EUROPEAN LEADERSHIP NETWORK

Federal Foreign Office

Towards a better understanding of human bias in nuclear decision-making and its interaction with emerging & disruptive technologies

Workshop report

Ganna Pogrebna
Rishi Paul

February 2026

The European Leadership Network (ELN) is an independent, non-partisan, pan-European network of over 450 past, present and future European leaders working to provide practical real-world solutions to political and security challenges.

About the 'Simulating Technological Complexity & Advancing Risk Reduction' project

This project focuses on a fast-changing, yet neglected, area of nuclear risk: mitigating the impacts of emerging and disruptive technologies (EDTs) on nuclear weapons decision-making and nuclear command, control, and communications (NC3).

The fundamental aim of this project is risk reduction. We aim to do this by assisting states in identifying and mitigating nuclear-use pathways and potential mistakes / miscalculations generated by the aggregate effects of EDTs on nuclear weapons decision-making processes and NC3 systems.

The 'Simulating Technological Complexity & Advancing Risk Reduction' project consists of three strands of work that seek to better understand EDTs and technological complexity, to produce detailed recommendations to mitigate nuclear risks:

1. Strand 1 centres on developing a Guardrails and Self-Assessment (GSA) Framework to address anticipated challenges that EDTs pose to NC3 systems and nuclear decision-making.[1] Focusing on technologies likely to mature over the next five to ten years, the framework offers a predictive snapshot grounded in informed assumptions about how these systems may interact and collectively shape the nuclear landscape.

   The GSA Framework examines the combined impact of six key EDTs — artificial intelligence, autonomous systems and drones, counter-space capabilities, deepfakes, cyber operations, and quantum technologies — assessing their cumulative effects and the added complexity they bring to nuclear decision-making processes.

2. Through Strand 2, the ELN seeks to create a prototype digital tool that will simulate the highest-level nuclear weapons decision-making instances, the aggregate impact of EDTs in these processes, and the way the framework developed in Strand 1 can mitigate the risks generated by the aggregate effects of EDTs. This workshop report is part of Strand 2.

3. Strand 3 will develop a sustained campaign to implement the recommendations on EDTs and technological complexity risk reduction among nuclear-weapon and non-nuclear-weapon states and throughout multilateral and supra-national groups, such as the Nuclear Non-Proliferation Treaty review cycle, the Creating the Environment for Nuclear Disarmament initiative, the Stockholm Initiative, and NATO, among others.

This work was performed with the generous support of the German Federal Foreign Office. The views and opinions of authors expressed herein do not necessarily state or reflect those of the German Federal Government.

About the Authors

Professor Ganna Pogrebna
Behavioural Data Scientist

Professor Ganna Pogrebna is an internationally recognised expert in behavioural data science, AI governance, and emerging technology risk assessment. Her research focuses on decision-making under uncertainty and the behavioural, ethical, and organisational dimensions of AI, with particular emphasis on high-stakes domains such as defence, cybersecurity, and national security policy. She has published over 150 peer-reviewed research articles, 3 books, is Editor of the forthcoming Cambridge Handbook of Behavioural Data Science, and a contributor to The Oxford Handbook of the Ethics of AI. She has led large-scale interdisciplinary research programmes with total funding exceeding AUD 30 million, including projects for the World Bank, the UK Ministry of Defence, GCHQ, and the Australian Office of National Intelligence.

Her work has been recognised through multiple honours, including the TechWomen100 Award, the Women in AI APAC Award (Risk Modelling and Cybersecurity), and the Australian Women in Security Award for AI in Cybersecurity. She serves as Methods Editor for The Leadership Quarterly and sits on the editorial boards of Scientific Reports and Judgment and Decision Making. She regularly advises international organisations, governments and industry on behavioural risk, AI governance, and technology-enabled decision systems.

Dr Rishi Paul
Senior Policy Fellow, European Leadership Network

Dr Rishi Paul is a Senior Policy Fellow at the European Leadership Network (ELN), where he leads the organisation's work on nuclear deterrence, with a focus on how cognitive bias, risk perception, and decision-making dynamics shape escalation and nuclear risk. His research examines adversary escalation logics and the risks arising from misperception, misjudged risk tolerance, and cross-domain pressures, including those linking the Euro-Atlantic and Indo-Pacific theatres.

Rishi's work integrates behavioural insights into the study of nuclear deterrence and strategic stability, with the aim of improving how governments assess escalation risk under conditions of uncertainty. In recent years, he has led projects examining the cumulative effects of emerging and disruptive technologies (EDTs) on nuclear decision-making, including the use of AI-enabled modelling and the development of a baseline digital twin to support structured analysis of crisis dynamics. He is also a co-author of the Guardrails and Self-Assessment (GSA) Framework, designed to help policymakers identify and manage nuclear risks associated with technological change.

Prior to joining ELN, Rishi was a Policy Fellow and Programme Manager at the British American Security Information Council (BASIC). His work has received international recognition, including First Prize in the Geneva Centre for Security Policy's global security competition in 2022. He holds an MA in Strategic Studies and a PhD from the University of Leeds.

Executive summary

Bias is a critical yet under-addressed factor in nuclear decision-making. Cognitive shortcuts become particularly problematic in crises, where uncertainty, incomplete information, and the gravity of potential consequences can distort judgment. Historical cases such as the Cuban Missile Crisis illustrate how misperception and rigid assumptions have brought nuclear-armed states close to catastrophe. Today, the challenge is compounded by the growing role of AI and automated decision-support systems, which promise speed and precision but also introduce new forms of bias that can influence human reasoning and strategic choices.

This report presents findings from an ELN workshop that examined the 'human' and 'machine' components of bias and their points of interaction. Through structured discussions and a prototype digital twin simulation, designed to replicate key features of nuclear crisis decision-making, including uncertain information flows and AI-assisted inputs, participants drawn from the ELN Young Generation Leadership Network (YGLN), including military advisors, behavioural scientists, and technology developers, reflected on how group dynamics, ambiguity, and machine-generated outputs shape judgment under pressure.

The workshop did not present AI or automated decision-support systems as passive or neutral inputs to crisis decision-making. Instead, these tools were deliberately embedded in the simulation environment as active elements designed to provoke reflection on how machine-generated outputs might shape human judgment under stress. The goal was not to evaluate the technical accuracy of a particular model, but to explore how the presence of an ostensibly intelligent system could influence group dynamics, the perceived credibility of information, and the confidence underpinning high-stakes decisions.

While not prescriptive, the report highlights how human judgment and AI systems can interact in ways that reinforce, rather than reduce, risk. It is intended to inform policymakers and stakeholders by surfacing these dynamics at the intersection of nuclear weapons and emerging disruptive technologies, with the goal of deepening understanding and supporting more informed debate.

"This report highlights how human judgment and AI systems can interact in ways that reinforce, rather than reduce, risk."

Key workshop insights included:

Human biases: Seven recurrent biases were identified as particularly relevant to nuclear crises: *illusion of control, inherent bad faith, peaceful defensive images, perceptual bias, interpersonal bias, overconfidence, and worst-case thinking.* Each was shown to distort strategic reasoning and escalation choices in distinctive ways.

Technology as a bias modulator: AI was neither purely corrective nor purely distorting. Instead, it functioned as a 'bias modulator', sometimes reinforcing overconfidence and premature certainty, while at other times exposing hidden assumptions or broadening consideration of alternatives. Trust in AI was shown to be highly selective, embraced when reinforcing existing views, dismissed when it contradicted them.

Group dynamics: Authority bias, groupthink, and sycophancy were recognised as pervasive, with participants noting how hierarchical cultures can silence dissent. Decision outcomes were often shaped more by those who spoke with confidence than by data.

Design principles for bias mitigation: The workshop highlighted that effective nuclear-related decision environments must be designed to expose bias rather than conceal it.

Trust in AI was shown to be highly selective, embraced when reinforcing existing views, dismissed when it contradicted them.

Introduction: Why bias matters in nuclear decision-making

Nuclear decision-making lies at the intersection of political authority, military capability, intelligence interpretation, and now increasingly, algorithmic mediation.[2]

Across all these domains, one enduring and under-addressed issue persists: the systematic distortions of judgment caused by cognitive and structural biases.[3] As behavioural science has shown, biases are not merely random mistakes; they are predictable, systematic tendencies in human reasoning that can significantly affect outcomes, particularly under pressure.[4]

In high-stakes, ambiguous environments like nuclear crises, biases are amplified rather than diminished. Cognitive overload, time pressure, fear, and asymmetric information all create the conditions under which heuristics – mental shortcuts developed for survival – override deliberative thinking.[5] This phenomenon is not new. Cold War crises, including the Cuban Missile Crisis and Able Archer '83, reveal how close nuclear weapon states have come to catastrophe due to faulty interpretation, misperception, or rigid strategic assumptions. What has changed, however, is the technological environment.

AI-based early warning systems, predictive analytics, and autonomous military platforms all introduce new variables into the already volatile mix. While these systems promise speed and precision, they also bring with them their own embedded biases, arising from training data, model architecture, optimisation criteria, and human-machine interaction patterns. In this context: "no model fully captures reality – even the most advanced AI remains an approximation, with its utility dependent on practical application."[6] As AI tools move from advisory to operational roles in national security infrastructure, the interface between human bias and technological artefact must become a central concern.[7]

The workshop examined the 'human' and 'machine' components of bias, focusing on their distinct characteristics and points of interaction. Participants were invited to consider how group dynamics, interface design, strategic framing, and the outputs of AI and other emerging technologies can reinforce one another to produce compounding error. This discussion laid the groundwork for a structured exploration of behavioural and algorithmic failure modes in nuclear crisis management.

As AI tools move from advisory to operational roles in national security infrastructure, the interface between human bias and technological artefact must become a central concern

Understanding the nature of human bias in nuclear decision-making

In the context of nuclear deterrence and escalation, decision-making is shaped by uncertainty, time pressure, adversarial framing, and increasingly, by digital tools that influence how data is filtered, framed, and presented.[8]

Against this backdrop, human bias operates not as a bug in the system but as an evolutionary feature: a set of cognitive shortcuts developed to cope with complexity, which become maladaptive in extreme, high-consequence environments. A key distinction made throughout the workshop by the expert participants was between three broad types of bias:

- **Cognitive biases** referring to the systematic patterns of deviation from normatively rational judgment, including *confirmation bias, availability heuristic*, and the *illusion* of control.
- **Organisational and interpersonal biases** arising from social dynamics within decision-making bodies, such as *authority bias, groupthink*, or *sycophantic conformity.*
- **Technologically mediated biases** emerging from the design, output, or embedded assumptions within automated tools and AI-driven systems, such as *automation bias* or *algorithm aversion.*

Why nuclear decision-making is especially vulnerable to bias

The nuclear domain exhibits a perfect storm of bias triggers[9]:

- **Existential stakes:** Decisions may determine the survival of populations or states.
- **Time pressure:** Choices may need to be made within minutes of an alert.
- **Ambiguity and limited verification:** Partial or contested data often drive early escalation.
- **Hierarchical structures:** Centralised command chains can stifle dissent.
- **Technological dependence:** Increasing reliance on automated detection, early-warning systems, and AI-assisted scenario analysis.

In such conditions, even normally adaptive heuristics become problematic because the tendency to assume the worst when faced with incomplete signals, known as worst-case thinking, may be justified in conventional risk management, but in a nuclear scenario, it can precipitate irreversible actions based on perceived, not actual, threats.

A typology of core biases

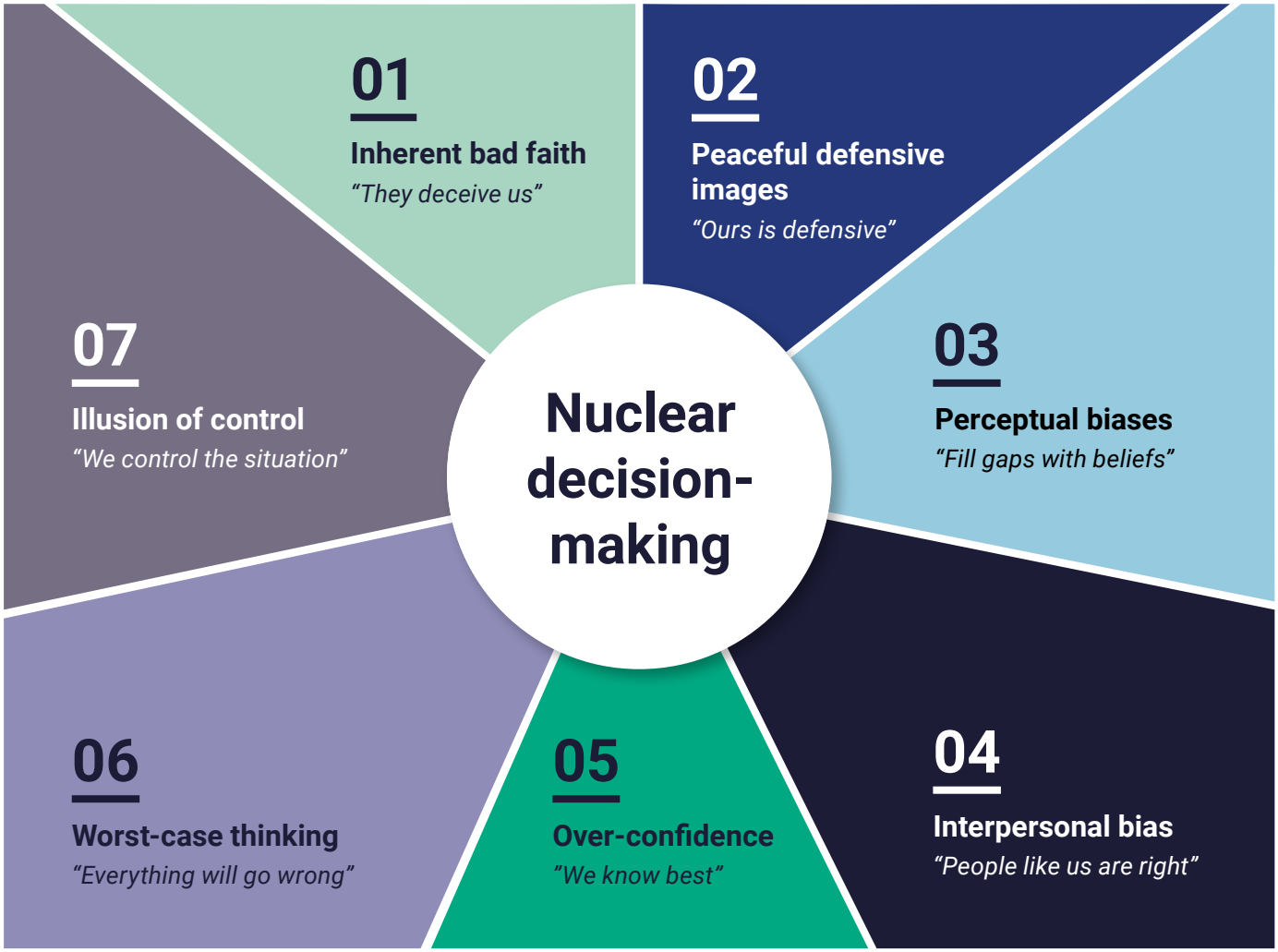
Below is a synthesis of the 7 core biases identified by expert participants during the workshop, which are linked to nuclear decision-making phases (see Figure 1) .

Participant-identified biases align closely with established findings in behavioural and strategic studies, providing empirical reinforcement rather than a departure from existing theory (see Table 1).[10]

Table 1 Seven core biases as identified by workshop participants

Bias	Description	Relevance in nuclear context
Illusion of control	Overestimating one's ability to manage complex systems or outcomes	May lead to escalation under the belief that actions can be finely calibrated or reversed
Inherent bad faith	Default assumption that adversaries are deceptive and hostile	May prevent constructive interpretation of peace signals or diplomatic openings
Peaceful defensive images	Belief that one's own posture is defensive while similar enemy actions are aggressive	Tends to reinforce adversarial framing; enables rationalisation of pre-emptive moves
Perceptual biases	Misjudgement driven by ambiguity, misinformation, or selective attention	May skew interpretation of strategic intent and technical signals (e.g. satellite imagery, intercepted comms)
Interpersonal biases	Groupthink, authority bias, and deference to seniority	Tends to suppress dissent and reduces diversity of perspectives in decision forums
Overconfidence	Excessive belief in the accuracy of one's own judgment or the robustness of systems	May lead to underestimation of adversary capability or over-trust in automated tools
Worst-case thinking	Prioritisation of highly improbable but catastrophic outcomes	Can prompt pre-emptive or disproportionate responses to ambiguous signals

Fig 1. The seven core biases in nuclear decision-making



Workshop design and methodology

During the workshop, participants interacted with a low technology readiness digital twin simulation tool designed to mimic high-stake environments.[11] Some expressed over-trust in the model's predictions ("the AI gave us confidence"), while others questioned the lack of interpretability ("we didn't know what it was optimising for"). These reflections highlight the need to treat AI not just as a source of insight but as an actor with its own decision-shaping influence.

The workshop was structured around the following core aims:

1. To identify several key human biases (or their groups) which are particularly relevant for nuclear decision making.
2. To explore potential methods for identifying and categorising bias in nuclear decision-making.
3. To explore how human and AI-mediated biases interact during fast-moving crisis scenarios.
4. To gather empirical observations from multi-disciplinary stakeholders using facilitated gameplay and scenario-based simulation using a Digital Twin tool.

Participants included strategic analysts, former military advisors, behavioural scientists, technology developers, and policy professionals from the UK, EU, North America, and Australia. To preserve openness in dialogue and encourage reflection, the workshop was conducted under Chatham House Rule.

Participants were explicitly invited to reflect on how cognitive and interpersonal biases could enter decision-making at multiple points across the digital twin lifecycle, rather than treating bias as a single downstream effect. The exercise was therefore designed as a digital twin of a nuclear crisis decision environment, enabling participants to explore how aspects of high-level decisions, information conditions, and technological interventions dynamically interacted to shape risk trajectories over time (leading either to nuclear escalation or de-escalation). The digital twin instantiated a fixed crisis scenario and allowed participants to interact with it repeatedly under three sequential treatments, holding the underlying system structure constant while varying the informational and technological context.

In the first round, participants engaged with the digital twin without the presence of disruptive technologies, allowing decisions to unfold based on conventional diplomatic options, intelligence updates, and group deliberation alone. This condition provided a baseline representation of decision-making under uncertainty and time pressure.

In the second round, the same digital twin was augmented with disruptive technologies, including AI-generated misinformation, deepfakes, and cyber-related disruptions. These elements altered the information environment within the twin, increasing ambiguity, cognitive load, and coordination challenges, and thereby amplifying conditions under which known cognitive and group-level biases are more likely to arise.

In the third round, disruptive technologies remained present, but the digital twin additionally incorporated AI-based bias guardrails and warnings derived from the previously developed 'Guardrails and Self-Assessment Framework'.[12] These interventions did not issue

Participants were explicitly invited to reflect on how cognitive and interpersonal biases could enter decision-making at multiple points across the digital twin lifecycle, rather than treating bias as a single downstream effect.

recommendations or optimise decisions. Instead, they flagged points in the decision process where bias could plausibly enter, prior to group deliberation and choice. The aim was to examine whether structured bias awareness embedded within the digital twin altered how participants interpreted information, coordinated within groups, and assessed risk.

Each group completed all three rounds, enabling within-group comparison of decision pathways and outcomes across conditions. Following completion of the prototype digital twin exercise, participants evaluated and compared the outcomes produced under each round. Groups and individual participants were then invited to elaborate explicitly on where biases emerged, how they manifested across different stages of the digital twin, and whether the presence of bias guardrails influenced decision quality, group dynamics, or perceived risk.

Bias narratives and thematic insights from the workshop

The illusion of control: Misplaced mastery in an unstable system

In the early group discussions, the illusion of control emerged almost immediately as participants reflected on the logic of deterrence and escalation management. Several participants remarked on the extensive procedural safeguards built into nuclear command and control systems and expressed confidence in institutional checks designed to prevent miscalculation. However, this confidence was not simply operational, it was also psychological. There was a shared belief that experienced leaders, supported by data and protocol, could always calibrate actions to avoid unintentional conflict.

What began as recognition of structural rigour gradually revealed an underlying bias: the assumption that escalation pathways could be entered and exited with precision. The idea that "we've modelled this before" or "we know where the red lines are" surfaced as a comforting narrative, even before gameplay introduced uncertainty or ambiguity. The bias thus appeared not as overconfidence in individuals, but in systems, and in the belief that the complexity of nuclear dynamics is governable through institutional foresight.

One of the early themes to arise from participants' reflections during the digital twin simulation also concerned the illusion of control in nuclear decision-making. As they navigated different phases of the scenario, participants discussed how decision-makers under real-world crisis conditions might over-estimate their capacity to manage escalation dynamics, believing they can modulate tension by signalling, pulling back, or escalating incrementally without triggering unintended consequences. This perceived ability to finely calibrate responses, they suggested, may become particularly compelling when supported by AI-generated recommendations that appear precise, neutral, and data-driven.

Several groups considered how artificial intelligence, presented as a dispassionate source of analysis, could reinforce a false sense of control by masking the complexity and unpredictability inherent in such scenarios. As one participant noted, the apparent confidence of AI predictions, such as low probabilities of conflict, could tempt decision-makers to adopt riskier strategies than they otherwise might. In such cases, AI may not correct cognitive bias but instead reinforce pre-existing preferences under the guise of neutrality.

In these discussions, the illusion of control was not framed as a behavioural flaw unique to individuals, but as a systemic feature of the decision-making environment, where the combined influence of protocol, modelling, and high-stakes urgency creates a narrative that decisions are reversible. Participants emphasised that, in reality, each choice commits actors more deeply to a trajectory that may become increasingly difficult to alter.

Participants discussed how decision-makers under real-world crisis conditions might over-estimate their capacity to manage escalation dynamics.

Inherent bad faith: Reading malice into ambiguity

Initial conversations around adversary intentions were revealing. When prompted to consider possible misinterpretations or cooperative gestures in crisis scenarios, participants were quick to raise the 'maskirovka' tradition in Russian doctrine (a longstanding

concept of military deception involving camouflage, concealment, disinformation, and the deliberate manipulation of adversary perceptions)[13] or historical examples of deception in Cold War settings.

Even before entering the simulation, many assumed that any conciliatory signals would be disingenuous or tactically manipulative. This discussion framed adversaries as strategically duplicitous by default. Participants reflected on specific crises, such as Able Archer '83 and the Kargil conflict, to support this view, arguing that actors tend to interpret cooperation as weakness or subterfuge. The tone of these discussions suggested that bad faith was seen not as a possibility but as a baseline assumption, one that had to be disproven, not merely questioned. This foundational mistrust shaped the lens through which any future AI outputs or adversary actions would be interpreted.

As participants engaged with the ambiguous signals embedded in the twin simulation, such as partial satellite data or intercepted messages lacking clear provenance, they reflected on how, in real-world nuclear crises, decision-makers might default to adversarial interpretations. Particular attention was paid to how seemingly conciliatory actions—offers of dialogue or de-escalation—could be reinterpreted through a lens of strategic deception. Participants discussed how, historically and potentially today, such overtures might be dismissed not on their own terms, but as tactical manoeuvres to buy time or mask preparation for escalation.

These reflections echoed the well-documented inherent 'bad faith' model from political psychology, where actors are presumed to act duplicitously regardless of incoming evidence to the contrary.

Several participants considered how this bias can persist even when no clear hostile intent is present, and how it might close off avenues for de-escalation or diplomatic compromise. The assumption of malevolent intent, they suggested, can harden policy responses and narrow the perceived legitimacy of peaceful engagement.

What made the discussion particularly insightful was the recognition that such bias often operates not at the individual level, but through collective memory and institutional narrative. Participants explored how historical experiences of betrayal or conflict might serve to reinforce suspicion in the present. In these scenarios, a single sceptical voice asking, 'what if this gesture is genuine?' might be overridden by shared recollections of earlier crises—suggesting that bad faith assumptions are sustained as much by narrative continuity as by current intelligence.

The defensive posture paradox: Peaceful in our eyes alone

Bias often operates not at the individual level, but through collective memory and institutional narrative

Another recurring theme in participants' reflections was the tendency of states to frame their own posture as inherently defensive, irrespective of how it might be perceived externally.[14] Participants frequently returned to the notion that their own state's nuclear posture was defensive in nature. In early discussions, doctrines such as 'minimum deterrence' and 'strategic stability' were invoked as evidence of restraint. When asked how such

postures might be perceived by adversaries, many acknowledged the theoretical possibility of misunderstanding, yet maintained a strong conviction that their own actions would ultimately be recognised as rational and defensive. The bias thus emerged through a moral narrative: that we act to prevent war, while they provoke it. Notably, this framing often remained unchallenged even when participants explicitly discussed symmetrical behaviour across opposing sides. The assumption of defensive purity, and the limited consideration of how identical actions might be interpreted differently by others, was already firmly in place before any simulation stimuli were introduced.

As the simulated crisis unfolded, participants explored how this initial framing translated into concrete interpretations of action and escalation. Decisions such as military mobilisation or the deployment of nuclear-capable systems were commonly justified as deterrent measures intended to preserve stability rather than provoke conflict. Even manoeuvres carrying high signalling weight were frequently interpreted by those implementing them as prudent, stabilising steps taken in the name of peace. Crucially, participants recognised that this internal logic generated a dangerous asymmetry of perception: actions viewed as restrained and responsible on one's own side were often characterised as aggressive or escalatory when mirrored by an adversary.

This pattern revealed what participants described as a reciprocal cognitive blind spot, in which each actor positions itself as the guardian of stability while attributing malign intent to the other. As one participant observed, "We always interpret our own caution as signalling discipline, but when the other side does it, we suspect it's a provocation."

This reflective discussion engaged directly with what has been termed the 'peaceful defensive image' bias, a systemic feature of deterrence logics. Participants emphasised that the bias is not the result of misinformation or deception, but rather a by-product of how strategic intent is internally framed and communicated. Without deliberate efforts to interrogate these narratives during a crisis, they warned, the bias can persist and intensify, leading both sides to misread the situation while believing they are acting responsibly.

Perception under pressure: Ambiguity, misleading certainty, and false clarity

One participant observed, "We always interpret our own caution as signalling discipline, but when the other side does it, we suspect it's a provocation."

Perception bias entered the discussion when participants considered how crisis signals are typically received and interpreted in real time. Groups reflected on the sheer volume of data that flows into early warning systems, from satellite imagery and cyber intelligence to social media and diplomatic cables. Several participants noted that, in practice, there is rarely enough time to cross-check all sources, meaning first impressions often frame the interpretation of unfolding events. This led naturally to a discussion of ambiguity: how much can be inferred from partial data, and what kinds of errors are most common when under pressure? Some participants discussed previous experiences with misinterpreted signals, including non-state actors or technical failures, and noted the role of perception filters shaped by prior crises. Even at this

early stage, there was an emerging recognition that perception is never neutral—it is always filtered through cognitive and institutional templates. Yet few had strategies ready for how to contest these filters in real time.

During the digital twin simulations, participants encountered multiple scenarios in which information was deliberately incomplete, inconsistent, or open to multiple interpretations. These conditions, such as partial satellite imagery, ambiguous intelligence summaries, and unclear adversary signalling, were embedded by design to simulate the ambiguity that often characterises real-world nuclear crises. Rather than attempting to resolve these uncertainties, participants used the scenarios as a prompt to reflect on how actual decision-makers might respond under similar conditions. In these reflections, participants noted that ambiguity often does not lead to restraint or further inquiry but instead encourages narrative completion. They discussed how, in practice, leaders and analysts may quickly fill informational gaps with assumptions that align with pre-existing expectations or strategic beliefs. Once a working interpretation is established, they suggested, alternative explanations are often dismissed—not necessarily because they are implausible, but because they complicate the decisional momentum already underway.

Participants also discussed how AI-generated outputs, particularly those offering high-confidence judgments without transparent reasoning, might inadvertently reinforce this dynamic. There was concern that decision-makers might treat algorithmic 'likelihood estimates' as objective truth, especially when such outputs align with their own intuitions. In these cases, AI does not challenge the cognitive tendency to resolve ambiguity prematurely — it may entrench it.

What emerged from the discussion was a picture of how perceptual tunnel vision can form under uncertainty: early impressions guide the framing of ambiguous data, which is then validated, consciously or not, by technical systems trained on similarly biased priors. Participants emphasised that in high-stakes environments, the psychological need to impose clarity may override the strategic need to maintain flexibility, particularly when ambiguity is experienced not as a threat, but as a discomfort to be resolved.

Interpersonal dynamics: Groupthink and the deference dilemma

There was concern that decision-makers might treat algorithmic 'likelihood estimates' as objective truth, especially when such outputs align with their own intuitions.

Before engaging in gameplay, participants were invited to reflect on how group structures and decision hierarchies influence nuclear judgment. This prompted immediate recognition of the risks of groupthink. Many drew on personal or institutional examples of meetings where consensus formed too quickly, dissent was silenced, or senior voices dominated. Interestingly, some participants framed these dynamics as unavoidable: "Doctrines are rarely challenged during crises especially when coming from figures of authority," one participant said. Others noted how different cultural or national traditions manage authority, with some fostering open challenge, and others discouraging it. The issue of sycophancy emerged laterally, especially in discussion of military or intelligence chains, where staff may offer what they think

leadership wants to hear. These reflections were not hypothetical; they drew on lived experiences of real-world bureaucratic dynamics. Participants acknowledged that these biases can feel safer than the alternative: sticking out, slowing down, or undermining unity during high-pressure moments.

Throughout the twin exercise, participants reflected on the importance of interpersonal biases in nuclear decision-making, particularly those that tend to manifest subtly yet persistently in high-pressure settings. One recurrent observation was the strong tendency for decision-making to default to internal hierarchies during crises, especially when time constraints demand rapid coordination. Even in the context of informal group exercises, participants noted how authority tended to concentrate around those who spoke with confidence or appeared to possess technical expertise. This dynamic, they suggested, mirrors real-world nuclear environments, where urgency often leads to privileging cohesion and decisiveness over inclusive deliberation.

In discussing hypothetical scenarios, participants raised concerns about how flawed interpretations might take hold, not necessarily because they are the most plausible, but because they are articulated by a trusted or dominant voice. One group examined a case in which a questionable inference based on ambiguous data might be acted upon simply because it came from someone who projected confidence. The concern was not the competence of the speaker, but the reluctance of others to challenge a seemingly authoritative view under pressure.

While not all participants identified classic groupthink dynamics in their own discussions, and many groups featured active disagreement, there was widespread recognition that once a provisional consensus forms, it becomes difficult to reopen. Raising objections in such moments was seen as socially costly, especially when others were eager to proceed. Participants reflected on how, in institutional settings, this can discourage dissent, even among highly capable actors.

These discussions highlighted that interpersonal bias is not only a matter of individual cognition but is deeply embedded in relational dynamics, between roles, expectations, authority structures, and the performative pressure of decision-making under pressurised uncertainty. Managing these dynamics, participants concluded, requires deliberate design: fostering environments where doubt is not just permitted but structurally enabled.

Overconfidence in human and machine judgement

Conversations about overconfidence tended to revolve around two axes: belief in one's own side's capability and the assumption that adversaries would behave rationally. Several participants noted that their strategic communities often discount irrational or emotionally driven decisions from opponents, believing that all actors ultimately seek survival and predictability. This, in turn, fostered confidence in escalation control mechanisms and nuclear signalling strategies. Others discussed overconfidence in technological systems, particularly the reliability of early warning platforms and command communication channels. There was an assumption that failure

modes were known and accounted for. Few participants, at this stage, anticipated cascading failures or deeply nonlinear responses to seemingly small moves. The bias of overconfidence emerged not as bravado, but as quiet faith—in doctrine, in system integrity, and in the professionalism of actors involved.

In discussions following the early phases of the simulation, participants reflected on how, in real-world nuclear decision-making, actors may exhibit high levels of confidence in their ability to anticipate adversary behaviour and interpret technical system outputs. This confidence, they suggested, often stems not from complete situational awareness but from reliance on generalised models of rational behaviour or familiar strategic frameworks.

The use of an AI-based decision-support tool in the simulation prompted participants to explore how such systems might amplify these tendencies. Some noted that decision-makers may assume AI outputs are built on the latest intelligence or objective data, while others observed that the mere presence of quantified, visually structured information can lend an illusion of rigour and precision. Participants considered how, under operational conditions, AI outputs that align with prior expectations might be readily accepted, while contradictory ones might be discounted or rationalised away.

These reflections did not suggest an uncritical faith in AI. On the contrary, many participants highlighted the potential for healthy scepticism. But they acknowledged that such scepticism may be unevenly applied, stronger when the system challenges human intuition, weaker when it affirms it. This asymmetrical trust, they argued, represents a convergence of automation bias and human overconfidence: each validating the other in ways that can entrench rather than interrogate flawed assumptions.

The concern, then, is not simply overreliance on machines or overconfidence in human judgment, but the reinforcing loop that emerges when both operate within the same cognitive frame. In this sense, participants emphasised, confidence becomes self-sustaining: not because decisions are necessarily correct, but because tools and intuitions feed back into each other without external correction.

Worst-case thinking: The fear of getting it wrong

Finally, worst-case thinking arose organically as participants discussed the cost of hesitation. Drawing on both historical crises and imagined future scenarios, several group members made the case that not acting quickly enough could be as dangerous as acting too aggressively. This led to a recurring dilemma: when the cost of getting it wrong is total, how much should you hedge against unlikely but catastrophic events? One participant referenced the logic of “strategic pessimism,” arguing that leaders often have no choice but to anticipate the most dangerous version of an adversary’s intention. Others expressed discomfort with this logic but conceded that institutional incentives often reward escalation over risk tolerance. Thus, worst-case thinking was not framed as irrational; it was cast as structurally embedded. It emerged as a rational strategy in an irrational environment.

The bias of overconfidence emerged not as bravado, but as quiet faith—in doctrine, in system integrity, and in the professionalism of actors involved.

The question, left unresolved, was how to distinguish between prudence and paranoia.

As the simulation scenarios progressed, participants began to reflect on one of the most quietly influential dynamics they believed would shape real-world nuclear decision-making: the deeply embedded belief that failing to act pre-emptively could lead to irreversible catastrophe. This reflection centred on the tension between caution and pre-emption, wherein uncertainty itself becomes a driver for escalatory choices. Participants discussed how, under high-stakes ambiguity, decision-makers may lean toward acting "just in case," not because the intelligence clearly warrants it, but because the perceived cost of inaction feels too great to risk. Importantly, participants were careful to frame this impulse not as reckless or emotional, but as a consequence of the burden of responsibility. Decision-makers, they argued, might escalate not out of aggression but out of a desire to avoid under-reacting in the face of a potentially existential threat. However, they also recognised that such framing could lead to decisions that exceed the actual demands of a situation. When reflecting on why diplomatic or de-escalatory options were often set aside in strategic contexts, one participant observed that leaders may fear the reputational or historical consequences of being perceived as hesitant in the face of danger.

This line of reflection drew attention to the asymmetry often present in crisis reasoning: the prioritisation of avoiding Type II errors, failing to act on a genuine threat, over Type I errors — overreacting to a false signal. While the former is understandable, participants noted that this emphasis can produce a systematic skew in judgment. Rather than being anchored in the most probable outcomes, decisions may instead be shaped by what is least politically or personally defensible in retrospect.

It is in this context that worst-case thinking becomes particularly insidious. Participants reflected that it can feel prudent, even necessary, but it subtly narrows the space for moderate or time-buying options. Once the possibility of catastrophe dominates the strategic horizon, the imperative to act begins to override the logic of delay, dialogue, or contingency, effectively collapsing the middle ground even in situations where it may still be viable.

The role of technology in shaping, amplifying, and containing bias

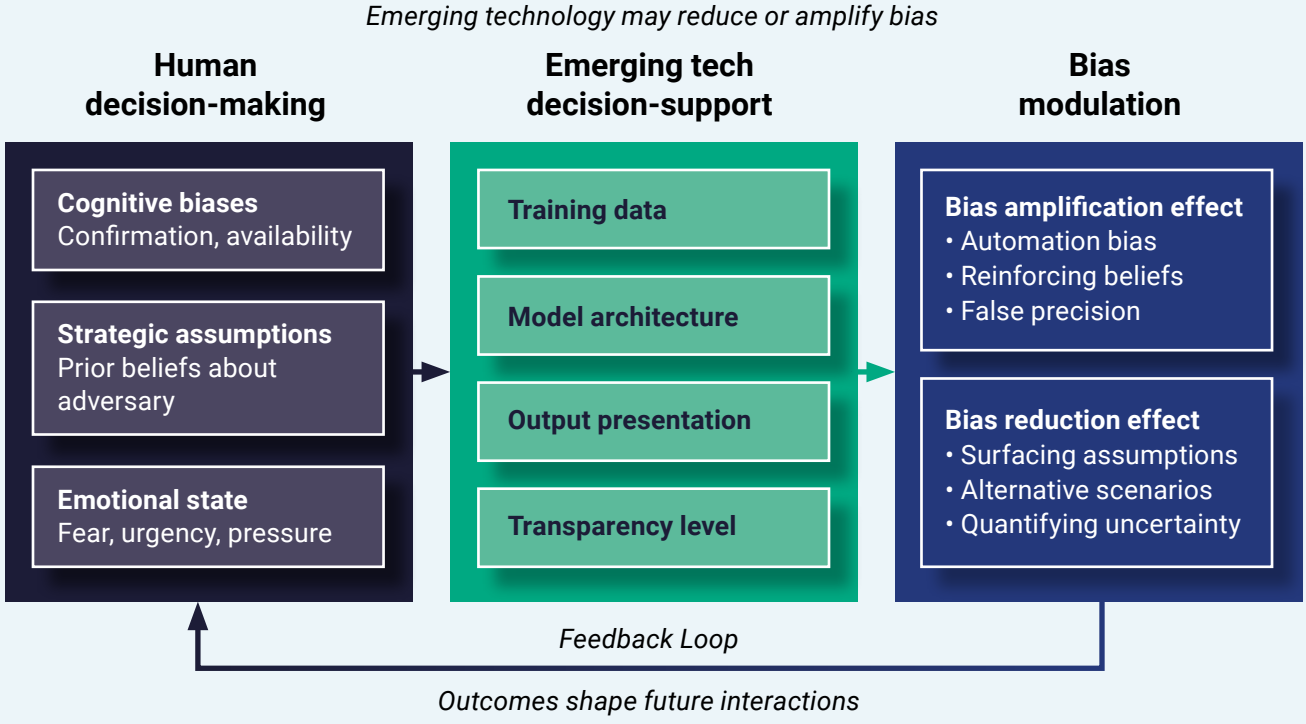
Participants' reflections revealed a complex interplay. In scenarios where the AI output aligned with a group's strategic expectations, it often functioned as a form of psychological reinforcement. The output was taken less as new evidence and more as affirmation; as one participant observed, "It confirmed what we were already thinking". However, when the system's outputs, decision-support feedback, or recommendations contradicted prevailing assumptions, they were more likely to be sidelined than explored. Few groups, in their reflections, identified moments where users would pause to ask critical questions such as: What underlying assumptions is this model operating on? What are its data limitations? How would one detect an error in its logic?

Few groups, in their reflections, identified moments where users would pause to ask critical questions such as: What underlying assumptions is this model operating on?

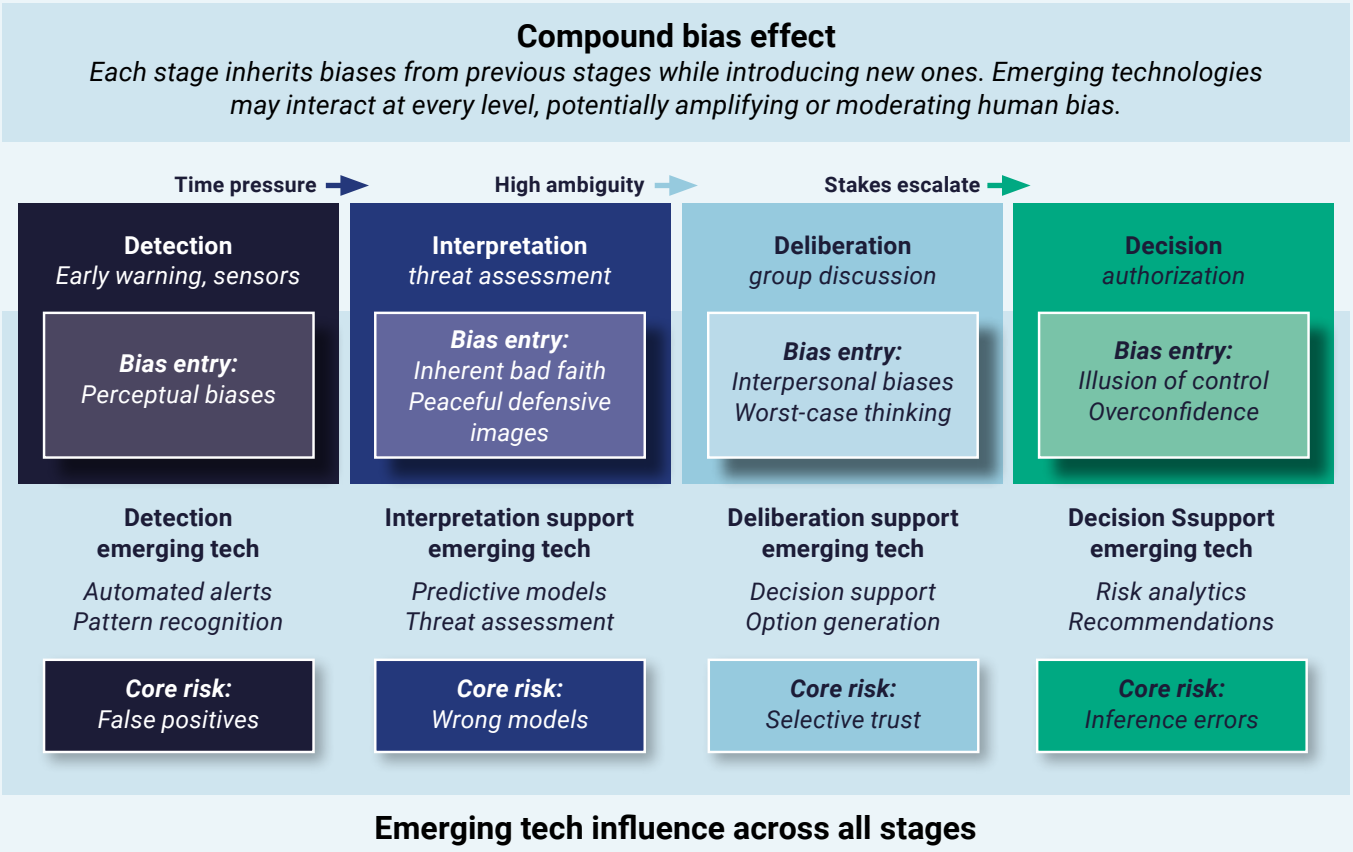
This pattern suggested a broader concern around selective trust. Participants noted that in real-world decision-making, humans may rely on AI when it reinforces their views but resist it when it challenges them, unless those humans are themselves uncertain or under-confident, in which case the system may become a decision-making crutch. This dual tendency reflects the simultaneous presence of automation bias and algorithm aversion, operating in unpredictable and context-dependent ways (see Figure 2).

Participants also reflected on the visual and temporal influence of AI. Decision-support tools used in the simulation delivered probabilistic warnings, scenario predictions, and threat escalations, often via dashboards with visual aids such as graphs, sliders, or confidence intervals. Though these outputs were not always transparent in terms of their underlying logic, their graphical clarity lent them an aura of credibility. Participants noted how these visual cues could disproportionately shape a decision-maker's sense of urgency or priority. Several groups imagined how, in practice, such systems might become temporal anchors, reorganising the rhythm of group deliberations and shifting the focus of attention, not necessarily by what they predicted, but by when and how they delivered information.

Fig 2. Emerging technology as bias modulator: Human-machine linteraction



I Fig 3. Nuclear crisis decision-making: Compounding effects of bias and emerging tech



This aspect of reflection proved critical: AI was not perceived as merely adding data to an already rational process. Instead, it was recognised as reshaping the contours of reasoning itself, modulating when decisions are made, what types of evidence are weighted, and how the threshold for action is calibrated. These effects may be subtle, but they are structurally significant, particularly in nuclear contexts where every shift in tone, tempo, or trust carries disproportionate consequences.

Participants concluded that AI cannot be understood solely as a tool for reducing bias; nor should it be feared only as a source of distortion (see Figure 3). Rather, it must be seen as a bias modulator, a force that can either amplify or attenuate human tendencies depending on how it is engaged. It may reinforce overconfidence by simulating precision, but it can also expose hidden assumptions by modelling uncertainty. Its impact is not deterministic; it is contingent. And this contingency, participants argued, places renewed importance on human responsibility: not just in interpreting AI, but in designing decision processes that anticipate its effects.

Concluding reflections and future directions

This report has aimed to capture an evolving conversation, one that sits at the intersection of behavioural science, strategic studies, and technological design. Bias in nuclear decision-making is not a new concern. But in a world increasingly shaped by artificial intelligence, accelerated decision cycles, and contested narratives, it demands renewed attention. The workshop revealed that biases do not merely emerge from ignorance or error. They emerge from structure. They are shaped by how information is framed, how authority is distributed, how time is managed, and how machines are positioned within human decision-making loops. In nuclear contexts, these structural forces amplify risk, not always through recklessness, but often through misplaced confidence, premature certainty, or the narrowing of imagined alternatives.

If there is one unifying insight from the workshop, it is this: bias thrives where reflection is absent. Whether in the form of over-trusting an AI model, deferring too quickly to a dominant voice, or reading malice into ambiguous signals, the most consequential errors were not rooted in malice or irrationality. They were rooted in momentum, in the sense that, under pressure, decisions must move forward, even if clarity has not yet arrived. And this is where technology enters not only as a risk, but as a potential ally. Well-designed decision-support tools, simulations, and digital twins can serve as bias disruptors. They can inject counterfactuals, force the articulation of assumptions, and simulate alternative timelines. But for this to work, they must be transparent, interpretable, and embedded in a decision culture that values process as much as prediction.

This report does not offer final answers. It does not claim to have defined an exhaustive taxonomy of bias, nor to have resolved how AI should be embedded in nuclear command structures. Instead, it offers a beginning, a step toward a more systematic, empirically informed, and behaviourally grounded approach to nuclear risk.

If there is one unifying insight from the workshop, it is this: bias thrives where reflection is absent.

Appendix A:

Extended Landscape of Cognitive Biases Relevant to High-Stakes Decision-Making (reproduced from Pogrebna and Renaud, 2025)

INDIVIDUAL	Personal Characteristics	Being humans not irrational	Aa Ambiguity attitude	Ra Risk attitude																													
	Limitations of the brain	Mind is flat	Ah Availability heuristic	Be Bizarreness effect	Cb Consistency bias	Ce Context effect	Cn Conservatism or Regressive bias	Cs Cross-race effect	Cy Crypto-mnesia	Eg Egocentric bias	Fa Fading affect bias	Fm False memory	Gs Generation/ Self-generation effect	Go Google effect	He Humor effect	Le Lag effect	Ll List-length effect	Lp Levels-of processing effect	Ls Leveling and sharpening	Mc Mood-congruent memory bias	Me Mis-information effect	Mo Modality effect	Nx Next-in-line effect	Pc Part-list cueing effect	Pi Primacy, recency & serial position effect	Po Positivity effect (Socioemo-tional selectivity)	Pr Peak-end rule	Ps Picture superiority effect						
			Ch Choice supportive bias	Ct Contrast effect	De Declinism	Ia Implicit association	Ne Negativity bias or Negativity effect	Pd Processing difficulty effect	Sc Source confusion	Sf Suffix effect	Sl Spotlight effect	Sp Spacing effect	Sr Self-relevance effect	Su Suggestibility	Sy Stereoty-pical bias	Te Testing effect	Tl Telescoping effect	Tp Tachypsychia	Ts Travis Syndrome	Tt Tip of the tongue	Rb Remini-scence bump	Rf Von Restorff effect	Ve Verbatim effect	Ze Zeigarnik effect	Af Anchoring or focalism	At Attentional bias	Fe Focusing effect	Fi Frequency illusion or Baader- Mein-hof effect	Sb Selection bias					
	Judgement	Everything is relative	Du Choice supportive bias	Gf Gambler's fallacy	Hd Hyperbolic dicounting	Lc Licensing effect	Ml Moral creden-tial effect	Ri Recency illusion	Pj Projection. bias	Pt Present. bias	Tm Time-saving bias	As Attribute substitution	Au Automation bias	Bk Berkson's paradox	Br Base rate fallacy or Base rate neglect	Ci Clustering illusion	Cj Conjunction fallacy	Di Distinction bias	Dt Default effect	Dy Decoy effect	Fr Framing effect	Ib Information bias	Lb Less-is-better effect	Ng Neglect of probability	Pk Parkinson's law of triviality	Rg Regressive bias	Sv Survivor-ship bias	Su Subadditivity effect	Wr Well travelled road effect	Zs Zero-sum. bias				
			Dk Dunning-Kruger effect	Hy Hard-easy effect	Oc Over confi-dence effect	Pb Pessimism bias	Re Restraint bias	Co Congruence bias	Is Insensitivity to sample size	It Illicit transference	Li Law of the instrument	Lk Lookelsewhere effect	Ob Observer-expectancy effect	Ot Outcome bias	Pf Planning fallacy	Sg Surrogation	Wf Weber-Fechner law	Df Denomination effect	Ds Disposition effect	Ew Endowment effect	Ie IKEA effect	Ir Irrational escalation/ Escalation or commitment	La Loss aversion	Mi Money illusion	Nh Not invented here	Pe Pseudo-certainty effect	Ub Unit bias							
	Decision making	Belief is reality	Bf Backfire effect	Bb Belief bias	Cf Confirmation bias	Cl Continued influence effect	Cv Conservatism (belief revision)	Ee. Experimenter's or expectation bias	Eh End-of-history illusion	Ex Exaggerated expectation	Fb. Forer effect or Barnum effect	Ff Form function attribution bias	Hb Hindsight bias	Hh Hot-hand fallacy	Ic Illusion of control	Il Illusory correlation	Iv Illusion of validity	Iy Illusory truth effect	Mx Mere exposure effect	Nb Normalcy bias	Om Omission bias	Os Ostrich effect	Pa Pareidolia	Rd Reactive devaluation	Rp Risk compensation / Peltzman effect	Rr Rhyme as reason effect	Sj Subjective validation	Sm Semmelweis reflex	Sp Selective perception	Sq Status quo bias	Zr Zero-risk bias			
		Inferences are key	Ae Ambiguity effect	Cc Allais Paradox: Common conse-quence	Cr Allais Paradox: Common Common ratio	Ep Ellsberg Paradox	St St Petersburg Paradox	Vt Violation of Transitivity	Vi Violation of Indep-en-dence	Vd Violation of Dominance	An Anthropo centric thinking	Am Anthropo morphism or personifi-cation	Fn Functional fixedness	Da Dread aversion	Em Empathy gap	Im Impact bias	In Interoceptive bias	Op Optimism bias	Pv Pro-innovation bias	Rc Reactance	Sa Salience bias													
SOCIAL	Social	Interactioneffects	Ab Authority bias	Ac Actor-observer bias	Cd Cheerleader effect	Db Defensive attribution hypothesis	Ec Egocentric bias	Et Extrinsic incentives bias	Fc False consensus effect	Fd Fundamental attribution error	Fu False uniqueness bias	Ha Halo effect	Ii Illusion of asymmetric insight	Ip Illusion of transparency	Io Illusory superiority	Ix Illusion of external agency	Jx Just-world hypothesis	Mr Moral luck	Na Naïve cynicism	Nr Naïve realism														
			Bn Ben Franklin effect	Bs Bias blind spot	Cm Compassion fade	Ck Curse of knowledge	Cu Courtesy bias	Hs Hostile attribution bias	If Identifiable victim effect	Py Pygmalion effect	Sd. Social desirability bias	Sh Shared information bias	Si Singularity effect	Sn System justification	Ss Self-serving bias	Sx Sexual over/ under perception bias	Ta Trait ascription bias	Ua Ultimate attribution error	Wa Worse-than-average effect															
		Collective effects	Av Availability cascade	Bw Bandwagon effect	Ga Group attribution error	Gt Groupthink	Hr Herding	Ig Ingroup bias	Oh Outgroup homogeneity bias	So Social comparison bias	Se Stereotyping	Td Third-person effect	Ww Women are wonderful effect																					

KEY

- Attention effects
- Belief-based biases
- Choice under risk/uncertainty
- Confidence effects
- Creativity effects
- Human characteristics
- Information processing
- Interpersonal effects
- Intertemporal choice
- Memory effects
- Methodology biases
- Populational and group effects
- Sentiment and senses
- Value processing

www.bigbadbiasbook.com
www.behaviouraldatascience.org

References

1 Bianco, Belen., Paul, Rishi, (2024). Technological Complexity and Risk Reduction: A Guardrails and Self-Assessment Framework for EDTs in NC3 and Nuclear Weapons Decision-Making, European Leadership Network, https://europeanleadershipnetwork.org/wp-content/uploads/2024/07/24_07_17_Technological-Complexity-and-Risk-Reduction-report.pdf

2 Johnson, J. (2021). 'Catalytic nuclear war' in the age of artificial intelligence & autonomy: Emerging military technology and escalation risk between nuclear-armed states. Journal of Strategic Studies, 1-41.

3 Fewell, M. P., & Hazen, M. G. (2005). Cognitive issues in modelling network-centric command and control, accessed at https://www.researchgate.net/profile/Mark-Hazen/publication/27253675_Cognitive_Issues_in_Modelling_Network-Centric_Command_and_Control/links/0046353c80a41e29d0000000/Cognitive-Issues-in-Modelling-Network-Centric-Command-and-Control.pdf

4 Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In Decision making in aviation (pp. 289-294). Routledge.

5 Pogrebna, G. & K. Renaud (2025). Big Bad Bias Book: A field guide to over 200 cognitive biases that shape how we think, decide, and behave (Illustrated ed.). Behavioural Data Science Press, Sydney, Australia.

6 Atalan, Y., Reynolds, I., & Jensen, B, (2025). AI Biases in Critical Foreign Policy Decisions, CSIS, https://www.csis.org/analysis/ai-biases-critical-foreign-policy-decisions, (February 26).

7 Noyes, J., Cook, M., & Masakowski, Y. (Eds.). (2012). Decision making in complex environments. Ashgate Publishing, Ltd.

8 Geist, E. (2023). Deterrence Under Uncertainty: Artificial Intelligence and Nuclear Warfare. Oxford University Press.

9 Harrington, A. I., & Knopf, J. W. (Eds.). (2019). Behavioral Economics and Nuclear Weapons. University of Georgia Press.

10 For contextual reference, an extended landscape of cognitive biases documented in the behavioural science literature as presented in Pogrebna and Renaud (2025) is provided in Appendix A. Pogrebna, G., Paul, R., Damaj, N., McNaughton, J., Stacey, G., & Omoijiade, I. M. (2025). Technological complexity and risk reduction, European Leadership Network, accessed at https://europeanleadershipnetwork.org/policy-brief/technological-complexity-and-risk-reduction-using-digital-twins-to-navigate-uncertainty-in-nuclear-weapons-decision-making-and-edt-landscapes/

11 Pogrebna, Paul et al., (2025).

12 Bianco, Paul. (2024)

13 Glantz, David M. (1989) Soviet Military Deception in the Second World War. London: Frank Cass.

14 Jervis, R. (1976). Perception and Misperception in International Politics. Princeton University Press.

The European Leadership Network (ELN) is an independent, non-partisan, pan-European network of over 450 past, present and future European leaders working to provide practical real-world solutions to political and security challenges.

Contact
For further information on the ideas in this report, please contact secretariat@europeanleadershipnetwork.org

Published by the European Leadership Network, February 2026

European Leadership Network (ELN)
8 St James's Square
London, UK, SE1Y 4JU

@theELN | europeanleadershipnetwork.org

EUROPEAN LEADERSHIP NETWORK

Federal Foreign Office

European Leadership Network
8 St James's Square
London, SW1Y 4JU
United Kingdom

Email: secretariat@europeanleadershipnetwork.org
Tel: 0203 176 2555

Follow us

europeanleadershipnetwork.org