Cognitive Blind Spots in Security Frameworks: From Cybersecurity to AI Governance

Vsevolod Shabad
University of Liverpool
v.shabad@liverpool.ac.uk

Author's note (Updated version — 18 October 2025):
This revised version incorporates additional evidence and sources clarifying the cost estimation methodology for coordinated patch deployment (based on Ponemon Institute's 2019 *Costs and Consequences of Gaps in Vulnerability Response* study). It also introduces a UK comparative reference — the National Cyber Security Centre (NCSC) advisory on active exploitation of Oracle E-Business Suite — alongside existing FBI examples to illustrate cross-jurisdictional urgency in patch management. Minor editorial refinements and reference updates have been applied throughout.

Executive Summary

Cybersecurity, OT security, and AI governance frameworks share the same structural blind spots: they often embed cognitive biases, such as anchoring on outdated assumptions and overconfidence in compliance. These weaknesses are amplified by the mismatch between slow governance cycles, usually tied to annual budgets, and the rapid pace of adversary innovation. Evidence from WannaCry, SolarWinds, and recent OAuth token exploits suggests that these gaps are systemic and costly, arising not from technical failings but from the governance design itself.

As AI governance matures, the risks become even greater. Frameworks such as ISO 42001 and NIST AI RMF mirror cybersecurity's structures but face additional challenges: exponential capability growth, immature measurement practices, and a lack of real-time intelligence infrastructure. This paper argues that governance should be seen as a form of "cognitive infrastructure" and proposes three interventions: work-in-progress limits to counter temporal mismatches, an AI-specific threat intelligence ecosystem, and bias correction protocols embedded in standards. Together, these measures can reduce structural vulnerabilities and help policymakers, regulators, and enterprises avoid repeating cybersecurity's costly blind spots.

Abstract

Cybersecurity and artificial intelligence (AI) governance frameworks share a structural vulnerability: they embed systematic cognitive biases that amplify security failures. These biases are not limited to individual decision-making but are reproduced and reinforced by assessment-first governance approaches. Evidence from high-profile cybersecurity incidents — including WannaCry, SolarWinds, and recent FBI warnings on OAuth token exploitation — demonstrates how anchoring and overconfidence biases persist even in well-researched domains supported by mature intelligence-sharing ecosystems. Previous work on flow-constrained risk management for operational technology (OT) security demonstrated that

temporal mismatches between annual governance cycles and rapidly evolving threats can be structurally mitigated. In this paper, we extend that analysis to AI governance, where frameworks such as ISO 42001 and NIST AI RMF replicate cybersecurity's structural patterns but operate without an established threat intelligence infrastructure. Recent AI-related parliamentary hearings — including investigations into Clearview AI in Canada, US Senate hearings on deepfakes, and UK debates on live facial recognition — indicate that AI security failures are already entering the same oversight cycle as cybersecurity incidents. We propose three complementary interventions: dynamic resource allocation through work-in-progress limits, AI-specific threat intelligence infrastructure, and systematic bias correction protocols embedded in governance standards. Without these interventions, AI governance risks replicating the cognitive blind spots that have repeatedly undermined cybersecurity frameworks.

This version was updated (18 Oct 2025) to include additional UK-source analysis and expanded cost benchmarking.

Keywords: Cognitive bias, security governance, AI governance, regulatory oversight, standards, cybersecurity, operational technology.

Introduction

Cybersecurity governance frameworks have been refined through decades of international standardisation and real-world testing. Despite this maturity, they remain vulnerable to systematic failures that arise not from technical deficiencies or isolated human errors, but from structural properties that generate anchoring and overconfidence biases, leading to dangerous temporal mismatches between governance cycles and threat evolution.

High-profile incidents have repeatedly exposed these vulnerabilities. In 2018, the UK Public Accounts Committee interrogated NHS officials following the WannaCry ransomware outbreak. The inquiry revealed that NHS organisations had anchored on a 2014 Windows migration plan that remained incomplete by 2017, leaving approximately 5% of systems vulnerable. Despite Microsoft releasing a critical security patch 58 days before the incident, there was "no formal mechanism" to ensure compliance across NHS trusts. Despite Microsoft releasing a critical security patch 58 days before the incident, the Department had 'no formal mechanism for assessing whether local NHS organisations had complied' [1]. The NHS ultimately incurred ≈£92m in disruption and recovery costs [2]. Industry benchmarks indicate organisations spend an average of $1.4 million (≈£1.1m) annually on vulnerability management activities [3], suggesting that coordinated emergency patching across the NHS would have cost a small fraction of the £92m ultimately expended in uncontrolled crisis response.

Similarly, in 2021, the US Senate Intelligence Committee questioned SolarWinds executives after a months-long compromise of government and private networks. Executives emphasised compliance with industry standards, but attackers had bypassed defences undetected, illustrating overconfidence generated by process-oriented frameworks rather than outcome validation [4].

More recently, in September 2025, the FBI issued FLASH-20250912-001 documenting campaigns by threat groups UNC6040 and UNC6395 that exploited OAuth tokens to bypass

multi-factor authentication. Within days of detection, the FBI provided indicators of compromise, attack signatures, and mitigation guidance [5]. This intelligence allowed organisations to update defences almost immediately, highlighting the corrective role of real-time intelligence in challenging organisational assumptions. Comparable domestic guidance in the UK is issued by the National Cyber Security Centre (NCSC) — for example, its "Active exploitation of vulnerability affecting Oracle E-Business Suite" advisory [6] — illustrating that such high-priority patch directives are a standard component of national cyber-defence coordination.

These patterns reveal a fundamental governance problem: many organisations align framework iterations with annual budgetary and planning cycles, creating structural lags while adversaries adapt techniques weekly or daily. This temporal mismatch amplifies cognitive biases, leaving governance decisions anchored in assumptions that are already outdated by the time resources are deployed.

Prior research on flow-constrained risk management for operational technology systems [7] demonstrated that structural interventions can address these temporal mismatches through work-in-progress limits and enforced quarterly reassessment. However, while such interventions are necessary in cybersecurity — a domain with decades of accumulated data, globally coordinated threat intelligence, and widely adopted standards — the challenges are magnified in AI governance.

AI governance frameworks, including ISO 42001 [8] and NIST AI RMF [9], replicate cybersecurity's structural approaches while facing exponential capability evolution, immature measurement science, and a lack of equivalent intelligence infrastructure. Documented AI security incidents have increased substantially, with Adversa AI cataloguing 233 cases in 2025 compared to 149 in 2024 [10], including data leakage from enterprise applications, false-memory injection attacks against conversational agents, and systemic prompt injection vulnerabilities. These incidents reveal governance blind spots that current frameworks are ill-equipped to address.

Political institutions are already responding. In May 2022, Canada's House of Commons Standing Committee on Access to Information, Privacy, and Ethics investigated Clearview AI's collection of 3 billion facial images, resulting in a finding that the company had violated federal privacy laws [11]. The hearings revealed systematic gaps in the oversight mechanisms for AI. In December 2024, the US Senate Judiciary Committee's hearing 'Oversight of AI: Election Deepfakes' examined AI-generated content in political campaigns, with senators expressing concern about regulatory frameworks lagging behind technological capabilities [12]. UK debates scrutinised the deployment of live facial recognition in public spaces [13]. These early hearings serve as warning signals: AI security failures are attracting the same legislative scrutiny that followed the WannaCry and SolarWinds incidents.

This paper makes three contributions. First, we extend cognitive bias research beyond individual decision-making to framework-level vulnerabilities, demonstrating how structural designs amplify anchoring and overconfidence. Second, we analyse how these vulnerabilities are magnified in AI governance, given exponential capability trajectories and an absent corrective intelligence infrastructure. Third, we propose three infrastructure-level interventions that can mitigate these vulnerabilities before AI incidents escalate to a disaster scale.

Related Work

1.1 Cognitive Bias Foundations

The systematic study of cognitive biases originates with Tversky and Kahneman's pioneering work on heuristics [14], which demonstrated consistent deviations from rational decision-making. Researchers have since catalogued numerous biases — including optimism, confirmation, herding, hindsight, status quo, among many others — that influence judgment under uncertainty across behavioural economics, psychology, decision theory, and risk management.

For security governance frameworks, two biases are especially relevant. Anchoring bias refers to the excessive reliance on initial information or assumptions, even when subsequent evidence renders them outdated or incorrect. Overconfidence bias refers to the tendency to overestimate one's knowledge, capability, or control over uncertain environments. These biases are not only observed in individual decision-makers, but, as we argue, are structurally embedded and amplified by the design of governance frameworks.

1.2 Cognitive Bias in Cybersecurity

Early work on cognitive bias in cybersecurity focused on individual decision-makers. Tsohou et al. [15] examined how cognitive and cultural biases influence compliance with security policies, showing that standards often emphasise content delivery rather than behavioural internalisation. Barre et al. [16], in a systematic literature review, identified six biases affecting governance, including optimism-pessimism bias and herding, both distorting organisational security decisions. De Wit and Meyer [17] surveyed practitioners, finding that 56.6% believed they could estimate risk consequences without complete information — an empirical demonstration of systematic overconfidence.

These studies demonstrate systematic cognitive biases in individual security practitioners, but they raise a deeper question: if frameworks are designed to mitigate human error, why do these biases persist at organisational levels? The answer, we argue, lies in how framework structures themselves can institutionalise and amplify the very biases they are intended to correct. Rather than eliminating cognitive bias, assessment-first approaches may systematically embed anchoring and overconfidence into governance processes.

1.3 Empirical Evidence from Cybersecurity Incidents

This framework-level bias amplification becomes evident when examining institutional responses to major security incidents. Recent FBI intelligence further demonstrates temporal mismatches between adversary innovation and governance cadence [5]. Cybersecurity benefits from extensive real-time threat intelligence ecosystems, including CISA's Automated Indicator Sharing (AIS) programme and the Common Vulnerabilities and Exposures (CVE) database. These mechanisms provide external correctives, preventing frameworks from drifting too far from operational reality.

1.4 Emerging AI Governance Literature

In contrast, research on AI governance remains relatively nascent. ISO 42001 [8] and NIST AI RMF [9] codify risk management structures modelled on cybersecurity frameworks but rely heavily on processes that create anchoring points while lacking corrective feedback loops.

Empirical evidence of AI security failures is growing. Adversa AI [10] documented a 56.4% increase in reported incidents between 2024 and 2025 (from 149 to 233 cases), though this growth may reflect improved incident reporting as much as absolute increases in vulnerabilities. Their analysis reveals diverse attack vectors, including logic flaws in enterprise AI deployments, false-memory injection in conversational agents, and systemic prompt injection vulnerabilities — risk classes that would not have been captured by governance frameworks designed in late 2024. The OWASP Agentic AI Security report [18] emphasises that agentic AI systems — characterised by autonomy, memory, and self-modification — introduce fundamentally new threat surfaces incompatible with static governance models.

1.5 Positioning Our Contribution

Existing research identifies cognitive biases at individual levels but does not explain how frameworks function as bias multipliers. Institutional reports document failures but do not generalise mechanisms by which frameworks amplify biases. AI governance literature highlights emerging risks but has yet to integrate them with cognitive bias theory.

Our contribution bridges these gaps by extending cognitive bias analysis from individual behaviour to governance structures, demonstrating how these structures systematically produce biases in cybersecurity, and showing how the absence of a corrective intelligence infrastructure amplifies vulnerabilities in AI governance.

Cognitive Bias Patterns in Cybersecurity Governance

Despite decades of refinement, cybersecurity frameworks systematically amplify cognitive biases. Two patterns dominate — anchoring in risk assessments and overconfidence generated by compliance metrics — which together create systematic temporal mismatches between framework cycles and adversary innovation.

1.6 Anchoring in Risk Assessments

Anchoring bias arises when organisations fixate on initial assumptions or reference points, even when evidence shows they are outdated. Frameworks such as NIST CSF [19] and ISO 27001 [20] institutionalise this risk by requiring organisations to define scope, context, and objectives at the outset of the assessment. These decisions shape subsequent analysis, resource allocation, and implementation of controls.

The WannaCry ransomware outbreak exemplifies the phenomenon of structural anchoring. NHS organisations had anchored on a 2014 migration plan to phase out Windows XP systems. By May 2017, approximately 5% of systems — including critical medical devices

— remained on XP. Microsoft had released patches 58 days prior, yet anchoring on the original plan meant vulnerabilities were not re-evaluated in light of emerging evidence [1]. In testimony before the UK Public Accounts Committee [21], NHS officials continued citing the 2014 timeline as if it remained adequate. Anchoring transformed a roadmap into a blind spot.

The economic asymmetry between preventive and reactive patching is well-documented in enterprise security research. Ponemon Institute's 2019 study of vulnerability response across 2,900 organisations found that the average organisation spends approximately 23,000 staff hours annually on vulnerability management activities — equivalent to $1.4 million at standard IT security labour rates ($62.50/hour fully loaded) [3]. Critically, this study documented that 60% of breaches occurred from vulnerabilities where patches were available but not applied, and that organisations require an average of 16 days to patch critical vulnerabilities due to coordination failures across organisational silos [3]. The Department's 'no formal mechanism' for compliance verification [1] represents precisely this coordination gap that transforms manageable patch deployment into catastrophic system failure.

SolarWinds provides a second example. Risk assessments had anchored on vendor trust and supply chain assurance, leaving organisations unprepared for compromise of software updates [4]. Frameworks encouraged assessments against predefined supply chain categories rather than adaptive validation of evolving vendor practices. Anchoring on "trusted supplier" status prevented detection of anomalies that should have triggered reassessment.

1.7 Overconfidence Through Compliance

Overconfidence bias occurs when confidence in one's security posture exceeds objective justification. Frameworks amplify this by emphasising compliance with prescribed processes rather than validated outcomes.

During WannaCry hearings, NHS officials cited the number of on-site cyber assessments conducted — 88 out of 236 trusts — as evidence of diligence [21], indicating that 63% of trusts had not undergone independent assessments. Despite this gap, officials expressed confidence in system-wide resilience. Overconfidence arose from treating compliance activities as proxies for actual security.

Similarly, SolarWinds executives highlighted compliance with NIST standards during Senate testimony [4]. However, attackers maintained undetected access for months, exploiting blind spots that compliance checklists failed to capture. This dynamic reflects what we term "compliance theatre": the performance of security through documented adherence, producing overconfidence divorced from actual system integrity.

1.8 Temporal Mismatches Caused by Cognitive Bias

Temporal mismatches occur when governance cycles — typically aligned with annual budgeting and planning processes — misalign with the actual threat evolution pace. While frameworks such as NIST CSF or ISO 27001 do not prescribe annual reassessment, organisations commonly tie risk assessments and compliance reporting to yearly budgetary cycles, creating a structural lag as adversaries innovate on a weekly or daily basis.

Prior research on flow-constrained risk management for OT systems [7] addressed these temporal mismatches by imposing work-in-progress (WIP) limits. By restricting concurrent initiatives, organisations are forced to complete or abandon existing projects before starting new ones, creating natural reassessment points and resource reallocation opportunities while reducing anchoring on outdated priorities. Evidence from critical infrastructure demonstrates that such constraints improve responsiveness to evolving threats.

The September 2025 FBI FLASH [5] further illustrates temporal mismatches. Within weeks, adversaries had developed techniques to exploit OAuth tokens, rendering annual control reviews obsolete. Organisations constrained by yearly budgets and planning cycles were structurally unable to incorporate this threat on time. These temporal mismatches magnify anchoring and overconfidence effects: organisations remain committed to outdated assumptions and unjustified confidence, while adversaries exploit the resulting lags.

1.9 Synthesis

The two biases and their structural consequence are not independent. Anchoring on outdated baselines, overconfidence in compliance, and temporal mismatch mutually reinforce one another. The result is systematic governance failure. While individual decision-making biases have long been studied, the evidence from WannaCry, SolarWinds, and recent FBI intelligence shows that frameworks themselves are structural bias multipliers.

Amplified Risks in AI Governance

AI governance frameworks inherit the structural biases of cybersecurity but face additional amplifiers: exponential capability evolution, immature measurement science, absence of corrective intelligence infrastructure, and cascading complexity across value chains. Emerging parliamentary scrutiny of AI security failures signals that these vulnerabilities are no longer theoretical.

The structural biases we identify apply broadly across AI applications, from enterprise chatbots to autonomous systems, with agentic systems providing obvious illustrations of these vulnerabilities. However, we focus particularly on agentic AI systems as illustrative examples because they represent the current frontier of AI deployment and demonstrate these vulnerabilities most acutely. Agentic systems — characterised by autonomy, memory, and self-modification — amplify governance blind spots precisely because their capabilities evolve during deployment in ways that initial risk assessments cannot anticipate.

1.10 Structural Replication of Biases

ISO 42001 [8] and NIST AI RMF [9] replicate the structural approaches examined above, requiring organisations to define risk categories, intended purposes, and contexts of use at the outset. These scoping decisions create anchoring traps. Once categories are fixed, subsequent risks are evaluated through that lens, even when capabilities evolve.

This parallels the NHS anchoring on the 2014 migration schedules. In AI governance, an organisation might scope a model as "customer service chatbot" in January. By July, the same model may have gained autonomous integration with financial systems. Focusing on the initial scope prevents reassessment, creating blind spots.

1.11 Accelerated Temporal Mismatches

Whereas cybersecurity threats evolve monthly, AI capabilities evolve exponentially. Stanford's AI Index [22] reports that training compute doubles approximately every five months, while inference costs fall by orders of magnitude. This accelerates risk assessment obsolescence. Adversa AI [10] documented incidents, including logic flaws in enterprise applications and false-memory injection attacks against conversational systems. Risk assessments conducted in late 2024 would not have captured these vulnerability classes. Temporal mismatches are therefore magnified: governance cycles tied to annual budgeting are structurally incapable of keeping pace with the acceleration of AI.

1.12 Measurement Immaturity and Overconfidence

Cybersecurity benefits from mature measurement science: vulnerability scoring (CVSS [23]), decades of incident data, and outcome-focused metrics. AI governance lacks comparable foundations. NIST AI RMF acknowledges the absence of validated benchmarks and recommends stakeholder consultation as an alternative to empirical thresholds [9].

This gap fosters dangerous overconfidence. Organisations generate risk registers with false precision, assuming that compliance with ISO 42001 or NIST AI RMF equates to security. Adversa AI's data show that 35% of AI incidents stem from prompt injection, a risk class not easily captured by current frameworks. The OWASP Agentic AI Security report [18] highlights that agentic systems disrupt assumptions of fixed functionality, further invalidating traditional metrics.

The result is dangerous: overconfidence is amplified precisely when uncertainty is most tremendous.

1.13 Absence of Corrective Intelligence Infrastructure

In cybersecurity, external intelligence ecosystems challenge organisational assumptions. CISA’s AIS programme, the CVE database, and vendor threat feeds provide continuous correction. The September 2025 FBI FLASH forced organisations to reassess confidence in MFA within days [5].

AI governance lacks such mechanisms. Existing incident repositories — the AI Incident Database, AVID, and AIAAIC — focus on retrospective cataloguing, often relying on voluntary submissions or media reports. They lack real-time, machine-readable feeds of indicators or vulnerabilities. Consequently, anchoring and overconfidence persist unchecked.

The recently published OWASP AIVSS framework [24] illustrates this bias amplification precisely. While pretending to mathematical rigour through detailed scoring methodologies, its core Agentic AI Risk Score relies entirely on unvalidated expert judgment without bias mitigation controls. This transforms subjective assessments into seemingly objective numerical scores that organisations will treat as authoritative, exemplifying how frameworks can institutionalise rather than correct cognitive biases.

The intelligence vacuum is structural. Without corrective feedback loops, AI frameworks drift further from operational reality with each governance cycle.

1.14 Cascading Complexity and Community Recognition

AI systems integrate across complex value chains, creating cascading failure risks that the OWASP community has proactively identified through their Agentic Security Initiative, including ASI08 Cascading Failures documentation [25]. However, even expert technical communities face temporal mismatches — their comprehensive risk assessments remain in draft form while AI capabilities evolve exponentially, illustrating how governance development cycles cannot keep pace with the rapid technological change.

1.15 Political Early Warnings and Regulatory Responses

Political institutions have begun recognising these governance challenges. Recent parliamentary hearings — including Canada's 2022 investigation into Clearview AI's privacy violations [11], the US Senate's scrutiny of election deepfakes [12], and the UK's debates on live facial recognition [13] — demonstrate that AI security failures are attracting the same legislative attention that historically followed major cybersecurity incidents. The EU AI Act and its Code of Practice represent the most comprehensive regulatory attempts, yet even these prescriptive frameworks face the structural bias challenges identified above. The UK's principles-based approach offers a contrasting regulatory philosophy, but both approaches must address the fundamental problem of governance cycles misaligned with AI evolution rates.

1.16 Early Regulatory Recognition and Intervention Opportunities

Parliamentary scrutiny in both the UK and EU demonstrates that AI governance failures are already attracting the same legislative attention that historically followed cybersecurity incidents. Canada's 2022 House of Commons investigation into Clearview AI's privacy violations [11], the US Senate's 2024 hearings on election deepfakes [12], and the UK's parliamentary debates on live facial recognition [13] reveal systematic gaps in oversight mechanisms that current frameworks cannot address.

The EU AI Act's Articles 9 and 10 mandate systematic risk management and bias detection protocols [26], yet these very requirements may embed the anchoring and overconfidence patterns we identify. Article 27's Fundamental Rights Impact Assessment requires updating "when any of the elements has changed," creating natural reassessment triggers that could mitigate temporal mismatches — if implemented without succumbing to the same cognitive traps.

The UK's emerging AI governance framework [27] faces parallel challenges. The AI White Paper's principle-based approach delegates risk assessment to sector regulators, potentially multiplying anchoring effects across different regulatory contexts. The pattern emerging from parliamentary hearings suggests UK regulators will face similar scrutiny: the same systematic bias amplification that led to WannaCry and SolarWinds inquiries now threatens AI governance.

Both jurisdictions' Codes of Practice mechanisms — Article 56 in the EU AI Act and the UK's forthcoming regulatory guidance — represent critical intervention points. These parliamentary warning signals indicate that a cognitive bias-resistant framework design is not

merely an academic theory, but an urgent practical necessity for avoiding the next generation of governance failures.

The implication is clear: AI security failures are no longer hypothetical. They are politically salient, socially consequential, and structurally magnified by governance frameworks.

1.17 Synthesis

AI governance inherits cybersecurity's structural biases but faces four amplifiers: exponential capability evolution, measurement immaturity, lack of corrective intelligence, and cascading interdependencies. The result is an environment where anchoring, overconfidence, and temporal mismatches are not only replicated but intensified. Parliamentary hearings demonstrate that these vulnerabilities are already attracting political oversight. Without intervention, AI governance will face failures of scale comparable to, or exceeding, WannaCry and SolarWinds.

Mitigation Strategies

The preceding analysis demonstrates that anchoring and overconfidence biases are structurally embedded in governance frameworks, creating systematic temporal mismatches. Cybersecurity has partially mitigated these through external intelligence feeds and adaptive practices, but AI governance lacks equivalent infrastructure. We propose three complementary strategies: dynamic resource allocation through work-in-progress limits, AI-specific threat intelligence infrastructure, and systematic bias correction protocols embedded in standards.

1.18 Dynamic Resource Allocation Through Work-in-Progress Limits

One mechanism for countering temporal mismatches restricts concurrent security initiatives through work-in-progress (WIP) limits. Borrowed from lean manufacturing and Kanban systems [28], WIP limits prevent resource entrenchment on outdated priorities by forcing regular reassessment.

Earlier work on flow-constrained OT security management [7] demonstrated the effectiveness of this approach. In critical infrastructure environments, WIP limits required organisations to timely complete or abandon existing initiatives before initiating new ones, producing natural checkpoints for threat reassessment. Empirical validation showed that shorter project cycles with enforced reassessment improved responsiveness to evolving threats.

Applied to AI governance, WIP limits could operate at both organisational and regulatory levels. Organisations could restrict the simultaneous activation of AI risk assessments or mitigation projects to fixed thresholds (e.g., no more than five concurrent high-priority controls). When novel risks emerge, such as new prompt injection vectors, managers should either reprioritise existing efforts or defer new projects. This forces an explicit reassessment of priorities rather than continued reliance on outdated risk registers.

At standards levels, regulators could encourage or mandate quarterly reassessment cycles tied to WIP limits. Instead of relying solely on budget-driven annual governance cycles,

organisations would demonstrate that resources had been reallocated to address newly identified risks at least quarterly. This would directly counter temporal mismatches, breaking cycles of anchoring on outdated baselines.

1.19 AI-Specific Threat Intelligence Infrastructure

Absent AI-specific threat intelligence structurally amplifies bias. Cybersecurity benefits from machine-readable feeds such as CISA's Automated Indicator Sharing (AIS) and the CVE programme, which provide continuous external validation of assumptions. In contrast, AI incident repositories are retrospective, voluntary, and qualitative.

We propose creating an international AI threat intelligence infrastructure with three core components:

1. **Machine-Readable Indicators of AI Compromise (IAOCs).** Analogous to indicators of compromise in cybersecurity, though adapted for AI-specific exploitation patterns, IAOCs would document technical artefacts of AI exploitation: malicious prompt patterns, model weights associated with emergent vulnerabilities, or adversarial input sequences. These should be published in structured formats (e.g., new extensions to existing STIX standards, adapted for AI-specific indicators) for automated ingestion.
2. **AI Vulnerability Disclosure Programme.** A standardised process for reporting and validating AI vulnerabilities is required, modelled on CVE. Researchers, vendors, and regulators would contribute verified entries to central repositories. Each vulnerability would be assigned severity scores analogous to CVSS [23], enabling prioritisation..
3. **Exploitability-Based Prioritisation.** AI threat intelligence should distinguish between theoretical risks (e.g., potential model inversion) and actively exploited vulnerabilities (e.g., real-world prompt injection campaigns). This parallels how cyber intelligence moved from cataloguing all vulnerabilities to prioritising based on live exploit data.

The AI Standards Hub [29] and OWASP [18] are well-positioned to coordinate early pilots of such infrastructure, leveraging their neutrality and international reach. Governments could mandate participation, as the European Union has done with cybersecurity incident reporting under the NIS2 directive [30].

Establishing an AI-specific intelligence infrastructure would not eliminate bias, but it would create corrective quick feedback loops. External data would challenge anchoring; overconfidence would be tempered by real-time evidence of exploitation; temporal mismatches would be reduced by continuous updates.

1.20 Systematic Bias Correction Protocols

The third intervention targets the frameworks themselves. ISO 42001[8] and NIST AI RMF [9] replicate cybersecurity's structural approaches, embedding anchoring and overconfidence by design. Unless corrected at standards levels, these frameworks will perpetuate structural vulnerabilities.

We propose embedding mandatory bias correction protocols into governance standards:

- **Periodic Assumption Challenges.** Organisations should be required to justify continued reliance on initial risk assessments at regular intervals. Independent review boards or red-team exercises could provide an external challenge.
- **Empirical Validation Requirements.** Compliance claims should be substantiated by evidence of security outcomes. For example, organisations claiming resilience against prompt injection should provide empirical testing results, not only documentation of process adherence.
- **Automated Reassessment Triggers.** Standards should specify conditions that trigger mandatory reassessment, such as model scale exceeding computational thresholds, deployment into new domains, or publication of verified vulnerabilities. These triggers would ensure that governance cycles remain aligned with the evolution of capabilities.
- **Bias-Aware Language in Standards.** While standards bodies prefer process-neutral phrasing, requirements could be framed as "systematic risk validation" or "assumption verification protocols." This would integrate bias correction without undermining the voluntary adoption of these measures.

These measures align with OWASP's [18] call for adaptive, continuous oversight of agentic AI systems. By embedding bias correction at standards levels, governance frameworks would move beyond awareness of cognitive limitations to systematic mitigation.

1.21 Integration of Strategies

The three strategies address different but complementary aspects of the bias problem. WIP limits mitigate temporal mismatches. Threat intelligence infrastructure provides corrective feedback against anchoring and overconfidence. Bias correction protocols embed resilience at framework levels.

Together, they form a multi-layered approach reducing systematic vulnerabilities while preserving the governance framework’s utility. Importantly, these strategies are mutually reinforcing: intelligence feeds inform WIP reprioritisation; reassessment protocols ensure new intelligence is operationalised; and WIP constraints prevent resource entrenchment.

Conclusion

Cybersecurity governance frameworks have repeatedly failed to prevent catastrophic incidents despite decades of refinement. Analysis of WannaCry, SolarWinds, and the September 2025 FBI FLASH demonstrates that structural approaches systematically amplify anchoring and overconfidence biases, creating dangerous temporal mismatches. Parliamentary hearings in the UK and US confirm that these failures are politically salient, requiring legislative scrutiny.

AI governance inherits these vulnerabilities but faces amplified risks: exponential capability evolution, immature measurement science, absent corrective intelligence infrastructure, and cascading dependencies across value chains. Recent political hearings on Clearview AI, deepfakes, and live facial recognition indicate that AI security failures are already entering the same oversight cycle as cybersecurity.

Our contribution is threefold. First, we extend cognitive bias analysis from individual decision-making to governance frameworks, demonstrating that frameworks themselves act as bias multipliers. Second, we show how AI governance amplifies these vulnerabilities, making current frameworks systematically inadequate. Third, we propose three interventions — WIP limits, AI-specific threat intelligence infrastructure, and embedded bias correction protocols — that address anchoring and overconfidence biases and their resulting temporal mismatches at structural levels.

The broader implication is that governance should be understood not merely as a technical or organisational challenge but as a form of cognitive infrastructure. Without corrective interventions, AI governance risks reproducing cybersecurity's blind spots on even larger scales, generating failures comparable to, or exceeding, those seen in past crises — including a preventive-to-reactive cost imbalance conservatively estimated at roughly 46:1 for WannaCry and the systemic compromise exemplified by SolarWinds. With such interventions, governance can evolve into adaptive, intelligence-driven systems capable of managing rapidly advancing technologies.

Future research should test these interventions empirically. Pilot programmes could measure WIP limits' effects on AI project prioritisation, simulate machine-readable AI threat intelligence feeds, and evaluate automated reassessment triggers' effectiveness in live deployments. These studies would provide the evidence base needed for embedding bias correction into international standards.

In conclusion, the lesson from cybersecurity is clear: governance frameworks are not neutral safeguards but cognitive infrastructures that can amplify or mitigate bias. As AI capabilities accelerate, addressing these blind spots becomes not just technically urgent but a matter of public interest. The development of cognitive-bias-resistant frameworks serves clear public interest objectives for AI safety and security. Parliamentary hearings across multiple jurisdictions demonstrate that AI governance failures attract political scrutiny precisely because they threaten societal welfare. Cognitive biases in governance frameworks are not merely academic concerns but structural vulnerabilities that can undermine public trust in AI systems and compromise collective security. By redesigning governance frameworks with bias resilience in mind, we can reduce the risk that tomorrow's AI crises become the subject of the next parliamentary hearing.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The author, being a non-native English speaker, used Claude (Anthropic) for language refinement, structural organisation of arguments, and literature synthesis assistance. Specific sections where AI assistance was used include portions of the Related Work section, transition sentences throughout, and overall readability improvements. All substantive intellectual contributions, research design, analysis, and conclusions remain entirely the work of the author.

Data Availability Statement

Due to the sensitive nature of critical infrastructure security implementation and confidentiality agreements with the participating organisation, raw data cannot be made publicly available. General implementation insights and anonymised methodological details are available from the corresponding author upon reasonable request.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. National Audit Office, "Investigation: WannaCry cyber attack and the NHS," HC 414, NAO, London, 2018. [Online]. Available: https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS.pdf

2. Department of Health & Social Care, "Securing cyber resilience in health and care. Progress update October 2018," DHSC, London, 2018. [Online]. Available: https://assets.publishing.service.gov.uk/media/5bbe1250ed915d732b99254c/securing-cyber-resilience-in-health-and-care-september-2018-update.pdf

3. Ponemon Institute, "Costs And Consequences Of Gaps In Vulnerability Response," Ponemon Institute, 2019. [Online]. Available: https://www.servicenow.com/premium/resource-center/analyst-report/ponemon-vulnerability-survey.html

4. U.S. Senate Select Committee on Intelligence, "Open Hearing: Hack of U.S. Networks by a Foreign Adversary", Washington, DC, USA, Feb. 23, 2021. [Online]. Available: https://www.intelligence.senate.gov/wp-content/uploads/2024/08/sites-default-files-hearings-chrg-117shrg45485.pdf

5. Federal Bureau of Investigation, "Cyber Criminal Groups UNC6040 and UNC6395 Compromising Salesforce Instances for Data Theft and Extortion," FBI, Washington, DC, USA, 2025. [Online]. Available: https://www.ic3.gov/CSA/2025/250912.pdf

6. National Cyber Security Centre, "Active exploitation of vulnerability affecting Oracle E-Business Suite," NCSC, London, 2025. [Online]. Available:

https://www.ncsc.gov.uk/news/active-exploitation-vulnerability-affecting-oracle-ebusiness-suite

7. V. Shabad, “Flow-Constrained Risk Management for Operational Technology Security: A Multi-Criteria Framework for Critical Infrastructure“, SSRN Working Paper, 5389934, 2025. [Online]. Available: https://dx.doi.org/10.2139/ssrn.5389934.

8. ISO / IEC, "Information technology — Artificial intelligence — Management system (ISO/IEC 42001:2023)”, Geneva, 2023. [Online]. Available: https://www.iso.org/standard/42001.

9. National Institute of Standards and Technology, "AI Risk Management Framework," NIST, Gaithersburg, MD, 2023. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

10. Adversa AI, "Top AI Security Incidents: 2025 Edition," Adversa AI, 2025. [Online]. Available: https://adversa.ai/top-ai-security-incidents-report-2025-edition/

11. Parliament of Canada, House of Commons Standing Committee on Access to Information, Privacy and Ethics, "Evidence — Meeting No. 18: Study of the Use and Impact of Facial Recognition Technology”, May 2, 2022. [Online]. Available: https://www.ourcommons.ca/DocumentViewer/en/44-1/ETHI/meeting-18/evidence

12. U.S. Senate Committee on the Judiciary, "Oversight of AI: Election Deepfakes”, 2024, Rep., [Online]. Available: https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-election-deepfakes

13. UK Parliament, House of Commons Library, "Police use of live facial recognition technology. Debate Pack (CDP-2024-0144)”, 2024, Rep., 11 November. [Online]. Available: https://researchbriefings.files.parliament.uk/documents/CDP-2024-0144/CDP-2024-0144.pdf

14. A. Tversky and D. Kahneman, “Availability: A heuristic for judging frequency and probability,” *Cognitive Psychology*, vol. 5, no. 2, 1973, pp. 179–195.

15. A. Tsohou, et al., “Analyzing the role of cognitive and cultural biases in the internalization of information security policies: Recommendations for information security awareness programs,” *Computers & Security*, vol. 52, 2015, pp. 128–141.

16. G. Barre, et al., “Towards Understanding Cognitive Biases in Cybersecurity Governance,” *Proc. 38th Bled eConference*, Bled, Slovenia, 2025, pp. 737–744.

17. J. de Wit and C. Meyer, “Uncovering Cognitive Biases in Security Decision Making,” *Security Management Magazine*, 2022.

18. K. Underkoffler, et al., "State of Agentic AI Security and Governance 1.0," OWASP Foundation, May 2025. [Online]. Available: https://genai.owasp.org/resource/state-of-agentic-ai-security-and-governance-1-0/

19. National Institute of Standards and Technology, "The NIST Cybersecurity Framework (CSF) 2.0", Gaithersburg, MD, USA, 2024. [Online]. Available: https://doi.org/10.6028/NIST.CSWP.29.

20. ISO / IEC, "Information security, cybersecurity and privacy protection — Information security management systems — Requirements (ISO/IEC 27001:2022)", Geneva, 2022. [Online]. Available: https://www.iso.org/standard/27001.

21. UK Parliament, House of Commons, Public Accounts Committee, "Oral evidence: Cyber-attack on the NHS," House of Commons, London, UK, 2018. [Online]. Available: https://committees.parliament.uk/oralevidence/10786/pdf/

22. N. Maslej, et al., "Artificial Intelligence Index Report 2025," Institute for Human-Centered AI, Stanford University, Stanford, CA, USA, 2025. [Online]. Available: https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf

23. FIRST.Org, "Common Vulnerability Scoring System version 4.0: Specification Document," FIRST.Org, Inc., Cary, NC, 2023. [Online]. Available: https://www.first.org/cvss/specification-document

24. K. Huang, et al., "AIVSS Scoring System For OWASP Agentic AI Core Security Risks v0.5," OWASP Foundation, 2025. [Online]. Available: https://aivss.owasp.org/assets/publications/AIVSS%20Scoring%20System%20For%20OWASP%20Agentic%20AI%20Core%20Security%20Risks%20v0.5.pdf

25. OWASP Agentic Security Initiative, "ASI08 Cascading Failures," OWASP Foundation, [Online]. Available: https://github.com/OWASP/www-project-top-10-for-large-language-model-applications/blob/main/initiatives/agent_security_initiative/agentic-top-10/Sprint%201-first-public-draft-expanded/ASI08_Cascading_Failures%20.md. [Accessed: 18 Sep 2025].

26. European Union, "Regulation (EU) 2024/1689 (Artificial Intelligence Act)", *Official Journal of the European Union*, vol. 2024/1689, pp. 2024

27. UK Government, "A pro-innovation approach to AI regulation: White Paper", London, 2023, Rep. 978-1-5286-4009-1, 2023/08/03. [Online]. Available: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

28. D.J. Anderson, *Kanban: Successful evolutionary change for your technology business*, Blue Hole Press, 2010.

29. AI Standards Hub, "AI Standards Hub," 14 Sep 2025, https://aistandardshub.org/.

30. European Union, "Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)", *Official Journal of the European Union*, vol. 333, pp. 80–152, 2022

Vitae

Vsevolod Shabad, CISSP, CCSP, is a Principal Enterprise Architect at BT Group and a Fellow of BCS, The Chartered Institute for IT. He has over two decades of international experience in cybersecurity and enterprise architecture, including senior leadership roles in the financial services and critical infrastructure sectors. He holds an MEng in Applied Mathematics and a PGDip in Information Security and is currently pursuing an MSc in Cybersecurity at the University of Liverpool. His research focuses on how cognitive biases and temporal misalignments affect governance frameworks across cybersecurity, operational technology, and AI. He has published preprints on SSRN and is preparing journal submissions while contributing to UK and EU policy discussions on AI and security governance.