Cognition as Perimeter: Toward a Doctrine of Information Resilience

Myrto J. Demetriou, MA

June 2026

Cognition as Perimeter: Toward a Doctrine of Information Resilience

Myrto J. Demetriou, MA — Co-Founder, Cyprus Security and Intelligence Institute (CySII)

Working Paper • Cognitive Security and Information Resilience • June 2026

Abstract

This paper proposes a framework for understanding cognitive resilience as a strategic security domain. It argues that contemporary influence operations target cognition directly, and that the defense of open societies requires extending the security perimeter inward ranging from physical territory and digital infrastructure to the cognitive substrate on which both ultimately depend. Drawing on cognitive psychology, educational science, and the inoculation and prebunking literature, the paper develops a three-layer model of cognitive defense: individual cognition, institutional cognition, and societal cognition. For each layer, the paper maps the principal vulnerabilities, the available evidence-based interventions, and the institutional capacities required to deploy them at scale. The paper concludes with a research agenda and a set of practical recommendations for states, alliances, educational systems, and intelligence and security services.

Introduction

The defense of open societies against disinformation has, for nearly a decade, been organized around the assumption that better detection systems would, in time, catch up to better manipulation. The assumption was reasonable when it was first articulated but it is becoming less reasonable every quarter. Generative artificial intelligence has driven the marginal cost of producing convincing manipulated content close to zero (Goldstein et al., 2023), while the institutional and technical capacity to detect, label, and slow such content has improved only incrementally. The economics are asymmetric, and present evidence on the structural limits of detection reliability (Sadasivan et al., 2023) gives no reason to expect them to improve. The implication, increasingly visible in policy debates across the European Union, NATO's southern flank, and the wider Atlantic community, is that

detection-first defense, however necessary, cannot by itself sustain the cognitive integrity on which open societies depend.

This paper proceeds from that diagnosis but is not principally concerned with restating it. The diagnosis parallels arguments made by Lewandowsky and van der Linden (2021), Roozenbeek and colleagues (2022), and Ecker and colleagues (2022) in the cognitive-psychology literature on misinformation and inoculation. What this paper proposes is one step further upstream. The argument advanced here is that the persistence of detection-first defense, despite increasingly clear evidence of its limits, reflects not a technical failure but a doctrinal one: open societies do not yet have a defensive doctrine that takes cognition seriously as a domain in its own right. The contribution this paper attempts to offer is one such doctrine.

The framework developed in Section 3, *Cognition as Perimeter*, organizes the existing cognitive-science literature into a three-layer model of cognitive defense: individual cognition, institutional cognition, and societal cognition. For each layer, the paper identifies the principal vulnerabilities, the available evidence-based interventions, and the institutional capacities required to deploy them at scale. The framework is doctrinal and does not produce new experimental findings; it proposes a way of recognizing and organizing existing findings such that they cohere into a defensive practice. This is a contribution that bridges literatures rather than extending any single one.

The paper therefore sits between three established literatures, none of which by itself produces the doctrine the framework proposes.

The first is the cognitive psychology of misinformation and influence. The experimental tradition that has, over four decades, produced robust findings on heuristics, motivated cognition, illusory truth, source confusion, the continued influence of corrected information, and the conditions under which inoculation builds durable resistance to manipulation (Kahneman, 2011; Hasher, Goldstein, & Toppino, 1977; Mitchell & Johnson, 2009; Pennycook & Rand, 2019; Roozenbeek & van der Linden, 2019). This literature establishes the cognitive substrate but does not, on its own terms, address the institutional and societal layers at which contemporary influence operations operate.

The second is intelligence studies. Particularly the tradition initiated by Richards Heuer's *Psychology of Intelligence Analysis* (1999) and continued in the structured-analytic-techniques literature (Heuer & Pherson, 2014) and the calibration-training research of the Good Judgment Project (Tetlock & Gardner, 2015; Mellers et al., 2014). This literature addresses the cognition of analytic institutions but treats the cognitive psychology it draws on as auxiliary rather than foundational, and it does not engage the societal layer at all.

The third is educational science. The research on how durable learning is constructed, how cognitive defenses are built through curriculum and practice, and how civic and information-literacy interventions can be designed to scale (Wineburg & McGrew, 2019; Caulfield, 2017; van der Linden, Roozenbeek, & Compton, 2020). This literature has produced the most directly applicable interventions for the societal layer but is rarely positioned within security thinking, and is frequently siloed in education ministries that do not coordinate with defense and intelligence institutions.

The framework developed in this paper is offered as one way of integrating these three literatures into a defensive doctrine that none of them alone produces. The doctrinal move is the contribution; the underlying empirical evidence belongs to fields whose authority is already established. The hypothesis advanced here is that the integration itself which is the recognition that these three literatures describe a single defensive domain across three layers, is what makes cognitive security operationally legible to the institutions whose business is the defense of open societies.

The remainder of the paper proceeds as follows. Section 2 develops, at greater depth than the inaugural commentary, the cognitive-psychology and economic evidence for the limits of detection-first defense. Section 3 develops the framework itself. Sections 4 through 6 examine the three layers of the cognitive perimeter in detail, identifying the principal vulnerabilities, the available interventions, and the institutional questions that arise at each. Section 7 turns to operationalization; the practical implications of the framework for states, alliances, educational systems, and intelligence and security services. Section 8 sets out a research agenda for the questions the framework opens but does not resolve and finally, section 9 concludes.

2. The Limits of Detection

The detection-first defensive posture against disinformation rests on an intuitive and previously defensible logic. If false content can be classified accurately, labeled visibly, slowed in its propagation, or removed from circulation before reaching large audiences, the cognitive damage of false content should be substantially reduced. The defensive architecture built around this logic, including automated detection systems, third-party fact-checking partnerships, platform-mediated labels and frictions and user-facing literacy nudges is the dominant mode in which open societies currently invest in disinformation defense.

This section argues that the logic, while not wrong in principle, encounters two structural limits that cannot be engineered around: a cognitive limit and an economic limit. The cognitive limit is the body of evidence, accumulated over four decades of experimental psychology, that detection and correction at the point of consumption operate against the strongest features of human cognition and rely on the weakest. The economic limit is the cost asymmetry between generation and detection in the generative-AI era. Both limits are structural rather than provisional; neither will be substantially reduced by technical improvement at the detection layer alone.

2.1 The cognitive limit

Three findings, each robust across decades of replication, jointly establish the cognitive limit on detection-first defense.

The ‘continued influence effect’ (Johnson & Seifert, 1994; Ecker et al., 2010; Lewandowsky et al., 2012; Ecker et al., 2022) refers to the observation that once a piece of false information has been encoded into a coherent mental model, subsequent correction

reduces its influence on inferential reasoning only partially, and sometimes negligibly. The mechanism is best understood through the cognitive architecture of mental models: information is integrated into causal-explanatory structures, and once integrated, it leaves behind inferential dependencies that persist even after the original belief is updated. Individuals who have learned a false claim, then accepted a correction, will frequently continue to make the inferences the false claim supported. The correction updates the *propositional* belief without unwinding the *causal model* that the false claim produced.

The 'illusory truth effect' (Hasher, Goldstein, & Toppino, 1977; Fazio et al., 2015; Unkelbach et al., 2019) refers to the systematic finding that repeated exposure to a statement increases its perceived accuracy, even when the statement is initially recognized as false and even when readers possess prior knowledge contradicting it. The mechanism is processing fluency: claims that have been encountered before are processed more easily, and the cognitive system misreads ease of processing as a signal of evidential weight. Detection systems that label false content but do not prevent repeated exposure leave this fluency-based mechanism intact, and the labeled exposures accumulate alongside the unlabeled ones.

The 'source monitoring failure' (Mitchell & Johnson, 2009; Loftus, 2005) refers to the differential decay rates of episodic and semantic memory: the *what* of a claim tends to persist in memory longer than the *who*, *where*, and *with what label* it carried. A user who saw a false video accompanied by a "fact-checked" label on Tuesday may, by the following week, retain the substance of the video and have lost the metadata about how it was labeled. Detection works against the strongest cognitive system which is recognition of meaningful content, and depends on the weakest which is durable source memory.

The three findings interact. A false claim that is encoded, corrected, but then encountered again is the case in which all three effects accumulate at once: the continued influence persists, the illusory truth strengthens, and the source memory fades. Detection-first defense, by design, operates inside the conditions under which these effects compound.

2.2 The economic limit

Beyond the cognitive limit lies a more recent structural problem. The cost structure of fabrication has been fundamentally altered by the diffusion of generative artificial-intelligence tools. Convincing manipulated text, image, and audio content can now be produced by actors at every level of capability, at marginal costs that approach the cost of computation alone. Recent analyses (Goldstein et al., 2023) have documented both the technical feasibility and the operational economics of generative-AI-enabled influence operations.

The detection side of the asymmetry is bound by a more demanding constraint. Detection systems must classify content they have never seen, against adversaries actively engineering against them, at the volume of platform-scale content streams, and at latency tolerances measured in seconds. The asymmetry is structural: generation requires only the production of a single instance; detection requires generalization across an open-ended adversary distribution. Nor is this asymmetry improving as recent work has documented

theoretical and practical limits to the reliability of AI-generated content detection under adversarial conditions (Sadasivan et al., 2023).

2.3 The doctrinal implication

The conclusion of this section is not that detection is unnecessary, but rather that the detection layer remains a necessary defensive component as it disrupts the distribution side of influence operations, it provides forensic data for attribution and policy response, and it imposes costs on adversaries who must engineer around it. The argument is that detection cannot, by itself, sustain the cognitive integrity of open societies under contemporary conditions.

The doctrinal implication is that defense must be added to, not substituted for, detection. The framework developed in Section 3 organizes the existing cognitive-science literature into the additional defensive layers that the limits of detection make necessary. The cognitive limit identified above indicates that defenses must reach the cognitive substrate before false content does, not after. The economic limit indicates that this is now structurally necessary rather than merely advisable.

3. Framework: Cognition as Perimeter

A defensive doctrine is, at its most basic level, a way of seeing; a way of recognizing what is being defended, where the relevant terrain lies, and what counts as defense rather than as something adjacent. Throughout the twentieth century, defensive doctrines have undergone a recurring pattern of expansion. A new domain becomes legible as defense not because the underlying terrain is new, but because the practice of defending it begins to require institutions, methods, and concepts of its own.

This paper proposes that cognition has now reached this threshold. The argument advanced here is that the defense of cognition, at the level of individuals, institutions, and societies, has become institutionally and conceptually distinct enough from adjacent activities such as information warfare, strategic communications, and counter-disinformation to warrant recognition as a defensive domain in its own right. The framework developed in this section, *Cognition as Perimeter*, is offered as one way to organize the existing cognitive-science literature into a doctrine capable of supporting that recognition.

3.1 The cognitive perimeter as defensive domain

I define the 'cognitive perimeter' as the set of cognitive processes, ranging from individual, institutional, and societal, through which information becomes belief, belief becomes inference, and inference becomes action. The perimeter is not a metaphor for cognition in general, but rather the more specific claim that there exists a definable defensive surface area, composed of the cognitive operations that translate signals from the information environment into the decisions, judgments, and behaviors of citizens, organizations, and

publics. That surface area is what contemporary influence operations target, and it is the surface area that a cognitive-security doctrine is organized to protect.

The argument for treating this surface area as a defensive domain follows the pattern by which earlier domains were recognized. Territory was always the perimeter of the state, but the practice of territorial defense was unified into a doctrine over centuries of military thought. Economic security was named as a domain over the course of the twentieth century, as the recognition crystallized that economic capacity was not merely the support for defense but a component of defense in its own right. Energy security emerged as a coherent doctrine in the 1970s, as the dependence of industrial states on stable energy supplies became politically visible. Cyber security took its current form in the early 2000s, when the digital substrate of modern life became too consequential to be defended only as part of telecommunications or intelligence. Each emergence followed the same logic: a domain previously treated as exogenous to defense became endogenous to it once the cost of leaving it undefended became apparent.

Cognitive security, on the argument advanced here, is the next such recognition. The cognitive substrate has always been the object of influence; what is new is the cost asymmetry between attacks against that substrate and the existing institutional capacity to defend it. The cost of fabricating and disseminating manipulated content has approached zero, while cognitive and institutional defenses against such content remain, by the standards of any of the earlier defensive domains, under-resourced and under-doctrinalized.

It is worth distinguishing cognitive security from several adjacent concepts with which it is sometimes conflated. 'Information warfare' is the offensive doctrine of using information to achieve strategic effect, while it shares cognitive security's subject matter but operates from the opposite side of the perimeter. 'Psychological operations' are operational techniques, typically deployed offensively or in tactical proximity to military operations, that share methodological roots with cognitive defense but differ in purpose. Strategic communications is the discipline of crafting and disseminating messages on behalf of an organization or state and is closer to persuasion than to defense. 'Counter-disinformation' describes the response activity of identifying and addressing specific false content and while it is one operational component of cognitive defense, it does not by itself constitute a doctrine.

The relationship between cognitive security and these adjacent concepts is the relationship between cyber defense and cyber attack: not the inverse activity, but the institutional domain organized around protecting the substrate that adversaries seek to exploit. The defensive doctrine has its own methods, its own institutions, and its own evidence base, drawn primarily from cognitive psychology, educational science, and the empirical study of how minds change under conditions of contested information.

3.2 Three layers

The framework distinguishes three layers of cognitive defense, each with its own subject matter, its own vulnerabilities, and its own evidence-based interventions. The layers are conceptually distinct but causally entangled, in ways developed in the following subsection.

'Individual cognition' is the layer of the citizen, analyst, voter, and decision-maker considered as a cognitively active individual. This is the layer at which most experimental cognitive psychology has been conducted, and the layer at which the densest evidence base is currently available. What is defended at this layer is the individual's capacity to encode information accurately, to update beliefs in proportion to evidence, to recognize manipulation, and to distinguish reliably between belief and inference. What is attacked at this layer is the same set of capacities, through the documented vulnerabilities of human cognition under load: heuristic substitution, motivated and identity-protective reasoning, illusory truth effects produced by fluency and repetition, source confusion produced by the differential decay of episodic and semantic memory, and the emotional arousal that drives sharing without comprehension (Kahneman, 2011; Pennycook & Rand, 2019; Fazio et al., 2015; Mitchell & Johnson, 2009; Brady et al., 2017).

'Institutional cognition' is the layer of the organization considered as a cognitive system in its own right. Intelligence services, ministries, electoral commissions, media organizations, courts, and university faculties all engage in collective reasoning processes whose properties are not reducible to the properties of their individual members. This is the layer at which the Heuer tradition in intelligence studies (Heuer, 1999; Heuer & Pherson, 2014) has operated for the past four decades, and at which the structured-analytic-techniques literature, the calibration-training research (Tetlock & Gardner, 2015; Mellers et al., 2014), and the organizational-decision-science literatures converge. What is defended at this layer is the institution's capacity for accurate collective judgment under uncertainty and time pressure. What is attacked at this layer both by adversaries, and by the institutional dynamics themselves is the same capacity, through groupthink and conformity cascades (Janis, 1972; Sunstein, 2006), anchoring under time pressure, organizational silence and the costs of dissent that produce documented intelligence failures (Bar-Joseph & McDermott, 2017), and the confirmation bias amplified by hierarchical structure.

'Societal cognition' is the layer of the public considered as a deliberative and epistemic system. Societies hold frames in common, distribute trust across institutions and individuals, and maintain shared epistemic infrastructures, such as public broadcasting, libraries, schools and courts amongst others, that allow expertise and evidence to circulate through public life. This is the layer studied by political psychology, communication science, and democratic theory. What is defended at this layer is the society's capacity to deliberate, to update collectively on evidence, and to maintain shared epistemic ground sufficient for democratic decision-making. What is attacked at this layer is the same capacity, through epistemic fragmentation across siloed information ecosystems (Benkler, Faris & Roberts, 2018), trust collapse in institutions and expertise (Hetherington & Rudolph, 2015), polarization-driven reasoning in which group identity overrides accuracy motivations (Iyengar et al., 2019; Kahan, 2017), and network amplification dynamics that reward emotional salience over evidential weight (Vosoughi, Roy & Aral, 2018; Brady et al., 2017).

3.3 The relationship between layers

The three layers are conceptually distinct but causally entangled. Each layer is partly upstream of the others and partly downstream. Individual cognition aggregates into

societal cognition through deliberation, networks, and public discourse. Institutional cognition shapes individual cognition through education, framing, and the epistemic standards institutions either uphold or erode. Societal cognition shapes institutional cognition by determining the public trust on which institutional authority depends.

This entanglement implies the methodological claim that cognitive security cannot be optimized at one layer alone, which is central to the framework. The hypothesis advanced here is that a society of well-trained individual reasoners embedded in failing institutional and societal cognition will, by different causal routes, exhibit vulnerabilities comparable to those of a society without such training, and that conversely, well-functioning institutional and societal cognition can compensate for substantial individual cognitive vulnerabilities, at least in the short term. Durable cognitive resilience, on the framework's argument, requires interventions across all three layers simultaneously, with the strongest emphasis directed at whichever layer currently functions least well.

This claim has direct consequences for policy and resource allocation, which Sections 4 through 6 develop in detail. It also has consequences for how cognitive-security research is conducted emphasizing that layer-specific findings, while valuable in their own right, cannot by themselves establish what a defensive doctrine should prioritize at a given moment in a given society. That judgment requires a cross-layer diagnostic capacity that does not yet exist in any institutional form. Developing it methodologically, empirically, and organizationally is one of the principal research-agenda items the framework opens, and will be examined in Section 8.

4. Layer One — Individual Cognition

This section examines the cognitive vulnerabilities and available interventions at the individual layer of the cognitive perimeter, defined in Section 3.2 as the citizen, analyst, voter, and decision-maker considered as a cognitively active individual. The argument advanced here is that the experimental evidence at this layer is sufficient to support specific defensive interventions at scale, that the institutional infrastructure to deploy these interventions is substantially underdeveloped, and that several open empirical questions, particularly around durability and generalization to generative-AI content, should be priority targets for cognitive-security research.

4.1 The principal vulnerabilities

The individual cognitive system is well adapted to ordinary informational environments and systematically vulnerable to environments designed to exploit its adaptations. Five vulnerabilities are most directly relevant to cognitive-security doctrine.

The first is heuristic substitution under cognitive load. Dual-process accounts of reasoning (Kahneman, 2011; Evans & Stanovich, 2013) describe a routine pattern in which individuals presented with effortful judgments substitute easier proxies including fluency, familiarity, source confidence, source congeniality. Recent empirical work (Pennycook & Rand, 2019, 2021) has demonstrated that susceptibility to fake news is better predicted by

lack of analytic engagement than by motivated reasoning. The intervention implication is consequential: even modest prompts toward deliberation can reduce misinformation belief and sharing at the individual level.

The second is motivated and identity-protective cognition (Kahan, 2017). When informational claims carry implications for group identity, individuals process them in ways that protect the identity rather than the accuracy of belief. The effect is documented across a wide range of domains and is particularly pronounced for politically polarized topics. Interventions that succeed under conditions of low identity salience may fail under conditions of high identity salience, which is a significant constraint on the generalizability of laboratory findings to politically charged contexts.

The third is illusory truth produced by repetition (Fazio et al., 2015; Unkelbach et al., 2019), discussed in Section 2.1. At the individual layer, the implication is that exposure-control interventions, meant to reduce the number of times an individual encounters false content, are foundational. No downstream intervention can fully compensate for high-volume repeated exposure.

The fourth is source confusion (Mitchell & Johnson, 2009; Loftus, 2005). The differential decay of source memory means that durable cognitive defense cannot rely on the assumption that individuals will remember which content was fact-checked, which came from credible sources, and which carried platform-level warnings. Interventions must operate at the level of recognized manipulative *forms*, not at the level of remembered metadata.

The fifth is emotional arousal and the sharing of unverified content. Work on the diffusion of moralized and emotionally charged content in online networks (Brady et al., 2017; Vosoughi, Roy & Aral, 2018) has documented systematic asymmetries in how rapidly and broadly emotionally charged content propagates. The individual-level implication is that arousal at the point of consumption is itself a vulnerability, and that interventions which restore a moment of reflective attention before sharing can reduce diffusion at the network layer.

4.2 Evidence-based interventions

Against these vulnerabilities, the experimental literature has produced four classes of intervention whose evidence base is now substantial enough to support deployment.

The first is ‘inoculation and prebunking’. The inoculation research program, whose origins can be traced to McGuire (1964), substantially revived over the past decade, and now supported by both laboratory work and large-scale field trials (Compton, 2013; Banas & Rains, 2010; Roozenbeek & van der Linden, 2019; van der Linden et al., 2020; Roozenbeek et al., 2022; Lewandowsky & van der Linden, 2021), demonstrates that brief, controlled exposure to weakened forms of manipulation, combined with explanation of the technique being used, builds durable cognitive resistance to subsequent encounters with the full-strength version. Recent field trials delivered through short video formats on major platforms have shown effects that persist for weeks and generalize across topics.

The second is 'accuracy nudges and deliberation prompts'. Pennycook and colleagues (2021) have demonstrated that briefly orienting attention toward accuracy, immediately before consumption or sharing, reduces the propagation of false content. The effect is modest in magnitude but easy to deploy at scale and complementary to other interventions.

The third is 'lateral reading and the SIFT methodology'. Research in educational psychology has documented that experienced evaluators of digital content and in particular professional fact-checkers, do not assess content by reading it more carefully. They assess it by reading laterally, leaving the source, evaluating it against external evidence, and returning only after the source's credibility has been independently established (Wineburg & McGrew, 2019). The SIFT methodology (Stop, Investigate the source, Find better coverage, Trace claims to original context) developed by Caulfield (2017) operationalizes lateral reading as a teachable practice. Both have produced positive effects in classroom interventions and are scalable through curriculum.

A fourth category, which I describe as 'deliberative scaffolding', draws on adjacent evidence rather than a single named research program. It encompasses interventions that introduce short, structured pauses between encounter and inference — breaking complex claims into sub-claims, offering quick reference to evidential standards, prompting the user to articulate why they believe what they are about to share. The evidence base is drawn from the broader literatures on deliberation prompts, dual-process reasoning, and working-memory support; its specific operationalization as cognitive-security infrastructure is one of the research-agenda items the framework opens.

4.3 Limits and research gaps

Three open questions condition the deployment of these interventions and are priority targets for cognitive-security research.

The first concerns durability. Most inoculation studies measure effects across days to weeks therefore the durability of effects across months and across multiple intervening encounters with un-inoculated content is less well established. The hypothesis advanced here, drawing on the broader educational-psychology literature on distributed practice and retrieval-based learning (Brown, Roediger & McDaniel, 2014; Karpicke, 2012), is that durability is achievable but requires recurrent rather than one-off intervention. Empirical confirmation across longer time horizons is needed.

The second concerns cross-cultural and cross-context generalization. The empirical base for individual-layer interventions is dominated by samples from North America and Western Europe. The extent to which inoculation and accuracy-prompt effects generalize to Mediterranean, Levantine, and broader regional contexts is a question of direct relevance to cognitive-security work in the East-Med region, and one to which CySII's research agenda commits.

The third concerns generative AI. Most of the experimental literature was developed before the diffusion of generative-AI tools capable of producing personalized, multimodal, and synthetic content at scale. The extent to which existing interventions generalize to AI-

generated video, voice, and personalized text is an open empirical question. Preliminary evidence suggests that the *form-based* logic of inoculation, which basically is training recognition of manipulative techniques rather than specific content, should remain effective, but this hypothesis requires direct empirical testing under contemporary conditions.

4.4 Hypothesis advanced

The hypothesis advanced in this section is that layer-one interventions are most effective when delivered as a recurrent, distributed-practice infrastructure rather than as one-off campaigns. The educational-psychology literature on durable learning (Brown, Roediger & McDaniel, 2014; Karpicke, 2012) is in broad agreement that spacing, retrieval, and transfer-conducive design are necessary for cognitive change that persists. Cognitive-security investments at the individual layer should be evaluated accordingly. The relevant question arising is not whether a campaign produced short-term effects, but whether the intervention infrastructure produces durable cognitive capacities across time.

5. Layer Two — Institutional Cognition

This section examines the cognitive vulnerabilities and available interventions at the institutional layer of the cognitive perimeter, defined in Section 3.2 as the organization considered as a collective reasoning system. The Heuer tradition has operated at this layer for the past four decades and provides much of the available intervention literature. The argument advanced here is that this tradition is now overdue for substantive integration with contemporary cognitive science, and that its operational utility will be substantially enhanced by such integration.

5.1 The principal vulnerabilities

Institutional cognition exhibits failure patterns that differ in kind, not only in degree, from the failure patterns of individual cognition. The collective reasoning of organizations is shaped by structures such as hierarchies, time pressures, norms of dissent and distributions of information, that introduce vulnerabilities the individual mind does not face in isolation. Four such vulnerabilities are central to cognitive-security doctrine at this layer.

The first is groupthink and conformity cascades. Janis's (1972) original formulation, subsequently extended in the political-psychology and organizational-decision-science literatures (Sunstein, 2006), describes how cohesive groups under pressure suppress dissent, converge on the in-group's apparent consensus, and produce decisions that no individual member would have endorsed in isolation. Information cascades (Bikhchandani, Hirshleifer & Welch, 1992) describe a related but distinct mechanism: individuals updating on others' inferred private information rather than their own evidence, with the result that group conclusions can drift far from the evidence available to any member.

The second is anchoring under time pressure. Time-pressed analytic and decision-making organizations exhibit systematic anchoring on initial assessments, with subsequent evidence updating only at the margin (Heuer, 1999). The effect is well documented in intelligence-failure case studies, where the cost of revising an established working hypothesis is high enough that contradicting evidence is reinterpreted to preserve coherence rather than to update belief.

The third is organizational silence. Bar-Joseph and McDermott (2017), drawing on case studies across multiple intelligence services, document the systematic patterns by which subordinates fail to report disconfirming information up institutional hierarchies. The costs of dissent be it career related, social or epistemic, are sufficient that even formally encouraged dissent is selectively suppressed. Intelligence failures often involve information that was available somewhere within the institution but did not reach the level at which it could change the institutional judgment.

The fourth is confirmation bias amplified by hierarchy. Individual confirmation bias is well documented; what is less appreciated in popular treatments is how hierarchical structure compounds it. Senior judgments propagate through the institution as priors for subordinate analysis; subordinate analysts who confirm senior priors are rewarded; the institutional information environment becomes selectively populated with confirming evidence. The mechanism is reinforced, not corrected, by routine bureaucratic practice unless deliberately counteracted.

5.2 Evidence-based interventions

The intervention literature at this layer is older and more institutionally embedded than at the individual layer. Four families of intervention are most clearly evidence-supported.

The first is structured analytic techniques (SATs) in the Heuer tradition (Heuer, 1999; Heuer & Pherson, 2014). Analysis of Competing Hypotheses, Key Assumptions Check, Indicators and Signposts, and related techniques operate by externalizing analytic reasoning to a format that exposes assumptions, alternatives, and inferential dependencies. The empirical evidence for SATs is mixed in its effects on outcome accuracy but consistently positive in its effects on the quality of analytic *process* in terms of assumptions surfaced, alternatives considered, reasoning made auditable. The hypothesis advanced here is that SATs are best understood not as accuracy-maximizing techniques in isolation, but as institutional infrastructure that makes analytic reasoning legible to subsequent review and external challenge.

The second is red-teaming and devil's advocacy (Zenko, 2015). Institutionally protected dissent such as analysts whose role is to argue against the prevailing institutional judgment, is one of the most consistently recommended and inconsistently implemented institutional cognitive defenses. The empirical evidence is suggestive rather than definitive, but the conceptual case is strong: institutionalizing dissent reduces the personal cost of disagreement enough that it actually occurs.

The third is calibration training and forecasting tournaments. The Good Judgment Project (Tetlock & Gardner, 2015; Mellers et al., 2014) has produced robust evidence that

probabilistic forecasting skills are trainable, that aggregated forecasts from trained individuals significantly outperform alternative methods, and that the institutional infrastructure to support calibration training is feasible. Application of these findings within intelligence and analytic institutions remains uneven.

The fourth is premortem analysis (Klein, 2007). A structured exercise in which the institution imagines that a contemplated decision has failed and reasons backward about why, the premortem produces forms of dissent that the unmodified group-decision process suppresses. It is low-cost, deployable in single meetings, and grounded in the cognitive-science literature on hindsight reasoning.

5.3 What an update to the Heuer tradition requires

The Heuer tradition, in its current form, is approximately four decades old. The cognitive science it draws on has been substantially extended since *Psychology of Intelligence Analysis* was first published. Three lines of development should be brought into a contemporary institutional-cognition doctrine.

The first is the integration of dual-process accounts of reasoning (Kahneman, 2011; Evans & Stanovich, 2013) into the analytic-techniques literature. Heuer's original treatment of cognitive bias drew on the heuristics-and-biases research as it stood at the time; the dual-process literature has since clarified when and why analytic reasoning succeeds or fails to override heuristic substitution. The implication for institutional design is that analytic structure must be calibrated to the cognitive load it imposes, analytic techniques that exceed working-memory capacity under time pressure will be silently abandoned in practice.

The second is the integration of the ecological-rationality literature (Gigerenzer, 2008). Heuristics, viewed through this literature, are not unilateral sources of error but environment-adapted strategies whose accuracy depends on the structure of the informational environment they are deployed in. An analytic institution that systematically corrects all heuristic reasoning will overcorrect; the more useful question is which heuristics are well-adapted to the contemporary intelligence environment and which are not.

The third is the integration of the debiasing literature. Sellier, Scopelliti, and Morewedge (2019) have shown that some debiasing trainings transfer to field decisions, while others do not. The implication is that not all debiasing is equally institutionally useful and that design choices matter, and institutions should evaluate the debiasing they invest in against the criterion of transfer to operational practice.

5.4 Hypothesis advanced

The hypothesis advanced in this section is that institutional-cognition interventions are most effective when embedded as ongoing institutional practice rather than as one-time training events, and when the institution maintains a measurement infrastructure for whether the practice is producing the cognitive outcomes it is intended to produce. The pattern observed in many institutions, the adoption of structured analytic techniques as a

procedural requirement, without measurement of whether they actually shift analytic outputs or whether they are silently abandoned under pressure is not cognitive defense, but cognitive theater. Doctrinal seriousness at this layer means treating institutional cognitive performance as a measurable property, and investing accordingly.

6. Layer Three — Societal Cognition

This section examines the cognitive vulnerabilities and available interventions at the societal layer of the cognitive perimeter, defined in Section 3.2 as the public considered as a deliberative and epistemic system. The argument advanced here is that this layer is the most often neglected by contemporary security doctrine, the most consequential for long-term resilience, and the layer at which the framework's central methodological claim (that cognitive security cannot be optimized at one layer alone) is most acutely visible.

6.1 The principal vulnerabilities

A society reasoning together is not simply the sum of its citizens reasoning separately. The properties of societal cognition, as in what claims can become common knowledge, how trust circulates, which institutions are credited and which discredited, what counts as evidence in public deliberation, emerge from the interaction of individual cognition, institutional cognition, and the structural features of the information ecosystem. Influence operations target these emergent properties directly, and the resulting vulnerabilities cannot be addressed by individual-layer or institutional-layer interventions alone.

The first is epistemic fragmentation across siloed information ecosystems (Benkler, Faris & Roberts, 2018). When subsets of a society draw their information from non-overlapping sources, the conditions for shared deliberation deteriorate. The vulnerability is not primarily that fragmented audiences hold different beliefs, which is possible, but that they lose the shared epistemic ground on which disagreement can be productively conducted. Adversarial influence operations are particularly effective in fragmented ecosystems because they can target sub-populations without triggering corrective signals from adjacent populations.

The second is trust collapse in institutions and expertise (Hetherington & Rudolph, 2015). When public trust in scientific, journalistic, governmental, and judicial institutions declines, the epistemic infrastructure on which durable public reasoning depends erodes. The trust literature documents a robust asymmetry between the conditions under which trust is built and the conditions under which it is lost: negative events impose disproportionate damage that subsequent accurate performance only slowly repairs (Slovic, 1993). Influence operations that seek to corrode rather than convince target trust directly, on the recognition that a populace which trusts nothing is more manipulable than a populace which trusts the wrong things.

The third is polarization-driven reasoning (Iyengar et al., 2019; Kahan, 2017). When group identity becomes the dominant frame through which informational claims are processed, accuracy motivations are systematically overridden. Polarization is not only a political

phenomenon but it is also a cognitive one, and it reshapes the layer-three terrain in which all layer-one and layer-two interventions must be deployed. The implication is that polarization itself is a cognitive-security problem, and that durable resilience at the societal layer requires attending to its underlying drivers, not only to the disinformation that exploits it.

The fourth is network amplification dynamics (Vosoughi, Roy & Aral, 2018; Brady et al., 2017). The architecture of contemporary information networks rewards emotional salience and novelty over evidential weight. False content propagates faster and farther than true content under network conditions that platform design has not yet meaningfully altered. This is a structural rather than incidental property of the current information ecosystem, and it conditions every other layer-three intervention.

6.2 Evidence-based interventions

Four families of intervention address the layer-three vulnerabilities directly.

The first is civic resilience education at scale, treated as defensive infrastructure rather than as an optional curricular enrichment. The educational-science literature on durable learning (Brown, Roediger & McDaniel, 2014; Karpicke, 2012), combined with the cognitive-science literature on inoculation, supports the design of civic-resilience curricula that build societal cognitive capacity over years rather than weeks. The hypothesis advanced here is that mandatory civic-resilience curricula at the secondary level, designed using the principles of distributed practice and transfer-conducive instruction, are among the highest-leverage investments available to states with the strategic horizon to make them.

The second is public-service inoculation campaigns through trusted broadcasters and platforms. The large-scale field trials of inoculation-based video interventions delivered through platform infrastructure (Roozenbeek et al., 2022) provide a proof of concept that inoculation can be delivered at societal scale with measurable effects. The question is whether public broadcasters and platforms treat such interventions as defensive infrastructure on a par with emergency broadcast systems, or as one-off campaigns among many.

The third is accuracy-prompt infrastructure built into platforms by design. The Pennycook et al. (2021) work on accuracy nudges suggests that platform-level redesign, making accuracy salient at moments of consumption and sharing, can reduce the propagation of false content without requiring any individual user to opt into a literacy intervention. This is the layer-three counterpart to individual-layer accuracy prompts: the same intervention, deployed as a structural property of the information environment rather than as a personal cognitive habit.

The fourth is the deliberate strengthening of shared epistemic institutions such as schools, libraries, public broadcasters, courts, independent journalism. These institutions function as common reference points whose credibility, when sustained, allows expertise and evidence to circulate across the otherwise fragmented information ecosystem. The strategic logic of cognitive defense suggests treating these institutions not as discretionary

public goods but as defensive infrastructure, with the funding stability and protection from political interference that designation implies.

6.3 The educational system as defensive infrastructure

The argument that the educational system should be reconceived as defensive infrastructure deserves explicit articulation. Education is the only institution that touches every citizen for an extended period under conditions designed for durable cognitive change. No other defensive layer has comparable reach. The educational-psychology literature on durable learning (Brown, Roediger & McDaniel, 2014; Karpicke, 2012) consistently identifies what is required: spacing, retrieval practice, transfer-conducive design, and recurrent exposure across time. Civic-resilience curricula designed on these principles, mandatory at the secondary level and reinforced at the tertiary level, are within the reach of contemporary educational systems and are not currently being prioritized as the cognitive-security investments they could be.

This argument carries an important caveat. State-supported civic-resilience education in plural democracies raises legitimacy questions that are not present in the same form at layers one and two. The state determining what counts as cognitive resilience is, structurally, the state determining what counts as legitimate cognitive disposition, and the risk of educational interventions sliding from defensive into ideologically prescriptive is real. The doctrine proposed here is that civic-resilience education should be designed around the recognition of *manipulative forms*, not around the endorsement of *particular conclusions*, and that the design process should be insulated from partisan political control through standard mechanisms of educational governance.

6.4 Limits, research gaps, and legitimacy concerns

Three open questions condition layer-three intervention and are explicit research priorities.

The first concerns adversarial interaction effects. How layer-three defensive interventions interact with sustained adversarial counter-interventions is poorly studied. Most existing evaluation of civic-resilience and inoculation programs occurs under conditions that are not actively adversarial. The durability of effects under sustained counter-pressure remains an open empirical question.

The second concerns the political economy of layer-three interventions. Who funds, designs, and delivers civic-resilience curricula and public-service inoculation campaigns is not a neutral question. Different institutional configurations produce different legitimacy properties. This is a domain in which empirical research must be paired with normative and institutional analysis.

The third concerns measurement. Layer-three outcomes in terms of societal trust, epistemic fragmentation and deliberative quality, are harder to measure than layer-one outcomes. The development of measurement infrastructure adequate to assess layer-three interventions is itself a research-agenda item, and one to which the framework explicitly commits.

6.5 Hypotheses advanced

Two hypotheses are advanced in this section. The first, parallel to those of Sections 4 and 5, is that layer-three interventions are most effective when treated as defensive infrastructure which is sustained, funded, and politically insulated rather than as periodic campaigns or discretionary public goods. The second, distinctive to this layer, is that societal cognition is the layer most often neglected by security doctrine and the most consequential for long-term resilience. A defensive posture organized around layers one and two without sustained investment in layer three is, on present evidence, insufficient. The institutions of shared epistemic life including schools, libraries, public broadcasters, courts and independent journalism amongst others, are not adjuncts to security, but under contemporary conditions, they are part of it.

7. Operationalization: From Framework to Doctrine

The framework developed in Sections 3 through 6 has direct operational implications for the institutions whose business is the defense of open societies. This section sets out those implications at four institutional levels which are states, alliances, educational systems, and intelligence and security services, and then turns to the funding architecture on which a coherent cognitive-security investment depends.

7.1 What states could do

States that take cognitive resilience seriously as a strategic priority would, at minimum, designate it as a named national-security domain, create or empower a coordinating function capable of operating across the education, communication, defense, and intelligence portfolios, and fund cognitive-resilience research at levels commensurate with other named security domains. Few states currently do any of these consistently; none, on present evidence, does all of them at the scale the framework suggests.

The specific institutional move that would most accelerate cognitive defense at the state level is the creation of a coordinating function that owns no single program but holds responsibility for the coherence of cognitive-security work across portfolios. This function would not produce content, training, or curricula, it would rather ensure that the content, training, and curricula produced by education ministries, public broadcasters, defense ministries, and intelligence services are mutually reinforcing rather than mutually contradictory. The closest existing analog is the cyber-coordination function established in many states over the past fifteen years, rendering the institutional precedent therefore available.

7.2 What alliances (EU and NATO) could do

The European Union and NATO occupy distinct but complementary positions in the cognitive-security domain. The EU's existing investments in disinformation defense materialized through a serious of measures such as the Digital Services Act, the European Democracy Action Plan, and the work of the European External Action Service's strategic-

communications task forces, constitute substantial infrastructure but remain predominantly organized around the detection layer. Reorienting a portion of this infrastructure toward upstream cognitive-resilience investment, particularly in coordination with member-state education ministries, is within reach of existing institutional capacity.

NATO's Strategic Communications Centre of Excellence has produced a substantial body of work on hybrid threats and on cognitive aspects of contemporary conflict. The extension of this work toward an explicitly defensive cognitive-security doctrine, with the same depth of institutional commitment that NATO has built into cyber defense over the past decade, would significantly advance the field. The institutional logic that cognitive resilience is to information warfare what cyber defense is to cyber-attack, is the same logic that produced NATO's cyber-defense investments.

The opportunity at the alliance level is also a diplomatic one. Joint cognitive-resilience research programs, shared assessment frameworks, and exercises that test cognitive defenses the way cyber exercises test cyber defenses would build the inter-allied infrastructure on which a coherent transatlantic cognitive-security doctrine depends.

7.3 What educational systems could do

The educational implications follow from Section 6.3. The specific operational moves are civic-resilience curricula embedded as recurrent rather than one-off components, beginning at the secondary level and reinforced at the tertiary, design of these curricula around the recognition of manipulative forms rather than the endorsement of particular conclusions; institutional protection of curricular design from partisan political control through standard mechanisms of educational governance, and the building of measurement infrastructure adequate to evaluate the cognitive outcomes the curricula are intended to produce.

Each is within the reach of contemporary educational systems but the barrier is recognition that cognitive resilience is a strategic priority on a par with literacy and numeracy, not a discretionary enrichment.

7.4 What intelligence and security services could do

Intelligence and security services have specific institutional opportunities that other actors do not. They have the analytical infrastructure to host the institutional-cognition interventions described in Section 5, they have the convening capacity to commission layer-one and layer-three interventions externally, and they have the practitioner depth to evaluate which interventions actually work under operationally meaningful conditions.

The doctrinal distinction that services should hold clearly is between defensive cognitive-resilience work and offensive psychological-operations work. The two are methodologically related and institutionally distinct. Conflating them risks both. Defensive work loses its legitimacy when it is suspected of being offensive work conducted by other means and offensive work loses its operational utility when it is constrained by defensive-legitimacy standards. Doctrinal seriousness requires organizational separation.

7.5 The funding architecture

Cognitive-resilience research currently sits across silos: education funding does not fund security research, security funding does not fund educational interventions, communication funding does not fund either. The framework's central operational implication for funders is that this funding fragmentation is itself a structural obstacle to the field.

A cognitive-resilience funding instrument perhaps at the EU or NATO level, or in the form of coordinated national instruments, that explicitly crosses the education-security-communication silos would do more to accelerate the field than any single program. The instrument need not be large in absolute terms. It would need to be deliberately structured to fund the work that does not naturally find a home in existing silos: cross-layer empirical work, institutional-cognition interventions in non-intelligence settings, and civic-resilience curricula evaluated against security outcomes.

8. A Research Agenda

The framework opens specific empirical and conceptual questions that CySII proposes to take up as the central research program of its cognitive security and information resilience stream. Six priority items follow.

The first is **durability and decay of inoculation effects** across periods longer than the existing literature has measured. Most field trials measure effects across days to weeks. The institutional implication of the framework is that durability across months and years is the relevant metric, therefore longitudinal field studies under realistic conditions are needed.

The second is **generalization to generative video and personalized synthetic media**. The existing inoculation literature was largely developed with text-based and short-video stimuli. Whether the form-recognition mechanism that drives inoculation transfers to generative video, voice cloning, and personalized synthetic content is an open empirical question with substantial practical implications.

The third is **comparative cost-effectiveness across layers**. Per citizen, what produces more durable cognitive resilience: investment at the individual layer (inoculation, lateral reading), the institutional layer (calibration, structured analytic techniques in the institutions that shape public information), or the societal layer (civic-resilience curricula, public-service inoculation campaigns)? The question has no general answer; it has answers conditional on starting conditions and intervention designs. Empirical work that compares interventions on common outcomes is needed.

The fourth is **the integration of structured analytic techniques with contemporary cognitive science of expertise and metacognition**. The Heuer tradition is overdue for substantive integration with the dual-process, ecological-rationality, and metacognition literatures developed since its founding. This work is conceptual as much as empirical.

The fifth is **the legitimacy and governance questions raised by state-supported cognitive-resilience interventions in plural democracies**. These questions are not resolvable by empirical work alone, but they require empirical and normative work conducted together. The institute commits to treating these questions as research priorities rather than as obstacles to be managed.

The sixth is **Cyprus- and East-Med-specific empirical work on cognitive vulnerability and resilience**. The empirical base for individual-layer interventions is dominated by samples from North America and Western Europe. CySII's geographic position uniquely qualifies it to extend the empirical base to the Eastern Mediterranean and the southeastern flank of the European Union, and to investigate the cross-cultural generalization questions identified in Section 4.3.

These six items constitute the explicit research program the framework opens. They are also the program against which the institute's contribution to the field will, in time, be assessed.

9. Conclusion

The argument advanced in this paper is that the defense of open societies against contemporary influence operations requires a defensive doctrine that takes cognition seriously as a domain in its own right. The framework developed here, *Cognition as Perimeter*, is offered as one such doctrine. It organizes the existing cognitive-science literature into three layers being individual, institutional, and societal, and identifies, for each layer, the principal vulnerabilities, the available evidence-based interventions, and the institutional capacities required to deploy them.

The framework's central methodological claim is that cognitive security cannot be optimized at one layer alone. Durable resilience requires interventions across all three layers, with the strongest emphasis directed at whichever layer currently functions least well. This claim has direct consequences for policy, for institutional design, and for the research agenda the field requires.

Detection-first defense remains necessary. The argument is not that detection should be abandoned, but that it has been organized to do work it cannot, by itself, do. The cognitive layer is where contemporary influence operations actually do their damage, and it is the layer where the most durable defenses are now available to be built. The evidence base for the interventions described is uneven across layers and intervention types, but for several core interventions such as inoculation and prebunking at the individual layer, calibration and structured analytic techniques at the institutional layer, accuracy-prompt infrastructure at the societal layer, it has, in our assessment, reached a threshold sufficient to justify deployment at scale, in coordination across the education, communication, and security portfolios that currently treat the cognitive substrate as someone else's concern.

The institutions of shared epistemic life, schools, libraries, public broadcasters, courts, independent journalism, analytic services as previously mentioned, are not adjuncts to

security. Under contemporary conditions, they are part of it. The doctrine proposed here is an attempt to make that recognition operationally legible, and to give the institutions charged with defending open societies a way of seeing what they are now being asked to defend.

This paper is offered as a beginning, not a conclusion. The framework's claims are framed throughout as hypotheses for empirical testing and institutional refinement. CySII commits to that program, and invites collaboration with academics, services, educational institutions, and civil-society partners willing to take the cognitive perimeter seriously.

References

- Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, 77(3). https://doi.org/10.1080/03637751003758193
- Bar-Joseph, U., & McDermott, R. (2017). *Intelligence Success and Failure: The Human Factor*. Oxford University Press.
- Benkler, Y., Faris, R., & Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.
- Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5). https://doi.org/10.1086/261849
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28). https://doi.org/10.1073/pnas.1618923114
- Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make It Stick: The Science of Successful Learning*. Harvard University Press.
- Caulfield, M. (2017). *Web Literacy for Student Fact-Checkers*. Pressbooks.
- Compton, J. (2013). Inoculation theory. In J. P. Dillard & L. Shen (Eds.), *The SAGE Handbook of Persuasion: Developments in Theory and Practice* (2nd ed.). SAGE Publications.
- Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, 38(8). https://doi.org/10.3758/mc.38.8.1087
- Ecker, U. K. H., Lewandowsky, S., Cook, J., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., Vraga, E. K., Wood, T. J., & Zaragoza, M. S. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1). https://doi.org/10.1038/s44159-021-00006-y
- Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3). https://doi.org/10.1177/1745691612460685

- Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5). https://doi.org/10.1037/xge0000098
- Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, 3(1). https://doi.org/10.1111/j.1745-6916.2008.00058.x
- Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv*. https://arxiv.org/abs/2301.04246
- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1). https://doi.org/10.1016/S0022-5371(77)80012-1
- Hetherington, M. J., & Rudolph, T. J. (2015). *Why Washington Won't Work: Polarization, Political Trust, and the Governing Crisis*. University of Chicago Press.
- Heuer, R. J. (1999). *Psychology of Intelligence Analysis*. CIA Center for the Study of Intelligence.
- Heuer, R. J., & Pherson, R. H. (2014). *Structured Analytic Techniques for Intelligence Analysis* (2nd ed.). CQ Press.
- Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22(1). https://doi.org/10.1146/annurev-polisci-051117-073034
- Janis, I. L. (1972). *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin.
- Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6). https://doi.org/10.1037/0278-7393.20.6.1420
- Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition. *Cultural Cognition Project Working Paper Series* No. 164, Yale Law School.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, 21(3). https://doi.org/10.1177/0963721412443552
- Klein, G. (2007). Performing a project premortem. *Harvard Business Review*, 85(9).
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3). https://doi.org/10.1177/1529100612451018
- Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2). https://doi.org/10.1080/10463283.2021.1876983
- Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4). https://doi.org/10.1101/lm.94705

- McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1). Academic Press.
- Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5). https://doi.org/10.1177/0956797614524255
- Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin*, 135(4). https://doi.org/10.1037/a0015849
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855). https://doi.org/10.1038/s41586-021-03344-2
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188. https://doi.org/10.1016/j.cognition.2018.06.011
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5). https://doi.org/10.1016/j.tics.2021.02.007
- Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5, 65. https://doi.org/10.1057/s41599-019-0279-9
- Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34), eabo6254. https://doi.org/10.1126/sciadv.abo6254
- Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv*. https://arxiv.org/abs/2303.11156
- Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training transfers to improve decision making in the field. *Psychological Science*, 30(9). https://doi.org/10.1177/0956797619861429
- Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk Analysis*, 13(6). https://doi.org/10.1111/j.1539-6924.1993.tb01329.x
- Sunstein, C. R. (2006). *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press.
- Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown.
- Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, 28(3). https://doi.org/10.1177/0963721419827854
- van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.566790
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380). https://doi.org/10.1126/science.aap9559

- Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, 121(11). https://doi.o