improve defensive monitoring, they are subject to structural and institutional constraints. Defensive AI systems require high-quality training data, produce a high rate of false positives, and require human direction within legal and administrative frameworks. Furthermore, defensive infrastructures are typically situated in complex public-private ecosystems, which impede coordination and response. By contrast, the offensive community enjoys fewer normative or institutional constraints. The imbalance is not only based on technical capability but also on organisational and regulatory environments that shape the use of AI. On the basis of this comparative analysis, the chapter goes on to examine the wider implications of international security. If AI further expands the offence defence gap in cyberspace, several implications follow. First, the risk of escalation could rise as the condensed time frames for detection and response leave less time for decision-makers to deliberate. Second, deterrence becomes more problematic as the delay between attacks and attribution trails the rapid execution of AI-enabled attacks. Third, the proliferation of superior capabilities to non-state actors further cloud state-centric visions of cyber stability. The chapter thus places AI in the context of debates on strategic stability, contending that cybersecurity must be understood not only as a technical challenge but also as a space of evolving power relations, and explores structural avenues for mitigating AI-driven asymmetry through architectural reform, collective defence coordination, improved attribution, and normative adaptation. The aim is to offer a framework by which policymakers and scholars might assess future technological developments without recourse to either alarmism or technological optimism.

Keywords: *Offense–Defence Asymmetry, Artificial Intelligence (AI), Cybersecurity, Cyber Conflict, Strategic Stability, Emerging Cyber Threats*

Introduction

The integration of artificial intelligence (AI) into digital infrastructure has been broadly described as a transformative moment for the field of cybersecurity. However, the current state of debate on this issue tends to vacillate between a vision of technological utopianism and a more alarmist vision of the future, without properly locating the phenomenon of AI within the existing body of literature on international security analysis. This chapter will contend that AI should not be considered simply a new “emergent risk” or “governance challenge” for cybersecurity. Rather, it should be considered a new “structural variable” that alters the balance of offence and defence in cyberspace.

The balance of offence and defence is a foundational concept in the field of International Relations (IR) theory, and it seeks to assess the prevailing strategic and technological conditions in a given international system as to whether they are more conducive to offensive or defensive action. When a system is characterised by a structural advantage for offence, this tends to increase the likelihood of escalation and undermine the stability of deterrence. Cyberspace has long been a system that is characterised by offence-preferential traits such as asymmetrical vulnerability, the difficulty of attribution, and the systemic interconnectedness of digital networks.

This chapter will argue that AI tends to further exacerbate these asymmetries by reducing the costs of innovation, increasing the tempo of operations, and diffusing advanced capabilities beyond the traditional state actor.

The main argument that will be made in this chapter is threefold. First, cyberspace is already a system that is characterised by a structural advantage for offence. Second, the current state of AI technology tends to increase this imbalance in a manner that is more pronounced on the side of offence than on the side of defence. Third, this has a number of implications for strategic stability, escalation, and the diffusion of power in international politics.

By locating the analysis of AI within the existing body of literature on offence-defence theory and applying it to the phenomenon of AI-enabled cyber operations, this chapter hopes to provide a more structured explanation of how technological change tends to alter the strategic equilibrium of international politics.

Section I

Reconsidering the Offence–Defence Balance in Cyberspace

The incorporation of artificial intelligence into digital systems has sparked an increasing amount of commentary speculating about the potential for transformation in the nature of cyber warfare. However, much of this literature is conducted without specifying the theoretical framework in which the potential for transformation is being measured. In order for AI to be considered a material factor, it must be considered not only as an innovation in technology but also as a factor within pre-existing frameworks of international security studies. Of these, the offence-defence balance is a particularly useful point of departure¹⁶². Rather than asking whether AI is “dangerous” or “useful,” the offence-defence balance asks a more systematic question: does AI change the relative ease of attack and defence in cyberspace?

The offence-defence balance has long been used in the conventional and nuclear contexts. At its heart is a simple but profound observation: the balance between offence and defence determines the strategic incentives for conflict, arms racing, and strategic miscalculation. When offence is perceived to be superior, states are more likely to pursue pre-emptive strategies and to view defensive preparations as threatening. When defence is superior, the price of aggression increases and deterrence stabilizes strategic interaction. The balance itself is not simply a descriptive tool for military capability but shapes strategic behaviour through perception and expectation.

Applying this framework to cyberspace demands a nuanced conceptual translation. In contrast to territorial space, cyberspace is not geographically bounded but rather defined by its networked architecture. Occupation does not involve control of territory but rather access to systems and data. The relevant strategic question is therefore whether it is more structurally feasible to gain access to and exploit these networks than to comprehensively protect them. By this measure, the offence in cyberspace is simply gaining unauthorized access, obtaining or manipulating information, or disrupting its functionality. Defence is simply preventing unauthorized access, detecting intrusion, and restoring system integrity.

¹⁶² Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations* (Oxford University Press 2017).

Even before the advent of more sophisticated AI systems, cyberspace has shown characteristics that make defensive primacy difficult. Digital systems are complex and interconnected. Software development is a layered process, involving third-party libraries and constant updates. Each layer is potentially vulnerable. Because digital systems are constantly evolving, the defender must deal with an ever-expanding attack surface. The defender's problem is therefore systemic and ongoing. The attacker, on the other hand, needs only one point of vulnerability. This is not a guarantee of success, but it certainly establishes an asymmetry.

The balance of offence and defence in cyberspace is also affected by information asymmetry. Attribution of cyber-attacks is often difficult and delayed. Unlike conventional attacks, which often betray their origin through force projection, cyber-attacks can be routed through multiple sources and masked by technology. This makes deterrence credibility difficult to establish. If retaliation is dependent on successful attribution, and attribution is difficult or delayed, the cost of the offence is perceived to be lower. Structural ambiguity therefore combines with vulnerability to tip the balance in favour of the offence.

Nevertheless, the existence of an offence bias in cyberspace does not necessarily mean that innovation in defence is a futile endeavour. The balance between offence and defence is not static. Encryption protocols, multi-factor authentication, intrusion detection systems, and international norms have all changed the character of cyber engagement over time. The challenge is not to assert that cyberspace is necessarily offensive in nature, but to assess how technological change shapes relative costs.

Artificial intelligence must be placed in the context of this constantly shifting balance of power. AI systems improve pattern recognition, facilitate analysis, and support adaptive responses. These characteristics are dual-use. They can improve anomaly detection and defensive monitoring; they can also support automated reconnaissance, exploit development, and deception¹⁶³. The strategic effects of AI depend on whether it upends the ratio of offensive to defensive effort. The role of effort must not be dismissed. In offence-defence theory, the key variable is not capability but cost. A technology that lowers the cost of performing offensive actions more than it lowers the cost of maintaining defence will shift the balance even if both sides are better off in absolute terms. Marginal change, not absolute change, is the key to structural advantage. In cyberspace, cost is more than just money. It includes time, expertise, coordination, and institutional constraint. Traditional cyber offence has required technical expertise and careful reconnaissance. Cyber defence requires constant monitoring, organizational coordination, and compliance with legal and regulatory requirements. In assessing the effects of AI, the key question is whether it changes these cost factors symmetrically.

A further conceptual clarification is needed on the role of technology and perception. The offence-defence balance of power theory highlights that what matters is not just material advantage but perceived advantage. If states perceive that AI-enabled cyber offence is gaining the upper hand, they may choose to counter by building their own cyber offence. These actions can create competitive spirals even in the absence of conclusive evidence that cyber offence has the

¹⁶³ Rebecca Slayton, 'What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment' (2017) 41 *International Security* 72.

advantage. In this way, AI can shape strategic stability through expectation effects as much as through capability effects.

Cyberspace also brings about complexities that do not exist in the traditional realm. Offensive and defensive capabilities can be the same. Research used for vulnerability identification for defensive patching can be used for offensive exploitation. Intelligence gathering for network defence is similar to intelligence gathering for attack. The dual-use nature of cyber capability makes it difficult to distinguish between the two, which were traditionally viewed as more distinct by offence-defence theory. AI makes this even more difficult because the machine learning models used for defensive anomaly detection can have the same underlying architecture as those used for offensive pattern recognition¹⁶⁴.

Moreover, cyberspace is inhabited not only by nation-states but also by corporations, criminal groups, and loosely connected collectives. The offence-defence balance thus applies at various levels of analysis. A technology that benefits non-state offensive actors can upset interstate relations indirectly. The spread of capability makes it difficult to apply traditional deterrence models, which rely on the existence of specific state-based adversaries. It is important to note that these complexities do not make the theory of offence and defence redundant. Rather, they highlight its versatility. The theory is a useful tool for examining the relationship between new technologies, incentives, and vulnerabilities. Rather than dwelling on the sensational implications of autonomous cyberwar, the offence-defence approach focuses on relative effort, scalability, and institutional constraints. The analytical argument made in this chapter is that artificial intelligence constitutes existing cyber asymmetries in ways that systematically favour offence. This argument is not based on technological determinism. Artificial intelligence does not necessarily enable attackers relative to defenders. Instead, it amplifies existing structural characteristics that already disadvantage defenders. Artificial intelligence decreases the marginal cost of reconnaissance, facilitates scalable deception, and accelerates adaptive exploitation, thereby reducing the marginal cost of offensive activities. Defensive systems are aided by artificial intelligence-assisted detection and response, but they remain limited by complexity and organizational factors.

The applicability of this argument extends beyond the technical realm of cybersecurity. If AI further exacerbates the offence-defence imbalance in cyberspace, it carries implications for strategic thinking among states¹⁶⁵. Perceived vulnerability may lead to pre-emptive strategies, expanded offensive capabilities, and heightened suspicion. Conversely, if defensive innovation does not keep pace, confidence in digital security may erode, with attendant economic and political consequences.

¹⁶⁴ Henry Farrell and Abraham L Newman, 'Weaponized Interdependence: How Global Economic Networks Shape State Coercion' (2019) 44 *International Security* 42.

¹⁶⁵ Martin C Libicki, *Cyber Deterrence and Cyberwar* (RAND Corporation 2009).

Section II

The Structural Foundations of Cyber Offence Advantage

Any hypothesis about the role of artificial intelligence in changing the balance of offence and defence in cyberspace must begin with a description of the underlying distribution of advantage. Otherwise, AI may be seen as the source of instability in itself, rather than a catalyst for the underlying conditions. The strategic nature of cyberspace was established well before the advent of massive machine learning systems. The architecture of cyberspace, its economic model, and its governance regime have created a space in which the superiority of defence has always been impossible to achieve.

The first structural basis for offence superiority is the architecture of digital systems themselves. Today's networks are not point solutions or self-contained systems. They are complex ecosystems that include operating systems, applications, firmware, third-party libraries, cloud infrastructure, and human interfaces. Each layer of the ecosystem introduces code, and each piece of code introduces the possibility of error. The size of today's digital systems ensures that vulnerabilities are not exceptional but statistical certainties. In such an environment, the defender's job is not simply to defeat an attack but to sustain systemic integrity through a constantly shifting technological landscape.

This is an architectural truth that creates a deep asymmetry of effort. Defence must be comprehensive. Offence only needs to be selective. An attacker needs to find one vulnerability that has not been noticed; a defender has to protect thousands of possible points of entry. The asymmetry is not merely one of degree but of logic. Failure for defence can occur through a single mistake; failure for offence does not preclude another attack. Since the cost of multiple attacks is low, persistence becomes a feasible strategy for attack. The complexity of digital systems thus embeds asymmetry at the level of system design.

The second basis for offence superiority is found in the economics of software development and cybersecurity investment. Cyber systems are built in a competitive market environment that favours innovation, speed, and functionality. Security is often an afterthought. Patching vulnerabilities takes time and effort that does not produce direct revenue. For many actors in the private sector, cybersecurity is a cost centre rather than a strategic investment. This economic incentive structure creates an uneven defensive posture across sectors and borders.

Attacks are driven by a different economic model. The development of exploits can be directly monetized through ransomware, data exfiltration, or denial-of-service attacks. Black markets enable the trading of vulnerabilities and malware¹⁶⁶. Once developed, exploit code can be repurposed and adapted at a low additional cost. The commoditization of offensive capabilities reduces barriers to entry and allows capability to be widely distributed outside of elite technical circles. In this way, the economics of the cyber domain favour successful exploitation over successful prevention.

¹⁶⁶ Michael N Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017).

Third, the attribution challenge undermines the credibility of deterrence. In traditional military conflict, tracing the source of attack is often an immediate or physically traceable process based on evidence of force projection. Cyber-attacks can be masked through proxy servers, compromised hosts, and routing techniques. Attribution is often a probabilistic process rather than a certainty. Even when technical analysis points to a likely source, political factors can complicate public attribution and retaliation.

Deterrence requires a credible threat of retaliation. When attribution is uncertain or delayed, there is a risk of miscalculation in response. This generates strategic ambiguity. Such ambiguity can be exploited by offensive actors who operate below the threshold likely to provoke escalation. Espionage, data manipulation, and limited disruption can be conducted in ways that test defensive resolve without triggering overt confrontation. Structural uncertainty therefore reduces the expected cost of certain forms of cyber aggression.

Temporal dynamics further reinforce asymmetry. Cyber intrusions are often gradual and may remain undetected for extended periods. During this time, attackers conduct reconnaissance, map networks, and prepare for disruption. Detection frequently occurs only after significant compromise has already taken place. The delay between intrusion and response provides attackers with informational advantage. Even when ultimately effective, defensive measures are reactive.

The defender's temporal burden is one of constant vigilance. Monitoring, updating, auditing, and responding require sustained organisational effort. Offensive actors can choose the timing of engagement strategically and are not required to maintain continuous visibility. This asymmetry of engagement places structural strain on defensive institutions.

A further imbalance arises from the diffusion of capability among both state and non-state actors. Unlike nuclear or advanced conventional weapons, cyber capabilities do not depend on large-scale industrial infrastructure. Skilled individuals, small groups, and criminal enterprises can develop impactful tools. The dissemination of knowledge through online platforms and open-source repositories accelerates this process. As capability spreads, the number of potential adversaries expands. Defensive actors must therefore guard against a diverse spectrum of threats, ranging from state-sponsored units to opportunistic criminal networks.

This diversity complicates strategic stability. Traditional deterrence models assume identifiable adversaries with centralised command structures. In cyberspace, potential actors include loosely organised networks whose motivations may not align with geopolitical rationality. Even if states achieve mutual deterrence, non-state actors may operate outside those constraints. The structural consequence is increased unpredictability.

Institutional and normative constraints further shape asymmetry. Defensive cybersecurity, particularly in democratic contexts, is embedded within legal frameworks governing surveillance, data retention, and automated response. Security measures must be balanced against privacy and civil liberties. Offensive actors, especially criminal organisations and authoritarian regimes, often face fewer normative constraints. This divergence affects the speed and intensity with which capabilities can be deployed. While defenders operate within compliance regimes, attackers iterate with comparatively fewer restrictions.

Interdependence compounds these structural factors. Digital infrastructure supports financial systems, healthcare networks, energy grids, and public administration. Civilian and military domains are interconnected. A vulnerability in one sector can cascade across others. The interconnectedness that enhances economic efficiency simultaneously amplifies systemic risk. Offensive actors benefit from this condition, as disruption in a single node can generate disproportionate consequences.

Taken together, these structural features—architectural complexity, economic incentive misalignment, attribution difficulty, temporal asymmetry, diffusion of capability, institutional constraint, and systemic interdependence—have produced a cyber domain in which defence faces intrinsic burdens. This does not imply the inevitability of offensive dominance; defensive innovation has strengthened resilience in numerous contexts. Nevertheless, the underlying asymmetry endures.

Establishing this baseline is essential for evaluating the impact of artificial intelligence. In a structurally neutral domain, the dual-use character of AI might generate symmetrical enhancement. In a domain already inclined toward offence, however, technologies that reduce operational cost or accelerate exploitation are likely to magnify imbalance. The analytical task that follows is therefore to determine whether AI interacts with these structural conditions in ways that amplify offensive leverage more than defensive resilience, and whether such amplification reshapes strategic incentives at the international level.

Section III

Artificial Intelligence as an Asymmetry Multiplier

Artificial intelligence does not operate in a neutral strategic environment. Instead, it is superimposed upon a cyber environment already marked by architectural vulnerability, economic misalignment, attribution uncertainty, and temporal imbalance. The relevant question, therefore, is not whether AI enhances capability on both sides—it clearly does—but how it interacts with these pre-existing asymmetries to alter their scale or character. The argument advanced here is that AI functions as an asymmetry multiplier¹⁶⁷. It does not create structural imbalance; rather, it amplifies it by reducing the marginal effort required for offensive action more substantially than it reduces the systemic burden of defence.

The concept of marginal effort is central to this claim. In strategic analysis, absolute increases in capability are less significant than relative reductions in cost. If both offence and defence become more sophisticated, the balance remains stable unless the cost structures shift unevenly. Artificial intelligence lowers specific categories of effort—time, labour, expertise, and iterative testing—associated with offensive cyber operations. Although defensive actors also benefit from automation, the structural constraints identified earlier limit the degree to which those benefits can counterbalance offensive acceleration.

One of the most consequential developments introduced by AI is the automation of analytical labour. Historically, cyber reconnaissance required extensive system mapping, code examination,

¹⁶⁷ Thomas Rid, *Cyber War Will Not Take Place* (Hurst 2013).

and behavioural analysis. Machine learning models excel at identifying patterns across vast datasets. When applied offensively, they can analyse code repositories, network metadata, and publicly available information at scales beyond manual capacity. The effect is not merely increased speed but reduced marginal cost. Expanding analysis to additional targets does not require proportional increases in personnel. Computational scaling substitutes for human scaling.

This development directly interacts with the architectural complexity of digital systems. Given that vulnerabilities are statistically inevitable in large-scale infrastructures, lowering the cost of discovering them produces disproportionate strategic effects. The defender's obligation remains comprehensive: all potential weaknesses must be addressed. The attacker's obligation is selective: a single viable point of entry is sufficient. AI reduces the cost of selectivity. As search becomes faster and less resource-intensive, the likelihood of identifying exploitable weaknesses rises without equivalent growth in expenditure. The asymmetry embedded in system design thereby becomes more readily exploitable.

Generative AI produces a comparable shift in the domain of social engineering. Cyber intrusions frequently rely not only on technical exploits but also on manipulation of human behaviour. Previously, a trade-off existed between scale and specificity: highly tailored deception required labour-intensive preparation, while mass campaigns sacrificed credibility. Generative language models collapse this trade-off by enabling contextually tailored communication at negligible marginal cost¹⁶⁸. Offensive actors can replicate institutional tone, incorporate contextual references, and adapt language to cultural settings without extensive human intervention.

The strategic consequence is enhanced scalability. When credible deception can be reproduced algorithmically, the volume of high-quality intrusion attempts increases. Defensive filtering systems confront dynamic variation rather than static templates. Machine-generated diversity challenges detection mechanisms reliant on known patterns. While defensive AI can analyse linguistic anomalies, it must minimise disruption to legitimate communication. Excessive sensitivity generates operational friction and economic cost. Offensive actors, by contrast, can tolerate failed attempts; unsuccessful experimentation carries limited penalty. This asymmetry in tolerance for error reinforces the broader asymmetry in cost reduction.

AI-driven iterative adaptation further widens the imbalance. Defensive cybersecurity often depends on identifying malicious signatures and deploying patches, processes historically marked by temporal lag. AI-enabled offensive systems can compress adaptation cycles by generating behavioural variants in response to detection. An offensive model can iteratively adjust until it identifies configurations that evade monitoring. Defensive responses, however, remain embedded within institutional processes of validation, coordination, and network-wide deployment. The obligation to maintain system stability constrains the speed of adaptation. Offence benefits from agility, whereas defence bears responsibility for continuity.

Lowered expertise thresholds also reshape the strategic landscape. AI-assisted coding and vulnerability analysis tools reduce technical barriers to conducting sophisticated cyber operations.

¹⁶⁸ Brandon Valeriano and Ryan C Maness, *Cyber War versus Cyber Realities: Cyber Conflict in the International System* (Oxford University Press 2015).

While skilled actors remain important, the pool of potential participants expands as marginal expertise requirements decline. Capability diffusion accelerates. More actors can attempt intrusion with less specialised training. Defensive institutions cannot equivalently reduce expertise thresholds without increasing systemic risk; effective security monitoring demands precision and oversight. The asymmetry between experimentation and reliability becomes more pronounced as AI tools proliferate.

Institutional context magnifies these dynamics. Offensive actors—particularly criminal organisations and loosely regulated entities—can deploy AI tools rapidly and without extensive oversight. They are not constrained by privacy regulations, procurement procedures, or public accountability. Defensive institutions, especially in democratic states and corporate settings, must integrate AI within regulatory and ethical frameworks. Concerns regarding transparency, bias, and legality mediate implementation. These governance requirements slow deployment and limit operational aggressiveness. Even where defensive AI offers technical advantages, its use is filtered through institutional friction.

It would be analytically unsound to claim that AI renders defence ineffective. Machine learning enhances anomaly detection, integrates threat intelligence across datasets, and supports predictive monitoring. In certain contexts, defensive automation significantly strengthens resilience. Nevertheless, these gains operate within a domain defined by comprehensive obligation. Defence must secure entire networks, ensure continuity of service, and minimise false positives that disrupt legitimate activity. Offence can concentrate on specific vulnerabilities, accept failure, and capitalise opportunistically on success. AI reduces the cost of opportunism more sharply than it reduces the burden of comprehensive protection.

The cumulative effect is a relative widening of offence–defence asymmetry in cyberspace. Both sides acquire enhanced tools, yet the structural characteristics of digital infrastructure enable offensive gains to translate more readily into scalable advantage. AI amplifies the selective logic of exploitation more effectively than it mitigates the systemic logic of protection.

The implications extend beyond technical cybersecurity. A perceived expansion of offensive advantage reshapes strategic incentives¹⁶⁹. States that interpret increased vulnerability may invest more heavily in offensive cyber capabilities, anticipating the need for pre-emption or retaliation. The belief that AI-enabled intrusion is becoming easier can stimulate competitive dynamics even where defensive capabilities improve concurrently. As offensive operations become cheaper and more scalable, the threshold for initiating limited cyber actions may decline, rendering strategic interaction more fluid and potentially more volatile.

Artificial intelligence thus operates as a multiplier of pre-existing imbalance. It does not transform cyberspace into an entirely new strategic domain, nor does it negate defensive innovation. Rather, it alters the marginal calculus of effort in ways that deepen structural asymmetry. Appreciating this

¹⁶⁹ Erik Gartzke, ‘The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth’ (2013) 38 *International Security* 41.

dynamic is essential for evaluating its broader consequences for escalation, deterrence, and strategic stability.

Section IV

Artificial Intelligence, Escalation Dynamics, and Strategic Stability

If artificial intelligence operates as an asymmetry multiplier in cyberspace, its relevance extends beyond questions of technical exposure. An imbalance between offence and defence matters because it reshapes strategic incentives. Classical security theory suggests that when offence is perceived to hold the advantage, actors experience greater pressure to adopt pre-emptive postures, to engage in competitive capability accumulation, and to interpret uncertainty as threatening. The effects are probabilistic rather than automatic: imbalance heightens the risk of instability. In the cyber domain, AI-induced shifts in relative advantage generate analogous concerns, albeit within a non-kinetic environment in which escalation manifests through disruption rather than territorial seizure.

Escalation in cyberspace differs in important respects from conventional military escalation. Cyber operations frequently remain below the threshold of armed conflict, encompassing espionage, intellectual property expropriation, infrastructure interference, and influence activities. Because such operations can be calibrated in intensity and obscured through ambiguous attribution, they enable incremental probing. Artificial intelligence reinforces this incremental dynamic by lowering the cost and increasing the speed of limited actions. As small-scale intrusions become cheaper to conduct, their frequency may increase, rendering strategic rivalry more persistent and less episodic.

A central concern lies in the compression of decision-making time. AI accelerates reconnaissance, exploit development, and adaptive evasion, thereby narrowing the window for detection and assessment. Defensive institutions may confront pressure to respond rapidly to ambiguous indicators¹⁷⁰. In contexts where attribution remains uncertain, accelerated response elevates the risk of miscalculation. The conjunction of speed and uncertainty is inherently destabilising. Strategic stability depends in part on deliberation—the time required to evaluate intent, determine proportionality, and coordinate response. AI-driven acceleration reduces this temporal buffer.

At the same time, the destabilising implications of speed require careful qualification. Cyber operations rarely generate immediate physical destruction, and many are reversible or containable. This has led some observers to argue that cyberspace possesses stabilising characteristics, permitting signalling without catastrophic escalation. Artificial intelligence complicates this assessment in two respects. First, by enhancing scalability, it expands the potential magnitude of disruption. Second, by diffusing capability across a broader array of actors, it increases the number of potential sources of instability. Stability grounded in restraint among major powers becomes more fragile when non-state actors can initiate cascading effects.

¹⁷⁰ Joseph S Nye, 'Deterrence and Dissuasion in Cyberspace' (2017) 41 *International Security* 44.

Deterrence in cyberspace is already characterised by ambiguity. Effective retaliation depends upon credible attribution and proportionate response. If AI widens offence–defence asymmetry, it may weaken deterrence by reinforcing perceptions that defensive systems are increasingly penetrable. States that perceive persistent vulnerability in their critical infrastructure, despite defensive investment, may pursue alternative strategies, including forward defence or expanded offensive postures. Such approaches risk normalising sustained cyber engagement rather than confining competition to discrete episodes.

Perception plays a decisive role in this process. Offence–defence theory emphasises that actors respond not only to objective capabilities but also to perceived advantage. If policymakers interpret AI-enhanced cyber capabilities as decisively favouring offence, they may prioritise offensive development accordingly. This perception can generate competitive spirals even when defensive capabilities improve in parallel. The security dilemma intensifies when defensive research into vulnerabilities is indistinguishable from offensive preparation. AI amplifies this ambiguity, as similar machine learning techniques may underpin both intrusion and detection.

Threshold ambiguity further complicates escalation dynamics. In conventional deterrence theory, escalation thresholds are often associated with physical destruction or territorial violation. In cyberspace, such thresholds are less clearly defined. AI-enabled operations may permit increasingly intrusive or disruptive activity while remaining below established markers of armed conflict¹⁷¹. The cumulative effect of sustained low-level intrusions may erode trust and heighten suspicion without provoking overt retaliation. An environment of continual probing fosters strategic fatigue and reactive positioning.

It would, however, be analytically reductive to conclude that widening asymmetry necessarily produces instability. Defensive adaptation is ongoing. As AI-enabled attacks grow more sophisticated, defensive institutions may reorganise, introduce redundancy, and cultivate more resilient architectures. Strategic stability does not require invulnerability; it requires predictability and credible signalling. Should major powers develop shared norms regarding acceptable cyber conduct, AI-enhanced capabilities may be integrated into managed competition rather than uncontrolled escalation.

Yet the normative dimension remains underdeveloped. International agreements governing cyber operations are limited and frequently non-binding. The pace of AI development exceeds diplomatic consensus. In the absence of clearly articulated rules, states interpret one another’s capabilities through worst-case assumptions, amplifying perceived imbalance. Uncertainty regarding AI’s future trajectory—its autonomy, adaptability, and scalability—contributes to strategic anxiety.

An additional source of instability arises from the interaction between cyber and conventional domains. As digital systems underpin military command, logistics, and communications, AI enhanced cyber intrusion into these networks could affect crisis management. If states believe their command-and-control infrastructure is vulnerable to AI-driven disruption, they may perceive

¹⁷¹ Miles Brundage and others, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (Future of Humanity Institute 2018).

incentives to act early in crises rather than risk incapacitation. This logic echoes concerns from nuclear strategy regarding vulnerability and pre-emption, although the modalities and stakes differ.

AI's influence on strategic stability thus operates through multiple interconnected pathways: acceleration of operational tempo, diffusion of capability, amplification of perception-driven insecurity, and deeper integration of digital infrastructure with critical systems. None of these dynamics makes escalation inevitable¹⁷². Each, however, increases the complexity of deterrence and the difficulty of sustaining equilibrium.

A widening offence–defence asymmetry does not necessarily foreshadow catastrophic cyber conflict. Rather, it points toward a strategic environment characterised by persistent contestation, compressed response windows, and heightened suspicion. Artificial intelligence intensifies the structural conditions under which miscalculation becomes more probable, particularly when combined with ambiguous attribution and indeterminate thresholds.

Ultimately, the consequences of these developments depend not solely on technological innovation but also on political interpretation and institutional adaptation. Strategic stability in an AI-enhanced cyber domain will hinge as much on norm development and governance as on technical capability. Nevertheless, if AI continues to reduce the marginal cost of offensive cyber operations more rapidly than it alleviates the systemic burden of defence, the underlying asymmetry will continue to exert pressure on established deterrence frameworks.

Section V

Rebalancing the Offence–Defence Equation in an AI-Enhanced Cyber Domain

If artificial intelligence deepens offence–defence asymmetry by lowering the marginal cost of offensive cyber operations more than that of systemic defence, then effective responses must address structural incentives rather than rely on isolated technical remedies. The challenge identified in this chapter is not merely the proliferation of AI-enabled threats, but the transformation of relative cost structures within cyberspace. Rebalancing therefore requires measures that raise the cost of offensive exploitation, reduce the burden of defensive maintenance, or recalibrate the strategic expectations that sustain competitive escalation.

The most foundational avenue of rebalancing lies in architectural reform. Cyberspace's offence bias is rooted in complexity and the retrospective addition of security to systems not designed with resilience as a primary objective. Artificial intelligence intensifies this condition by automating vulnerability discovery across expansive infrastructures. A structural response must therefore intervene at the design stage. Secure-by-design standards, mandatory code auditing, and liability regimes that sanction negligent software practices would elevate baseline security. By decreasing the density of exploitable vulnerabilities, such reforms increase the cost of offensive discovery

¹⁷² Lucas Kello, *The Virtual Weapon and International Order* (Yale University Press 2017).

even within an AI-augmented environment. The aim is not to eliminate vulnerability entirely, but to narrow the asymmetry generated by systemic fragility.

A second pathway involves redistributing the defensive burden. At present, defenders carry comprehensive responsibility for safeguarding complex networks, while attackers exploit isolated weaknesses. Collective defence mechanisms—shared threat intelligence platforms, synchronised patch management, and cross-sector response coordination—can reduce duplication and lower marginal defensive costs. AI-based detection systems perform more effectively when trained on aggregated datasets rather than siloed organisational logs. In this respect, cooperation functions as a defensive multiplier, counterbalancing the offensive multiplier effect produced by AI.

Yet technical coordination alone cannot rectify structural incentives. The persistence of asymmetry is reinforced by limitations in deterrence, particularly those stemming from attribution uncertainty. Improvements in forensic AI that enhance attribution reliability may strengthen deterrent credibility. Although perfect certainty is unattainable, reducing ambiguity alters strategic calculation. If offensive actors anticipate more consistent identification and coordinated response, the perceived cost of aggression increases. Enhanced attribution thus operates as an indirect mechanism for elevating offensive risk.

Legal and regulatory reform also has structural implications. Current market incentives in software production do not fully internalise security externalities. Clearer liability standards for severe security failures would shift costs from victims to producers, altering investment incentives. When firms face tangible repercussions for inadequate safeguards, preventive investment becomes economically rational. Over time, strengthened baseline resilience diminishes the exploitable surface available to AI-assisted reconnaissance.

International norm development addresses the perceptual dimension of asymmetry. Offence–defence theory underscores the influence of expectations on strategic behaviour. If states perceive AI-enhanced cyber capabilities as unconstrained and readily deployable, they are more likely to invest pre-emptively in offensive instruments. Norms delineating protected targets—such as critical civilian infrastructure—can stabilise expectations even amid technological change. While norms do not eliminate capability, they restrict its legitimate application, thereby moderating worst-case assumptions.

The governance of AI dissemination presents an additional structural consideration. Open distribution of advanced generative or code-producing models lowers expertise thresholds broadly, including for malicious actors. Conversely, excessive restriction may impede defensive innovation. A calibrated approach—incorporating safety testing, red-teaming protocols, and monitored deployment—can introduce friction against misuse without suppressing legitimate research. Such mechanisms moderate the pace at which offensive marginal costs decline.

Rebalancing does not necessitate the complete eradication of offensive advantage. In international politics, perfect equilibrium is rarely attainable. The objective is to prevent widening asymmetry from destabilising strategic expectations. Investments in resilience, redundancy, and rapid recovery capacity can diminish the strategic payoff of intrusion. Systems designed to degrade gradually rather than fail catastrophically reduce the expected benefit of attack, thereby lowering incentive even where technical penetration remains feasible.

Adaptation must also occur at the doctrinal level. Cyber defence should be integrated into comprehensive strategic planning rather than treated as a peripheral technical function. Clear articulation of red lines, proportional response frameworks, and escalation management mechanisms can mitigate ambiguity. As artificial intelligence compresses operational tempo, doctrine must compensate by clarifying response logic in advance. Predictable signalling reduces uncertainty and constrains escalation dynamics.

The central analytical insight is that AI-induced asymmetry is neither fixed nor inevitable. It arises from the interaction between technological capability and structural context. By reforming architecture, strengthening collective defence, enhancing attribution, realigning economic incentives, developing norms, and clarifying doctrine, actors can influence the trajectory of the offence–defence balance. Although the effectiveness of such measures will vary across political systems, their shared objective is to reshape relative marginal costs rather than merely respond to discrete threats. Rebalancing is therefore a process of structural adaptation. Artificial intelligence magnifies existing asymmetries; deliberate institutional innovation can mitigate that amplification. Whether such adaptation can keep pace with technological advancement remains uncertain. Absent structural intervention, however, the logic of asymmetry outlined in this chapter is likely to continue shaping the evolution of the cyber threat environment.

Conclusion

This chapter has examined whether artificial intelligence alters the offence–defence balance in cyberspace. Rather than approaching AI as a discrete technological breakthrough, the analysis has situated it within the enduring structural conditions that define the cyber domain. The central argument has been that cyberspace exhibited offence-leaning asymmetry prior to the diffusion of advanced AI systems, and that AI deepens this asymmetry by reducing the marginal cost of offensive action more substantially than it alleviates the systemic burden of defence. The structural baseline is therefore decisive. Digital infrastructures are complex, interdependent, and in constant evolution. Defenders carry comprehensive responsibility for securing systems in which vulnerabilities are statistically unavoidable. Attackers require only selective success. Attribution remains uncertain, deterrence constrained, and capability diffusion ongoing. Artificial intelligence enters this environment as a force multiplier. By automating reconnaissance, enabling scalable deception, accelerating adaptive exploitation, and lowering expertise thresholds, AI reinforces the selective logic of intrusion. Although defensive actors also benefit from automation, their gains are circumscribed by institutional constraints, reliability requirements, and the imperative of systemic continuity. The outcome is a widening of relative advantage rather than a balanced transformation. AI magnifies pre-existing imbalance because the underlying architecture of digital systems remains vulnerability-dense and economically misaligned. Rebalancing, therefore, does not entail suppressing AI development but reshaping the structural context within which it operates. Secure-by-design principles, collective defence coordination, improved attribution mechanisms, realigned economic liability, normative boundary-setting, and doctrinal clarification all operate upon the cost structures that underpin the offence–defence balance. By raising the expected cost or diminishing the anticipated payoff of offensive cyber operations, such measures can temper the asymmetry intensified by AI. The strategic trajectory of cyberspace will depend less on the intrinsic sophistication of AI systems than on the institutional adaptations that

accompany their integration. Offence–defence theory underscores the dynamic nature of balance: technological innovation may shift relative advantage, but governance, coordination, and perception can recalibrate it. Should states interpret AI-enhanced intrusion as an uncontested advantage and respond primarily through competitive offensive expansion, instability is likely to deepen. Conversely, if adaptation prioritises structural resilience and calibrated normative frameworks, asymmetry may be moderated even amid rapid technological change. Artificial intelligence does not convert cyberspace into an entirely new strategic arena. Rather, it sharpens and accelerates existing characteristics. It hastens processes that were previously slower, scales activities that were once limited, and diffuses capabilities that were formerly concentrated. Whether this sharpening results in sustained instability or in managed competition will depend on the extent to which political actors confront the structural foundations of imbalance rather than merely its technological expressions. The offence–defence balance in cyberspace is not a fixed conclusion but a continually evolving equation. AI has altered its variables. The eventual outcome will hinge on the effectiveness of institutional responses to that alteration.

-----*****-----