CCDCOE
NATO COOPERATIVE
CYBER DEFENCE
CENTRE OF EXCELLENCE

A Structural Model of the Human Agent in Cognitive Warfare

Dr Fedir Korobeynikov, Dr Andrii Davydiuk, Prof Volodymyr Mokhor

NATO Cooperative Cyber Defence Centre of Excellence
G.E. Pukhov Institute for Modelling in Energy Engineering National Academy of Sciences of Ukraine
State Service of Special Communications and Information Protection of Ukraine

Tallinn 2026

About the authors

Fedir Korobeynikov is a PhD in Information Security, director of digital technologies and information security at System Capital Management, founder of the Security Studies and Research Center (Kyiv, Ukraine), and doctoral student (D.Sc. program) at the G.E. Pukhov Institute for Modelling in Energy Engineering, National Academy of Sciences of Ukraine

Andrii Davydiuk is a PhD in Cybersecurity, Ambassador at NATO CCDCOE, deputy branch head in the State Cyber Protection Centre State Service of Special Communications and Information Protection of Ukraine, senior scientific research staff, doctoral student (D.Sc. program) at the G.E. Pukhov Institute for Modelling in Energy Engineering National Academy of Sciences of Ukraine.

Volodymyr Mokhor is a DrSc, Professor in Math Modelling, CyberSecurity and Risk Science, corresponding member National Academy of Sciences (NAS) of Ukraine, Vice Chair at the Division of Energy and Energy Technologies NAS of Ukraine.

CCDCOE

The NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) is a NATO-accredited knowledge hub offering a unique interdisciplinary approach to the most relevant issues in cyber defence. The heart of the CCDCOE is a diverse group of international experts from military, government, academia, and industry, currently representing 39 nations.

The CCDCOE maintains its position as an internationally recognised cyber defence hub, a premier source of subject-matter expertise and a fundamental resource in the strategic, legal, operational, and technical aspects of cyber defence. The Centre offers thought leadership on the cutting edge of all aspects of cyber defence and provides a 360-degree view of the sector. The Centre encourages and supports the process of mainstreaming cybersecurity into NATO and national governance and capability, within its closely connected focus areas of technology, strategy, operations, and law.

The Tallinn Manual, prepared at the invitation of the CCDCOE, is the most comprehensive guide for policy advisers and legal experts on how international law applies to cyber operations carried out between and against states and non-state actors. Since 2010, the Centre has organised Locked Shields, the biggest and most complex technical live-fire cyber defence challenge in the world. Each year, Locked Shields gives cybersecurity experts the opportunity to enhance their skills in defending national IT-systems and critical infrastructure under real-time attacks. The focus is on realistic scenarios, cutting-edge technologies, and simulating the entire complexity of a massive cyber incident, including strategic decision-making and legal and communication aspects.

The CCDCOE hosts the International Conference on Cyber Conflict, CyCon, a unique annual event in Tallinn, bringing together key experts and decision makers from the global cyber defence community. The conference, which has taken place in Tallinn since 2009, attracts more than 600 participants each spring.

The CCDCOE is responsible for identifying and coordinating education and training solutions in the field of cyber defence operations for all NATO bodies across the Alliance. NATO-accredited centres of excellence are not part of the NATO Command Structure.

www.ccdcoe.org
publications@ccdcoe.org

Disclaimer

This publication is a product of the CCDCOE. It does not necessarily reflect the policy or the opinion of the Centre or NATO. The Centre may not be held responsible for any loss or harm arising from the use

of information contained in this publication and is not responsible for the content of the external sources, including external websites referenced in this publication.

Digital or hard copies of this publication may be produced for internal use within NATO and for personal or educational use when for non-profit and non-commercial purpose, provided that copies bear a full citation.

1. Table of Contents

2. Abstract 5
3. Introduction 6
4. The Reduction of the Subject: Why Standards Remain Blind 8
5. The Topology of Identity: A Formal Model of the Agent 9
6. The Anatomy of Cognitive Attack 11
   6.1. Cognitive Biases as Structural Vulnerabilities 11
   6.2. Weaponisation: The Mechanism of Cognitive Attack 11
   6.3. The Topology of Attack: Three Forms of Influence 12
   6.4. The Choice of Point of Application: System, Agent, and Crystallisation 13
   6.5. A Running Example: Confirmation Bias as an Attack Vector 14
7. Operationalisation: Towards Early-Warning Systems for Cognitive Threats 16
8. Towards Topological Defence 18
9. Conclusion 21
10. Bibliography 22

2. Abstract

This article develops a structural model of the human agent in cognitive warfare. In doing so, it extends the structure-centric approach previously applied at the level of socio-technical systems. Its point of departure is the constitutive invisibility of cognitive influence: unlike kinetic, cyber, or space operations, a successful cognitive attack need not appear to the targeted agent as an attack at all, but may be experienced as the natural evolution of that agent's own views. Existing doctrines of cognitive defence remain largely content-oriented: however sophisticated, they operate at the level of information – its veracity, volume, or provenance – rather than at the level of the structures through which information is processed. Such approaches are insufficient where the object of attack is not what the agent thinks, but the architecture through which perception, judgement, and decision are formed.

The article therefore models the human agent not as an atomic node within a socio-technical network, but as a multiplex structure composed of layer-specific states stabilised by personal invariants. Within this framework, cognitive attack is conceptualised as a perturbation of the connection that governs the agent's transitions between contexts. The attack exploits the natural asymmetries of this connection – latent zones of reduced resistance, of which cognitive biases are one instance – driving inter-layer incoherence beyond the threshold of transmorphance, up to the point at which the invariant core that sustains identity begins to erode.

The article distinguishes three structurally distinct forms of cognitive attack: the disruption of transition coherence, the accumulative transformation of identity through altered holonomy, and the degradation of the invariant core or of invariant compatibility between agents. It further argues that the invariants of a socio-technical system are derivative of the personal invariants of its agents; consequently, the erosion of agent coherence and invariant compatibility may precede institutional fragmentation even while formal structures remain intact. On this basis, the article outlines two practical directions: early-warning systems based on behavioural misalignment and correlation adaptometry, and a doctrine of topological defence aimed at protecting the coherence of decision-making agents, the compatibility of their invariants, and the points of crystallisation upon which systemic cognitive security depends.

Keywords: cognitive warfare, socio-technical systems, human agent, multiplex identity, topological defence, transmorphance, personal invariants, cognitive security, reflexive control, early-warning systems, cognitive decoherence

Acknowledgements

The authors would like to express their gratitude to the NATO CCDCOE Ambassador, Dr. Sigurður Emil Pálsson, for his review, and to the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) for their invaluable support throughout the development of this research. They are also grateful to Dr David Atkinson of York St John University for valuable discussions that contributed to the refinement of the philosophical framing of this study.

3. Introduction

All five recognised domains of military operations – land, sea, air, space, and cyberspace – share a common property: sooner or later, an attack in each of them leaves an observable trace. Kinetic action produces physical destruction. An operation in space is registered by monitoring and surveillance systems. A cyberattack may remain latent for a time, yet its consequences – disruption, data exfiltration, or infrastructure failure – eventually render the fact of attack evident. Cognitive influence is deliberately deprived of this property: the targeted system does not realise – and, by design, must not realise – that it is under attack, perceiving both the attack and its consequences as the organic evolution of its own convictions. It is precisely this constitutive invisibility that distinguishes cognitive warfare from other forms of confrontation and confers upon it its strategic effectiveness.

Over the past decade, doctrinal thinking has increasingly recognised that purposeful influence upon an adversary's mind, directed at disrupting the cycle of observation, orientation, and decision-making, constitutes a distinct dimension of military conflict. In 2022, the NATO Science and Technology Organization (STO) identified the cognitive sphere as a key area of strategic research [1]; by 2025, the contest for 'cognitive superiority' had come to be recognised as a strategic challenge in its own right for the Alliance [2], whilst cognitive operations have increasingly been conceptualised as a distinct, sixth domain of operations [3]. Such institutional recognition indicates that the phenomenon has outgrown the status of a tactical instrument and now demands fundamental academic theorisation.

Yet recognition of the threat has advanced more rapidly than the capacity to counter it. Existing protective mechanisms remain predominantly content-oriented: they detect disinformation, refute false narratives, and flag unreliable sources. This approach rests on the mistaken assumption that cognitive influence consists in the insertion of false content, and that defence, accordingly, consists in its refutation. The aim of such influence, however, is not merely to alter what a person thinks, but irreversibly to alter how that person thinks [4–5]. The object of influence is the set of stable structures – whether of a system or of a human agent – that determine how information is processed and how decisions are made. Information security is blind to this level: in refuting a false message, it leaves unprotected the very architecture of perception at which the attack was directed. Existing prescriptions – the cultivation of 'critical thinking', or the promotion of information hygiene – lack a rigorous analytical apparatus capable of identifying vulnerability before the threat is realised, and are for that reason insufficient. As critics have rightly observed, to define the phenomenon solely through its effects invites 'conceptual stretching' [6], whereby a model registers what has occurred, yet fails to explain why some systems remain resilient whilst others prove vulnerable.

A response to the limitations of the existing theoretical framework was offered by the structure-centric approach developed within the research programme of the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) [7]. That approach established that the targets of cognitive warfare are systemic invariants – epistemic, axiological, identificatory, social, and teleological structures that hold a socio-technical system together as a coherent whole. Cognitive influence operates through the exploitation of inter-layer linkages within a multiplex architecture, and its strategic aim is cognitive decoherence. Yet, whilst this approach accounted for cognitive vulnerability at the level of the system, it did not disclose its foundations at the level of the human agent.

The present article extends the structure-centric approach to the modelling of the human agent in socio-technical systems. Its point of departure is that the agent cannot be reduced to an elementary, or atomic, node. It must instead be understood as a multiplex structure in which the cognitive sphere is represented by a matrix of layer-specific states, activated across different institutional and social contexts, and held in a condition of dynamic integrity by a set of personal invariants. The identity of the agent is determined not by the coincidence of its layer-specific states, but by the invariance of the structure of their mutual transformations. The mode through which identity is preserved is transmorphance – the capacity to vary layer-specific states whilst preserving the invariant core [8].

A key structural consequence follows from the proposed model: the invariants of any socio-technical system – an alliance, a state, an organisation, or an agency – are derivative of the invariants of the agents

comprised within it. Institutional forms crystallise through the interaction of agents whose personal invariants possess a high degree of compatibility. Consequently, the degradation of the personal invariants of key agents necessarily precedes the erosion of institutional cohesion, whilst the disruption of agent coherence functions as an early predictor of the subsequent decoherence of the socio-technical system as a whole.

To avoid conflating scales, a methodological qualification is required. Whereas the systemic multiplex architecture describes actors as nodes and the types of relations between them as layers, the present study describes the internal architecture of the agent itself, whose layers are constituted by the contexts of its activity. It is precisely the nesting of these scales that explains how an attack upon the agent translates into systemic decoherence.

The remainder of the article is structured as follows. Section 4 exposes the structural blindness of existing doctrines, which reduce the agent to an atomic node. Section 5 formalises the model of the agent as a multiplex structure; the full mathematical and tensorial apparatus of the model – including the matrix-tensor representation, connection, curvature, holonomy, and the formal description of cognitive influence – is set out in a separate mathematical companion [9]. Section 6 proposes a topology of cognitive attack, classifying forms of influence according to their structural targets and observable traces. Section 7 translates the model into the domain of early-warning indicators, and Section 8 formulates the principles of proactive topological defence.

4. The Reduction of the Subject: Why Standards Remain Blind

Formal models for describing socio-technical systems have attained a considerable degree of mathematical maturity; yet one fundamental reduction persists within the dominant formalisation. Even in the standard multiplex model – not to mention classical graph-theoretic and other network architectures – the agent is still represented as a single node present across all layers of the network, on the implicit assumption that its state remains static with respect to the context of interaction [10–11]. In classical multilayer network theory, the environment is rendered complex, yet the agent is not. This ontological simplification reduces the human being to an atomic component endowed with a fixed set of attributes, disregarding shifts in norms, levels of responsibility, and regimes of interpretation.

The inadequacy of this assumption is readily apparent in practice. The same officer may act, in the course of a single day, across fundamentally heterogeneous contexts: at a staff meeting, subject to the formal logic of hierarchy; when assessing intelligence, guided by the epistemic norms of the analytical community; when engaging with civilian authorities, operating under political constraints and thresholds of acceptable risk. In each case, a different configuration of cognitive strategies and motivational dispositions is activated. What is at stake is not merely a change of social role, but a transition between distinct operational regimes of the subject. A conception of the systemic agent as the bearer of a single state cannot express this internal differentiation.

It is telling that the internal plurality of the agent has long been studied outside security engineering. Sociology describes it through the divergence of role expectations [12] and through situational dramaturgy [13]. Psychology treats the psyche as a system of interacting modes [14–15], whilst cognitive science establishes the dual nature of the mechanisms of information processing [16]. Different disciplines thus converge upon a single conclusion: the human being exists as a context-differentiated structure. Yet in formal security models this plurality is not rendered as a topology linking the architecture of the agent to the architecture of the system.

A critical mismatch therefore arises between the human sciences and the science of security. In sectoral standards such as ISO 22316, ISO/IEC 27001, and the NIST Cybersecurity Framework 2.0, the human being figures predominantly as a procedural unit: trained or untrained, compliant with security policy or in breach of it [17–19]. The concept of 'identity' is reduced to technical access management – Identity and Access Management – rather than treated as a structure that synthesises values, memory, interpretation, and will. An analogous paradox is discernible in military doctrine: cognitive superiority is recognised as a strategic priority [2], yet the internal constitution of the agent whose resilience such doctrine seeks to protect remains without a formal language of description. Doctrine postulates the necessity of protection, yet lacks a structural model of what, precisely, is under attack, and by what means.

The consequences of this conceptual blindness extend far beyond the theoretical. The atomic-node model is capable of registering effects at the level of the system, but cannot detect the predictors of degradation in the structure from which those effects arise. It identifies failure post factum, whilst missing the latent transformation of the coordinates within which the agent comes to perceive trust, threat, the ally, or his own role differently. It is precisely this latent phase that constitutes the operational window of cognitive attack: influence directed at the invariant foundations of the agent demands fewer resources than an attack in any other domain, thereby realising Clausewitz's formula of imposing one's will upon the adversary at minimal cost, compelling the system to reproduce its own disintegration from within.

Consequently, external organisational and infrastructural stability ceases to be a guarantee of viability. A system may retain resources, procedures, and formal marks of governability at the very moment when its nodes have already lost the capacity to interpret reality coherently and reproduce shared invariants of action. The disintegration of the system begins with the dissolution of agent coherence. Countering this requires a transition to a new analytical framework – a model of the human agent as the bearer of multiplex identity.

5. The Topology of Identity: A Formal Model of the Agent

Overcoming the limitations of the existing paradigm requires not a further extension of current models, but a reconsideration of the very ontology of the agent. The standard representation of the agent as an atomic node endowed with a fixed set of characteristics is merely a special case of a more general reduction, in which all layer-specific states of the agent are forcibly identified with one another. This assumption is convenient for computational algorithms, but destructive for the analysis of cognitive security, since it effaces the internal architecture of the agent entirely.

Once these limitations are removed, space opens for a model capable of reflecting the complexity of the human agent. In a multiplex socio-technical system, the agent should be formalised not as a point or a vector, but as a plurality of states, each activated within a particular institutional, professional, or social context. These states form a structured configuration – the matrix of layer-specific states M_i, whose rows correspond to layers of interaction and whose columns correspond to state parameters. The identity of the agent is secured not by the coincidence of these states, but by a set of personal invariants I_i preserved across transitions between contexts.

The personal invariants I_i constitute the core of identity: stable value orientations, basic epistemic dispositions, decision-making architectures, and models of the future. Unlike layer-specific states, invariants do not depend directly upon context; rather, they determine the admissible forms of variation among states. The identity of the agent should therefore be understood not as a bijective coincidence of its states across different contexts – something structurally precluded by the multiplex nature of the system – but as the invariance of the structure of their mutual transformations. An officer who operates with different cognitive strategies at headquarters and at home does not thereby lose his identity; he loses it when the rule linking one of his states to another collapses. Invariants are precisely this rule: they do not eliminate the variability of layer-specific states, but delineate the boundaries within which the agent can adapt to different institutional contexts without loss of functional identity. Formally, the personality of the agent is the ordered pair $P_i = (M_i, I_i)$, where M_i is the totality of layer-specific states, and I_i is the invariant structure defining the admissible space of their transformations [9, Eq. (C.5)].

The measure of internal consistency is inter-layer incoherence Δ_i: high values of Δ_i indicate a misalignment of operational logics across layers. Yet inter-layer variation is not, in itself, pathological; it is the normal condition of a multiplex agent. Vulnerability arises only when such variation exceeds the admissible space and leads to the degradation of the invariant structure I_i.

This limit is formalised through the measure of transmorphance τ_i – the agent's capacity to vary layer-specific states widely whilst preserving the invariant core. Formally, $\tau_i$is the supremum of admissible inter-layer incoherence Δ_i compatible with preservation of the invariant structure [8 – 9, Eq. (C.7)]. Here τ_i functions precisely as a measure of the agent's adaptive potential, rather than as a description of its current state.

On this basis, the agent's state space may be partitioned into three regimes. In the regime of adaptive variation, Δ_i may be considerable, but remains admissible because the invariants are preserved: the agent varies from context to context, yet does so coherently. In the regime of rigidity, Δ_i is low by virtue of limited adaptivity: the agent varies his behaviour only minimally, whatever the context, preserving coherence at the cost of functional flexibility. In the regime of fragmentation, high Δ_i is accompanied by the degradation of invariants I_i: variation exceeds the admissible space, and the structure M_i loses its integrity. It is this third regime that constitutes the primary object of cognitive security, for it is here that the structural integrity of the agent is at risk of irreversible dissolution.

The model thus constructed describes an individual agent; its full significance, however, becomes apparent only when one moves to the level of the system. If each agent bears a set of personal invariants I_i, then agents whose invariants are compatible are able to form stable ties and reproducible practices. It is from

precisely this kind of interaction that the institutional layers of any socio-technical system crystallise. Formally, the systemic invariants I_S result from the aggregation of the personal invariants I_i [9, Eq. (C.8)].

The key conclusion follows: systemic invariants are derivative of personal invariants. The degradation of I_i in key agents entails the degradation of I_S with a delay determined by the inertia of organisational structures, such that the disruption of agent coherence systematically precedes the erosion of system coherence.

Herein lies the specificity, and the danger, of cognitive attacks. A kinetic attack, or a cyberattack, strikes manifestations, leaving the system's points of crystallisation intact: where agents bearing undamaged invariants persist, so too do those points of crystallisation, and with them the system's potential for recovery. An attack upon personal invariants, by contrast, erodes the foundations of the system itself, limiting its capacity to reproduce institutional coherence even after the technical infrastructure has been restored.

6. The Anatomy of Cognitive Attack

6.1. Cognitive Biases as Structural Vulnerabilities

Cognitive warfare operates with an extensive arsenal of influences directed at different classes of invariants: epistemic procedures of verification, architectures of trust, teleological orientations, and related structures [7]. The present section analyses only one mechanism through which such influences may be realised – an attack that exploits cognitive biases – since it is through this example that the general logic of cognitive influence becomes most readily apparent. This does not mean that cognitive warfare is reducible to the manipulation of biases; rather, the case serves as a particularly revealing instance.

A cognitive bias is a systematic and reproducible deviation of judgement from normative models of rationality. It arises through reliance upon heuristics – compressed rules for processing information under conditions of uncertainty, time pressure, and limited cognitive resources [20, 16]. Its character is systemic: it is not an accidental malfunction, but a predictable pattern stably reproduced under particular conditions. In the practice of military command, such heuristics are well known. Confirmation bias leads an officer to assign greater weight to intelligence that accords with an already-adopted hypothesis concerning the adversary's intent, and to discount evidence that contradicts it; the availability heuristic leads the probability of a threat to be assessed by the ease with which a recent or emotionally salient incident comes to mind. Such biases are not signs of incompetence: they are as characteristic of an experienced commander as they are of a qualified analyst.

A conclusion follows that substantially alters the approach to security. From the perspective of ecological rationality, heuristics appear not as defects of thought, but as evolutionarily consolidated algorithms for rapid and economical decision-making, effective within the environments for which they emerged [21–22]. Bias arises not because a heuristic is erroneous in itself, but because it is applied beyond its adaptive domain. In other words, this is not a 'breakdown' of the agent's cognitive apparatus, but a normal mechanism of cognitive economy that becomes a vulnerability only when the informational environment is deliberately altered.

This qualification calls for a reconsideration of doctrines of cognitive defence. The question is not how to 'eliminate biases' – a task manifestly impossible, since they are embedded in fundamental cognitive architecture – but what occurs when the environment is artificially distorted in such a way that a consolidated heuristic begins to work against the agent.

It is precisely here that the proposed model gives cognitive bias a structural reading. In terms of the geometry of transitions between the agent's contexts [9, Sec. 7], a bias may be described as a natural inhomogeneity of the agent's connection ∇_i: it renders some routes of transition cognitively 'cheaper' than others, reducing the energetic cost of information processing and inducing the agent to move along the path of least resistance. An agent in whom a given heuristic is strongly expressed possesses a specific, asymmetric geometry of transitions, in which the invariants I_i define non-equivalent routes. This inhomogeneity is not, in itself, pathological: it does not destroy the connection, but specifies its asymmetry – latent zones of reduced resistance that, under normal conditions, remain fully compatible with the integrity of the agent. Inter-layer incoherence Δ_i thereby remains within the limits of transmorphance τ_i: the structure carries its asymmetries without loss of identity. Yet it is precisely these natural zones that constitute the architecture of vulnerability towards which the adversary may direct destructive influence.

6.2. Weaponisation: The Mechanism of Cognitive Attack

The model set out above makes it possible to formulate a precise and, in a certain sense, counter-intuitive definition of cognitive attack. An attack does not create a vulnerability – it exploits one that already exists. The adversary does not implant a new defect into the agent, but directs focused influence upon a natural inhomogeneity of the connection: the latent zone of reduced resistance which, under normal conditions, remains fully compatible with the integrity of the agent.

In formal-geometric terms, this is described as a perturbation of the connection, $\nabla_i \rightarrow \nabla_i + \delta\nabla$, where $\delta\nabla$ expresses a targeted alteration of the rules by which the agent moves from one context to another [9, Eq. (C.11)]. Cognitive bias here serves not as the target, but as a 'wedge' – the point of application at which minimal informational effort on the adversary's part produces maximal structural displacement.

The mechanics of this displacement are as follows. Under directed informational pressure, the perturbation $\delta\nabla$ increases the curvature of the connection $R(\nabla_i)$, whose observable trace is the growth of inter-layer incoherence Δ_i – for example, a widening divergence between the objective content of intelligence and the analyst's tendency to interpret it in favour of an already-adopted hypothesis. So long as this divergence remains within the adaptive limit τ_i, the structure of invariants withstands the tension: the agent forms suboptimal judgements, but preserves the connection, and the agent's cognitive asymmetries remain structurally absorbed.

The aim of the attack is to drive Δ_i beyond this adaptive limit and thereby destroy the invariant structure I_i. By applying a perturbation $\delta\nabla$ precisely targeted at a specific asymmetry, the adversary artificially accelerates incoherence, generating a flow of data resonant with the agent's initial predisposition and driving the agent ever further from a coherent description of reality, along a route of lowered cognitive resistance. When the divergence between the induced logic and reality becomes critical – as it typically does under a forced change of context – the required adaptive shift abruptly exceeds τ_i. The regime of structural fragmentation then sets in: the personal invariants I_i degrade, and the capacity for rational meaning-making collapses.

From this follows the property that makes the mechanism universal rather than tied to any single type of bias. The choice of the point of influence is derivative of the topology of the target. The adversary determines which asymmetry in the given agent affords the greatest structural 'leverage', and directs a resonant perturbation into that locus, deploying the relevant arsenal of cognitive biases or other forms of influence.

The specific combination of exploitable vulnerabilities through which the attack is conducted is determined by the individual topology of the target agent's invariants and that agent's position within the system. Cognitive biases thereby function only as variable channels of influence, whilst the mechanism of identity destruction itself remains structurally invariant.

6.3. The Topology of Attack: Three Forms of Influence

The perturbation of the connection $\delta\nabla$ is not homogeneous. Depending upon the element of topology at which it is directed, it assumes different forms, each leaving its own observable trace. The distinction between these forms is not merely classificatory, but strictly operational: it separates influence upon the content of beliefs from influence upon the cognitive structure through which the agent moves between contexts. It is precisely this structure that remains the 'blind spot' of traditional information security.

The first form is an increase in transition curvature: disorientation. The perturbation renders the outcome of a transition path-dependent: the agent's behaviour across different institutional layers becomes mutually contradictory, although declared beliefs may remain formally unchanged. An officer need not alter any professed views, yet may cease to be consistent in applying them: assessments and decisions that retained their logic in one context begin to diverge radically upon transition to another. The observable trace is a sharp rise in inter-layer incoherence Δ_i.

The second form is an alteration of the holonomy of the connection: accumulative shift. Here, influence exploits closed routes of transition, or closed loops. By conducting the agent through a sequence of institutional or informational contexts and returning that agent to the initial point, the adversary brings about a condition in which the agent's state no longer returns to its original configuration. The accumulated effect of such cycles is inscribed in the topology as a stable displacement of behavioural patterns, experienced subjectively as the organic evolution of the agent's own views. The observable trace is a persistent, non-reverting shift in value patterns following passage through closed contextual loops.

The third form is the destruction of the invariant structure: fragmentation. This is the most destructive form, since it is directed at the invariant core itself. Here it is crucial to distinguish between two adjacent objects of attack.

The first is the degradation of the core of an individual agent. If the attacked agent occupies a structurally significant position – functioning as a node of high inter-layer connectivity and brokerage (betweenness centrality) – the destruction of that agent's personal invariants I_i initiates a cascading degradation of the invariants of the entire organisation, by virtue of the derivative relation between systemic and personal invariants.

The second is the destruction of invariant compatibility. Institutional coherence rests upon the topological condition $d(I_i, I_j) < \varepsilon$: the invariants of interacting agents must be sufficiently close to generate coordinated practices [9, Sec. 5]. An attack may destroy this compatibility without compromising the integrity of any individual agent taken in isolation: the invariants of each agent undergo divergent drift – a slow drawing apart – whilst the internal coherence of each structure is preserved. Each officer remains cognitively stable, consistent, and competent; yet the topological distance between their invariants is artificially driven beyond the threshold ε. As a result, the unit loses the capacity to produce joint decisions.

Thus, the object of strategic protection is not only the structural integrity of the invariant core of individual agents, but also the topology of their compatibility – a systemic property irreducible to the properties of the constituent agents and requiring its own apparatus of observation.

These forms are not mutually exclusive. A prolonged campaign is, as a rule, combined in character: disorientation, or Form 1, creates the environment in which non-reverting shifts, or Form 2, accumulate, preparing the ground for core degradation and the rupture of compatibility, or Form 3. What unites all three forms is that they act upon the structure of identity rather than upon content, and therefore remain invisible to security systems that analyse informational content alone.

6.4. The Choice of Point of Application: System, Agent, and Crystallisation

If the choice of the point of application is a function of the target's topology, then this function has two arguments: the profile of the agent and the macrostructure of the system into which that agent is integrated. Structure-centric analysis has shown that the logic of cognitive influence varies according to the type of socio-technical system: against rigidly centralised architectures tending towards functional stability, the injection of informational entropy into the dominant layer is effective; against pluralistic, evolutionarily-dynamic systems, by contrast, the rupture of inter-layer coherence becomes the more effective mode of influence [7]. The type of system thereby defines the desired structural effect – which of the forms of perturbation discussed above it is expedient for the attacker to realise.

The profile of the agent, in turn, defines the channel through which this effect is delivered. As shown in Section 6.2, one and the same result – for example, the loss of coherence in an analytical staff group – may be achieved through different natural asymmetries of the invariants I_i among its members. The intersection of these two axes forms a targeting matrix: one axis answers the question of what structural result is to be sought, determined by the type of system; the other answers through which vulnerability that result is to be produced, determined by the topology of the target.

Yet the truly strategic target is not an isolated individual agent, but the foundational invariants around which the system crystallises. Since institutional invariants are derivative of personal invariants, and since institutional connection arises where the invariants of agents are compatible [9, Eq. (C.8)], the greatest structural leverage is obtained by attacking an invariant shared across a group – the point of crystallisation around which a unit, organisation, or state has cohered into a whole.

Influence upon such an invariant does not neutralise a single node; it disrupts the condition of compatibility for a plurality of agents. The shared invariant is displaced, the topological distance between agents' invariants passes beyond the threshold ε, and the connection disintegrates not because one of its constituent nodes has been struck, but because the very foundation upon which it rested has been destroyed. The defeat of such an invariant therefore imperils not the individual transmorphance of a single agent, but the coherence of the entire system held together by that point of crystallisation.

The nodes that bear foundational invariants by no means always occupy the upper levels of the administrative hierarchy. In the terms of Social Network Analysis, they are 'key players' – nodes whose significance is determined not by formal status, but by their contribution to connectedness: either through inter-layer connectivity, or through their potential to fragment the network [23]. They may be a

sergeant or a veteran through whom the rituals of unit cohesion are reproduced; a liaison officer who ensures coordinated transitions between layers; or an informal leader whose influence rests upon competence and the capacity to translate intent into joint action [24]. The cognitive defeat of such agents is critically dangerous precisely because it deactivates nodes essential to maintaining the compatibility of invariants across the group.

6.5. A Running Example: Confirmation Bias as an Attack Vector

To demonstrate the mechanism described above in operation, this section traces the exploitation of a single cognitive bias – confirmation bias – through all three forms of structural perturbation. It is chosen only because it has been extensively described in the classical academic literature and belongs to the canonical repertoire of cognitive biases [25–26]. The example is deliberately hypothetical and refers to no specific operation; its purpose is solely to show how the logic set out above unfolds into a coherent trajectory of attack.

Consider an intelligence officer within a headquarters staff who has formed an early hypothesis concerning the adversary's intent. In the terms of the model, the officer's connection is asymmetric: the transition from data collection to planning occurs with little resistance when incoming information confirms the hypothesis, and encounters sharply increased resistance when that information contradicts it. This is a latent asymmetry, not a defect: under normal conditions, the officer makes corrections, requests additional data, and keeps inter-layer incoherence Δ_i within the limits of transmorphance τ_i. Vulnerability arises not because the officer is inclined to 'believe what is false', but because the route of transition from data to plan contains a natural inhomogeneity of resistance.

Exploiting this topology, the adversary need not construct the attack as a crude fabrication of false information. A more subtle form of influence, corresponding to Form 1, consists in constructing an environment of interpretation: selectively amplifying certain signals, delaying others, altering the order in which they arrive, increasing time pressure, and using authoritative channels through which ambiguous data come to appear as firm confirmation of the hypothesis. The officer moves along a route of artificially lowered cognitive resistance. The divergence between the accepted picture and reality grows, yet is subjectively experienced not as incoherence, but as increasing confidence. What is observed is a rise in Δ_i whilst declared beliefs remain formally unchanged.

If the influence assumes a cyclical form – the rhythmic delivery of selective reports, recurring emphases, and coordinated signals from different sources, whilst the facts themselves remain strictly accurate – Form 2, accumulative shift, is realised. Passing repeatedly through the closed loop 'data → interpretation → confirmation', the officer returns each time to the initial context cognitively altered: each cycle imperceptibly shifts the critical threshold beyond which contradictory information is still admitted for consideration. The accumulated displacement becomes fixed irreversibly – what changes is no longer an isolated judgement, but the algorithm by which evidence is weighed, that is, a basic invariant – and the officer experiences this as a growth in professional competence. The holonomy of the connection has been altered: the state of the agent no longer returns to its initial configuration, even after the external informational influence has weakened or ceased altogether.

The critical phase, corresponding to Form 3, occurs under a forced change of context – when the reality of the theatre of operations enters into undeniable contradiction with the accepted picture. A radical adaptive shift is then required; yet because the connection has been overloaded by accumulated bias, and because the invariant core I_i has been held together only by an artificially reinforced logic, the required compensatory shift abruptly exceeds the limit of adaptation τ_i. Fragmentation sets in: the invariants I_i degrade, and the capacity for rational meaning-making collapses. The officer begins to adjust reality to the accepted picture, no longer able to decentre perception without loss of cognitive integrity. This is not the result of simple disinformation, but the engineering of structural overload – driving the agent into a cul-de-sac in which confrontation with reality is structurally bound to exceed the limit of adaptation.

The individual officer is only the point of application. The true target is not that officer, but the structural connectedness of the headquarters as a whole. This connectedness rests upon the compatibility of the invariants of its members – a shared normative contour for assessing the situation, trust in sources, thresholds of acceptable risk, and a common model of the desired outcome. It is precisely the

maintenance of the topological condition $d(I_i, I_j) < \varepsilon$ that makes the group a coherent socio-technical system.

A destabilisation campaign conducted against several key officers through individually selected vulnerabilities – one through confirmation bias, another through sensitivity to authority, a third through loss aversion – may drive their modes of perception so far apart that the topological distance between their invariants exceeds the threshold of compatibility ε. The rupture also passes through the teleological layer – the shared image of the desired outcome that had previously synchronised the group's efforts. The resulting effect is unattainable either by disinformation alone or by the physical removal of individual specialists: the number of officers and their qualifications remain unchanged, yet the headquarters loses the capacity to form a coherent picture and to produce a coordinated decision. The command body has not been physically destroyed; it has been defocused, demotivated, and deprived of the common cognitive frame within which data are transformed into joint action. This is the strategic aim of cognitive attack.

7. Operationalisation: Towards Early-Warning Systems for Cognitive Threats

The proposed model has direct practical significance; at this stage, however, it does not offer ready-made metrics of cognitive vulnerability – neither universal units of measurement nor an empirical procedure capable of rigorously estimating inter-layer incoherence Δ_i or establishing the threshold of transmorphance τ_i. The development of such metrics constitutes a separate research programme. The purpose of the present section is different: to show that the formal model defines a vector of operationalisation, indicating what kinds of indicators should be sought and why, in principle, they are observable. This distinction is essential: to conflate a class of possible indicators with a completed measurement procedure would repeat the very error of premature formalisation that this study seeks to avoid.

The point of departure is that the structural quantities of the model, although not directly measurable, possess observable traces. The curvature of the connection $R(\nabla_i)$ is not directly measurable, yet its growth, other things being equal, manifests as growth in inter-layer incoherence Δ_i, which may in turn be assessed through patterns of the agent's behaviour across different institutional contexts. This shifts diagnosis from the plane of content – what the agent believes – to the plane of topology: how consistent the agent's assessments, decisions, and actions remain across transitions between contexts. At least three classes of observable indicators may be outlined, corresponding to the forms of structural influence analysed above.

The first class of indicators is intra-agent behavioural misalignment. Growth in Δ_i manifests as a widening divergence between the decisions and actions of one and the same agent across different institutional layers: assessments that previously retained strict coherence begin to contradict one another. This is the structural analogue of the early-warning signals described for critical transitions across a broad class of complex systems: the approach to a threshold reveals itself in the statistics of behaviour before the threshold is actually crossed [27]. The anticipatory signal here is not the informational content of the agent's judgements, but the growing topological inconsistency of their structure.

The second class is inter-agent misalignment. This indicator tracks the growth of topological distance $d(I_i, I_j)$ between the invariants of agents who had previously acted as a single mechanism. Divergence in situational assessments, priorities, and criteria of acceptable risk among officers who, until recently, had produced a common decision without visible effort is the observable trace of the erosion of the compatibility upon which group cohesion rests. What is essential is that the indicator registers processes of structural degradation before they manifest in a critical institutional failure.

The third class of indicators is correlation–dispersion dynamics, or adaptometry – the measurement of adaptive potential. At this level, the model converges with an existing formal apparatus [28–29], which analyses the dynamics of groups under increasing load. Correlation adaptometry has established a stable effect: in an ensemble of similar systems under a common influence, as stress increases, both the variance of parameters and the correlation between them rise simultaneously – and this dynamic precedes the overt symptoms of crisis. As the load increases further, a second threshold is reached: correlation, having attained its maximum, begins to decline, whilst variance continues to grow. This serves as a formal sign of irreversible structural fragmentation.

This apparatus, validated on biological populations and applied to the diagnosis of social stress, appears to be a natural candidate for the operationalisation of group cognitive monitoring – a rigorous analytical interface between the topology of multiplex identity and early-warning systems. Growth in correlation and variance in the responses of a group of key agents under cognitive pressure may serve as an observable marker of approach to the compatibility threshold ε, whilst passage through the second threshold may indicate its irreversible rupture. Such a transfer is, of course, not automatic: establishing the precise mathematical correspondence between the variables of adaptometry and the quantities I_i and Δ_i requires separate formalisation, and is identified here precisely as a direction for future research.

On the basis of these three classes of indicators, an architecture of early-warning systems for cognitive threats of a fundamentally new kind begins to take shape. Existing security protocols are focused on the informational layer – disinformation, propaganda, phishing. The approach proposed here, by contrast, is oriented towards monitoring adaptive potential, both of the system as a whole and of individual key agents [8]: that is, towards estimating the reserve of admissible incoherence before the threshold of transmorphance τ_i is reached, and towards monitoring the compatibility of their invariants. Such an approach shifts the focus of preventive measures from the surveillance of personnel to the monitoring of structural indicators of coherence in decision-making – the consistency of assessments and actions, rather than their content. Since disruption of agent coherence precedes disruption of system coherence, such monitoring is capable of detecting an attack at a stage when classical security metrics still register nominal stability. The practical significance of the transition from content defence to topological defence lies in the capacity to diagnose degradation where it originates – in the cognitive architecture of the agent – rather than to register it post factum, at the moment of institutional paralysis.

8. Towards Topological Defence

The proposed model does not abolish the traditional layers of information defence, but reveals their insufficiency in the face of influence directed at the coherence of agents. If a cognitive attack is a directed perturbation $\delta\nabla$, altering the rules by which the agent moves between contexts, then security architecture must include a distinct structural perimeter – topological defence – protecting both the connection that governs coherent transitions within the agent and the compatibility of invariants between agents. The object of defence is not an isolated element, such as a message or a channel, but the topology of relations itself.

This concept accords naturally with the development of current NATO doctrine. If cognitive superiority is the capacity to influence the perception and behaviour of an adversary [2], then cognitive defence cannot be reduced to the refutation of false content: it must protect the very architecture within which perception is formed. Topological defence thus serves as the structural basis of a doctrine of cognitive superiority, answering the question of which precise structure adversarial influence seeks to alter.

From this follow three functional levels of topological defence: anticipatory mapping – where is an attack possible?; the counter-reflexive loop – which form of perturbation is being realised?; and the maintenance of adaptive coherence – how can the irreversibility of perturbation be prevented?

Anticipatory mapping. Defence begins with the construction of a topological map of the structure that the adversary seeks to exploit. Here the model intersects with the theory of reflexive control – the covert transmission to an adversary of grounds for making a decision desired in advance by the initiator [30–31]. The condition of such control is knowledge of the target's filters of perception and routes of interpretation, which, in the terms of the present model, means an understanding of the system of personal invariants I_i. Influence is thereby reduced to the selection of a perturbation $\delta\nabla$ adapted to a known natural asymmetry of the agent's connection.

The defending system must therefore perform upon itself the same analytical procedure that the adversary performs upon it: the identification of the systemic invariants I_S, the personal invariants I_i of key agents and, above all, the points of crystallisation – those shared invariants around which the system has formed – together with the identification of the agent-nodes that bear them, the key players through whom the coherence of the group is reproduced (see Section 6.4). To localise the points of crystallisation is to localise the probable direction of cognitive attack. If such mapping is to possess predictive value, it must rest upon a continuously updated catalogue of typical perturbations of the connection, correlating known attack vectors with the revealed topology of the system's own vulnerabilities.

Counter-reflexive loop – the institutional core of defence. This is the capacity of the system to reconstruct the model of imposed influence and return it into the decision-making loop as an object of reflection. If early-warning indicators – growth in Δ_i or data from correlation adaptometry – register only the structural trace of an anomaly, countermeasures are activated at the moment when this anomaly is compared against the catalogue, and the model of attack is reconstructed and integrated into the practice of the agent in the form of red-team review or preventive training. In the terms of the OODA loop – Observe, Orient, Decide, Act – the counter-reflexive loop tests whether the orientation phase has been covertly channelled towards a decision predetermined from without. Boyd emphasised that victory depends upon the capacity to preserve one's own orientation whilst disrupting that of the adversary [32–33]. Modern defence must therefore verify not only the factual reliability of observations (Observe), but also the topological structure of the cognitive transition from data to decision (Orient).

The effectiveness of such reflection follows from the very nature of the attack. Structural cognitive influence remains effective only for as long as the induced distortion is experienced by the agent as a natural evolution of the agent's own views. Reconstructing the model of attack transfers the imposed logic from the category of autonomous conclusions to that of an external influence upon the decision-making process, thereby restoring the agent's control over judgement. The asymmetry of the connection that formed the path of least resistance loses its destructive force as soon as its mechanism becomes an object of conscious reflection. At the individual level, this is achieved through psychological inoculation, which trains the agent to recognise structural routes of connection distortion [34–35]; at the institutional level, through the function of critical decision audit, or Red Teaming [36].

A more radical move is also possible – reflexive inversion. Identification of the attack vector unmasks the reflexive model that the adversary has constructed of the defending side: the image of its decision-making algorithms upon which the adversary relied in initiating the influence. Having gained access to this model, the defender turns the logic of reflexive control against the initiator [30]: whilst preserving the appearance of predetermined behaviour, the defender constructs a simulated asymmetry of the connection – a structural trap.

Such a strategy can be described rigorously in terms of active inference: a theory that treats adaptive systems, from individual agents to collective structures, as minimising variational free energy – that is, the mismatch between a model of the world and incoming signals [37]. Within the present model, free energy functions as the neurodynamic equivalent of inter-layer incoherence Δ_i. By supplying the adversary with the signals that the adversary's own model expects, the defender induces that model to increase the precision weighting of false priors: free energy within the attacker's loop is reduced, and the attacker's planning system settles upon policies optimised for the imposed simulation.

Once the adversary has become institutionally drawn into this scenario, false grounds for critical decisions are produced. On this model, a sudden change of operational context by the defender should induce a sharp rise in free energy – one that is all the more difficult to compensate through Bayesian updating, the higher the precision weighting of the false model has been artificially raised. The theory predicts here a cascading increase of incoherence within the adversary's hierarchical control loops: the adversary's own analytical apparatus, calibrated to the imposed picture, becomes a source of internal disintegration. In the terms of active inference, this corresponds to a disruption of the integrity of the adversary's Markov blanket – the statistical boundary through which a system separates itself from its environment and filters incoming signals. In the terms of the present model, this boundary corresponds to the invariant core I_i, the loss of whose integrity signifies fragmentation.

This is the highest form of defensive reflection: the closure of the attack back upon the source of influence, whereby the adversary's confidence in the success of the operation turns the adversary's own analytical apparatus into an instrument of cognitive and structural disintegration.

Maintenance of adaptive coherence. The third level prevents the irreversibility of attack – the transition to structural fragmentation. The capacity to withstand an attack without loss of coherence is secured by increasing transmorphance τ_i. The ideal of a fully rational agent, invulnerable to cognitive attack, is unattainable in principle: heuristics are a fundamental part of the structure of judgement, not a removable defect. Defence is therefore built not upon the elimination of vulnerabilities, but upon the capacity to withstand their exploitation without loss of coherence.

The protection of invariant compatibility $d(I_i, I_j) < \varepsilon$ thereby becomes critical. The capacity of the agents of a system to produce a unified decision rests upon practical interdependence and trust [24], which are themselves conditioned by the compatibility of their invariants. Yet the threat to this compatibility is not always external: leaders themselves may undermine the coherence of their own system, exploiting the same topological vulnerabilities in order to retain personal power. To fragment a group for the sake of immediate manageability artificially drives its invariants beyond the threshold ε, so that short-term control is purchased at the cost of future systemic decoherence.

The imposition of an image of the future grounded in exclusivity and isolation closes the system's teleology. By excluding the 'other', it severs the variety pool upon which the system's capacity for adaptation depends and, deprived of that variety, redirects its resources not towards development but towards self-reproduction: open becoming – poiesis – degenerates into homeostasis [38]. Confined within such a teleology, the system preserves basic survival, but forfeits the potential for development.

This lends a particular significance to Weber's concept of legitimacy. Political legitimacy in Weber's sense [39] admits a direct reading in the terms of the present model: as the compatibility of invariants between the governors and the governed. Influence that drives these invariants beyond the threshold ε – whether external or originating from authority itself – undermines legitimacy not through open contestation, but by destroying the very topological condition upon which legitimacy rested.

The protection of invariant compatibility should therefore be regarded as a task comparable in importance to the protection of physical infrastructure. Infrastructure can be restored if the coherence of the system is preserved; a destroyed topology of connectedness is incomparably harder to rebuild.

Defence, in this sense, is not the conservation of immutability, but the capacity of a system to adapt deeply to its environment whilst preserving the invariant core that secures its coherence.

9. Conclusion

This study has substantiated a structural approach to cognitive security, shifting the analytical focus from content-oriented defence – the countering of disinformation and hostile narratives – to the protection of the topology of the decision-making agent. The necessity of such a shift follows from the constitutive invisibility of cognitive influence: unlike an attack in any of the recognised operational domains, it need not leave a reliable trace of attack at the level of content, and the targeted agent may experience the induced changes as the natural evolution of the agent's own views. For the security of socio-technical systems at every level – states, organisations, and governance bodies – this means overcoming a critical vulnerability: traditional doctrines centred upon protection against disinformation, fact verification, and the filtering of false narratives are, in principle, unable to protect such systems from attacks directed not at content, but at the architecture of decision-making itself.

The structural model of multiplex identity proposed here addresses this problem at both levels: that of the system and that of the human agents through which it is constituted. By describing the agent as a matrix of layer-specific states stabilised by a core of personal invariants, the model provides a rigorous conceptual apparatus for analysing covert cognitive influence. It shows that cognitive attacks exploit latent asymmetries of the connection – among them the cognitive biases of the agent – through three structurally distinct forms of perturbation: the disruption of transition coherence, the accumulative transformation of identity, and the degradation of the invariant core itself. Each form produces a directed increase in inter-layer incoherence, up to the point at which the required structural reconfigurations exceed the agent's transmorphance: the capacity to adapt to changing conditions whilst preserving identity.

It has further been shown that the invariants of a socio-technical system are derivative of the invariants of its agents. Consequently, the disruption of the coherence of personal invariants, and the erosion of their compatibility, lead necessarily to the structural disintegration of the system itself.

The operationalisation of this model opens two practical directions. The first is the development of early-warning systems for cognitive threats, drawing upon the methods of correlation adaptometry to identify latent structural tension and the erosion of compatibility at stages preceding institutional failure. The second is the formation of a doctrine of topological defence: the anticipatory mapping of invariants, functional layers, and points of crystallisation; the integration of counter-reflexive loops into the OODA decision-making cycle; and the targeted protection of the structural coherence of governance cadres and personnel. At its outer limit, such a doctrine permits not only the neutralisation of an attack, but also its reflexive inversion – the closure of the imposed logic back upon its source, whereby the adversary's own analytical apparatus, calibrated to the image it has itself projected, becomes an instrument of the adversary's cognitive and structural disintegration.

Further research should be directed towards the empirical grounding of threshold values for indicators of structural tension, and towards the integration of topological-defence algorithms into the actual circuits of state and military governance. Achieving superiority in the cognitive domain therefore requires the methodological recognition that the critical vulnerability lies not in hostile informational content, but in the latent asymmetries of the structure of the agent itself – and that the maintenance of cognitive coherence, which at the systemic level constitutes the proper object of adaptive security, must first be secured within the architecture of the human agent on which that coherence ultimately rests.

10. Bibliography

1. NATO Science & Technology Organization. (2022). 2021 highlights: NATO Science and Technology Organization. NATO. https://www.nato.int/content/dam/nato/legacy-wcm/media_pdf/2022/4/pdf/2021-NATO-STO-Highlights-web.pdf
2. NATO Science & Technology Organization. (2025). NATO Chief Scientist's report: Cognitive warfare. https://www.sto.nato.int/wp-content/uploads/chief-scientist-report-cognitive-warfare-4.pdf
3. Le Guyader, H. (2022). Cognitive domain: A sixth domain of operations. In B. Claverie, B. Prébot, N. Buchler, & F. du Cluzel (Eds.), Cognitive warfare: The future of cognitive dominance (pp. 1–5). NATO Collaboration Support Office. https://hal.science/hal-03635898v1
4. Du Cluzel, F. (2020). Cognitive warfare. NATO ACT Innovation Hub. https://innovationhub-act.org/wp-content/uploads/2023/12/20210122_CW-Final.pdf
5. NATO Allied Command Transformation. (2023, May 9). Cognitive warfare: Beyond military information support operations. https://www.act.nato.int/article/cognitive-warfare-beyond-military-information-support-operations/
6. Deppe, C., & Schaal, G. S. (2024). Cognitive warfare: A conceptual analysis of the NATO ACT cognitive warfare exploratory concept. Frontiers in Big Data, 7, Article 1452129. https://doi.org/10.3389/fdata.2024.1452129
7. Korobeynikov, F., Davydiuk, A., & Mokhor, V. (2026). Ontological foundations of cognitive warfare. NATO Cooperative Cyber Defence Centre of Excellence. https://ccdcoe.org/library/publications/ontological-foundations-of-cognitive-warfare/
8. Korobeynikov, F., & Mokhor, V. (2026). Adaptive security: Strategic principles for complex socio-technical systems. Royal Society Open Science, 13(1), Article 251481. https://doi.org/10.1098/rsos.251481
9. Korobeynikov, F. (2026). Multiplex identity: A formal model of the human agent in socio-technical systems. Mathematical companion [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.20640642
10. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. Journal of Complex Networks, 2(3), 203–271. https://doi.org/10.1093/comnet/cnu016
11. Boccaletti, S., Bianconi, G., Criado, R., del Genio, C. I., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., & Zanin, M. (2014). The structure and dynamics of multilayer networks. Physics Reports, 544(1), 1–122. https://doi.org/10.1016/j.physrep.2014.07.001
12. Merton, R. K. (1957). Social theory and social structure. Free Press.
13. Goffman, E. (1959). The presentation of self in everyday life. Doubleday.
14. Berne, E. (1961). Transactional analysis in psychotherapy. Grove Press.
15. Schwartz, R. C. (1995). Internal family systems therapy. Guilford Press.
16. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

17. International Organization for Standardization. (2017). Security and resilience – Organizational resilience – Principles and attributes (ISO 22316:2017). https://www.iso.org/standard/50053.html

18. International Organization for Standardization. (2022). Information security, cybersecurity and privacy protection – Information security management systems – Requirements (ISO/IEC 27001:2022). https://www.iso.org/standard/82875.html

19. National Institute of Standards and Technology. (2024). The NIST cybersecurity framework (CSF) 2.0 (NIST CSWP 29). https://doi.org/10.6028/NIST.CSWP.29

20. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

21. Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 103(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

22. Gigerenzer, G. (2008). Why heuristics work. Perspectives on Psychological Science, 3(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

23. Borgatti, S. P. (2006). Identifying sets of key players in a social network. Computational & Mathematical Organization Theory, 12(1), 21–34. https://doi.org/10.1007/s10588-006-7084-x

24. King, A. (2013). The combat soldier: Infantry tactics and cohesion in the twentieth and twenty-first centuries. Oxford University Press.

25. Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

26. Pohl, R. F. (Ed.). (2017). Cognitive illusions: Intriguing phenomena in thinking, judgment and memory (2nd ed.). Routledge.

27. Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. Nature, 461(7260), 53–59. https://doi.org/10.1038/nature08227

28. Sedov, K. R., Gorban, A. N., Petushkova, E. V., Manchuk, V. T., & Shalamov, E. N. (1988). Adaptometry as a method of clinical examination of the population. Vestnik of the USSR Academy of Medical Sciences, (10), 69–75.

29. Gorban, A. N., Smirnova, E. V., & Tyukina, T. A. (2010). Correlations, risk and crisis: From physiology to finance. Physica A: Statistical Mechanics and its Applications, 389(16), 3193–3217. https://doi.org/10.1016/j.physa.2010.03.035

30. Lefebvre, V. A., & Lefebvre, V. D. (1984). Reflexive control: The Soviet concept of influencing an adversary's decision making process. Science Applications.

31. Thomas, T. L. (2004). Russia's reflexive control theory and the military. Journal of Slavic Military Studies, 17(2), 237–256. https://doi.org/10.1080/13518040490450529

32. Boyd, J. R. (2018). A discourse on winning and losing (G. T. Hammond, Ed.). Air University Press.

33. Osinga, F. P. B. (2007). Science, strategy and war: The strategic theory of John Boyd. Routledge.

34. McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 1, pp. 191–229). Academic Press. https://doi.org/10.1016/S0065-2601(08)60052-0

35. Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. Palgrave Communications, 5, Article 65. https://doi.org/10.1057/s41599-019-0279-9

36. UK Ministry of Defence. (2021). Red teaming handbook (3rd ed.). Development, Concepts and Doctrine Centre. https://www.gov.uk/government/publications/a-guide-to-red-teaming

37. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787

38. Atkinson, D. M. (2023). Reimagining capitalism: Applying negative dialectics for a better future. Vernon Press.

39. Weber, M. (1978). Economy and society: An outline of interpretive sociology (G. Roth & C. Wittich, Eds.). University of California Press. (Original work published 1922).