Map of a Crisis Discipline: Vulnerabilities of Human Cognition

Zubair Ansari[1] and Naveed A.A.[1]

[1]Research Division, PakCrypt NPO, Pakistan,
zubair.ansari@pakcrypt.org

Abstract. The convergence of algorithmically curated information environments, generative artificial intelligence, and industrial-scale influence operations has transformed long-known features of human cognition into systemic security vulnerabilities. This paper proposes a three-layer taxonomy—**Individual**, **Collective (Emergent)**, and **Cultural**—and catalogues thirty-nine cognitive vulnerabilities with their seminal empirical foundations and contemporary exploitation archetypes. We argue that most exploitable biases are best understood as *mismatched ancestral adaptations*—heuristics that were ecologically rational in small-group Pleistocene ecologies but that misfire in algorithmically structured information environments. Collective pathologies—cascades, polarisation, hypernormalisation—are framed as *weakly emergent* properties of populations of biased agents on engineered networks. AI systems amplify these vulnerabilities through three mechanisms: hyper-personalised persuasion, synthetic media, and engagement-optimised recommender systems that operationalise Goodhart's Law at civilisational scale. We close with a defensive architecture grounded in inoculation theory, the EU Digital Services Act and AI Act, and the NATO concept of cognitive warfare as a sixth operational domain.

Keywords: cognitive security · cognitive warfare · evolutionary mismatch · emergence · disinformation · deepfakes · generative AI · recommender systems · inoculation theory

1 Introduction

The most consequential attack surface of the early twenty-first century is not silicon but neural. Where twentieth-century information warfare targeted communication channels, contemporary adversaries target the *inferential machinery* that processes those channels—and they do so with industrial precision.

In January 2024, a finance employee at the British engineering firm Arup wired HK$200 million across fifteen transactions after a video call in which deepfaked reconstructions of his CFO and colleagues, generated from public footage, instructed him to act [29]. In March 2022, a fabricated Zelensky surrender video appeared on hacked Ukrainian outlets within weeks of Russia's invasion. In September 2023, AI-generated audio of Michal Šimečka discussing electoral

fraud spread during Slovakia's legal pre-election silence period. None of these operations would have succeeded without exploiting *specific, well-documented features of human cognition*—authority bias, fundamental attribution error, the illusory truth effect, and the negativity-weighted reactivity of System 1.

This paper argues that human cognition contains a stable inventory of vulnerabilities operating at three distinguishable layers—individual, collective-emergent, and cultural—and that AI now permits their exploitation at unprecedented scale, personalisation, and velocity. **The biases are not new; the asymmetry between attacker and defender is.** Bak-Coleman et al. [8] characterise human collective behaviour under digital communication as a crisis discipline comparable to conservation biology: poorly understood, accelerating in importance, and lacking the evidentiary base needed for stewardship.

The remainder of this paper is structured as follows. Section 2 establishes theoretical foundations. Section 3 catalogues twenty-nine individual-layer vulnerabilities. Section 4 covers six emergent collective-layer phenomena. Section 5 covers four cultural-layer vulnerabilities. Section 6 examines AI-specific amplification mechanisms. Section 7 outlines the cognitive-security policy frontier. Section 9 concludes.

2 Theoretical Foundations

2.1 Are Cognitive Biases Adaptations or Spandrels?

The debate over the evolutionary status of cognitive biases organises into three positions.

Adaptationism. Gigerenzer, Todd and the ABC Research Group [50] hold that heuristics such as Take-the-Best and the recognition heuristic outperform Bayesian models under uncertainty and small samples. Cosmides and Tooby's [31] cheater-detection module demonstrates that conditional-rule performance rises from ≈10% to 65–80% when framed as social contracts. Haselton and Nettle's [61] *Error Management Theory* generalises this: under uncertainty with asymmetric error costs, selection favours decision rules biased toward the cheaper error, producing the "paranoid optimist"—a cognitive architecture of pervasive, functional misperceptions.

Byproduct theory. Gould and Lewontin [54] provide the foundational critique of pan-adaptationism. Buss et al. [21] supply evidentiary criteria distinguishing adaptations, exaptations, and spandrels. Within cognitive science of religion, Boyer [17] and Atran [6] treat supernatural and conspiracy cognition as byproducts of evolved theory-of-mind, hyperactive agency detection, and teleological reasoning.

Evolutionary mismatch. The most productive framework here is the *evolutionary mismatch hypothesis* [82]: mechanisms calibrated to small-group, face-to-face Pleistocene ecologies misfire when scaled to algorithmic information environments

where social signals are detached from reputation, reciprocity, and embodied context. Mercier and Sperber's [87] *argumentative theory of reasoning* sharpens this: reasoning evolved for dialogical argument production and evaluation, not solitary truth-tracking. Confirmation bias and motivated reasoning are therefore *adaptive features of an interactionist mechanism* that malfunction when deployed against curated, homophilic feeds.

The mature consensus is **pluralist**: some biases satisfy strict adaptationist criteria; some are best modelled as byproducts; and many classical Kahneman–Tversky phenomena are best understood as *ancestral adaptations operating in environments they were never selected for*. This last category—mismatch—is the analytical centre of gravity for this paper.

2.2 Emergence in Collective Behaviour

The bridge from individual biases to civilisational pathology runs through emergence. Following Bedau [11] and Chalmers [23], we distinguish *weak emergence*—macro-states derivable from microdynamics only by simulation, exhibiting irreducibility in practice but not in principle—from *strong emergence*, which posits in-principle irreducibility and downward causation [71]. For the empirical phenomena treated here, **weak emergence is sufficient and philosophically defensible**.

The complex-systems lineage from Anderson's [2] "More Is Different" through Holland [63] supplies the formal toolkit: phase transitions, self-organised criticality, power-law cascades. Castellano, Fortunato and Loreto [22] apply this machinery to opinion dynamics and crowd behaviour. Bikhchandani, Hirshleifer and Welch [14] prove that even fully Bayesian agents observing predecessors' actions ignore their own private signals after a short prefix, producing fragile emergent conformity *without any individual irrationality*. Granovetter's [55] threshold model demonstrates that collective outcomes depend non-monotonically on the *distribution* of individual thresholds. Sloman and Fernbach's [120] *community of knowledge* and Hutchins's [65] distributed cognition complete the picture: cognition is properly a property of socio-technical systems, and the *illusion of explanatory depth* [109] lets individuals confuse community knowledge for personal understanding.

The pipeline is therefore: **individual cognitive mechanisms → interaction on algorithmically structured networks → emergent collective pathologies.**

3 Individual-Layer Vulnerabilities

We catalogue twenty-nine individual-layer vulnerabilities, clustering them into five functional families. Each entry provides the standardised definition, seminal evidence, and one or two AI-age exploitation archetypes.

3.1 Belief-Formation Biases

Belief formation is rarely a purely logical endeavor. Instead, it is a process of integrating information in a way that minimizes cognitive effort and maximizes internal consistency.

Confirmation Bias. Confirmation bias is the most famous of these vulnerabilities. It is the unwitting selectivity with which we seek, weight, and recall evidence that aligns with our prior beliefs. Nickerson [92] defines confirmation bias as the unwitting selectivity by which evidence is sought, weighted, and recalled in alignment with prior belief. Wason's [133] 2-4-6 rule-discovery task established the phenomenon empirically; participants generated almost exclusively confirming triples. *Exploitation:* Cambridge Analytica's psychographic micro-targeting (2014–2018) used OCEAN profiles imputed from Facebook data to deliver belief-congruent advertising [72, 85]. Internal Facebook test accounts (Haugen disclosures, 2021) showed algorithmic recommendation pushing a conservative profile into QAnon groups within two days.

Illusory Truth Effect. The "illusory truth effect" further complicates this. Hasher, Goldstein and Toppino [60] showed truth ratings rise monotonically with statement repetition; Fazio et al. [42] demonstrated this occurs even when participants demonstrably know the correct answer (*knowledge neglect*). The mechanism is fluency-mediated. *Exploitation:* Russia's "firehose of falsehood" doctrine [97] weaponises volume plus repetition across RT, Sputnik, and Telegram. AI-generated content collapses the marginal cost of repetition toward zero.

Processing Fluency / "Truthiness." Reber and Schwarz [104] showed that raising colour contrast on statements increases truth ratings, isolating fluency from repetition. McGlone and Tofighbakhsh [86] found rhyming aphorisms judged more accurate than non-rhyming paraphrases. *Exploitation:* Kreps, McCain and Brundage [74] found GPT-2-generated news rated nearly as credible as authentic articles. Nightingale and Farid [93] found StyleGAN faces more trustworthy than real ones. **LLM-generated propaganda is grammatically polished by construction**, raising fluency-based credibility independently of factual content.

Belief Bias. Evans, Barston and Pollard [41] demonstrated that syllogism-validity judgements are biased by conclusion believability. *Exploitation:* RLHF-trained chatbot sycophancy [115] reinforces this loop: LLMs tailor reasoning to user-believable conclusions, supplying motivated reasoners with technically polished "logic" for their prejudgements.

Motivated Reasoning. Kunda's [76] review established that directional motives selectively activate beliefs and inferential strategies likely to yield desired conclusions, constrained by the need for apparently reasonable justifications. *Exploitation:* Lewandowsky, Ecker and Cook [81] document organised actors supplying technical-sounding justifications that motivated audiences uncritically adopt. Bakshy, Messing and Adamic [9] found Facebook's News Feed and individual choice jointly produce ideologically congenial information diets.

Cognitive Dissonance Reduction. Festinger's [44] theory holds that inconsistent cognitions generate aversive arousal motivating rationalisation. Festinger's *When Prophecy Fails* [43] showed disconfirmed believers intensified proselytising. *Exploitation:* QAnon's repeated prophecy failures produced deeper commitment among adherents [15]; pig-butchering crypto scams (FBI IC3 2023: $4.57 billion in losses) extend victim engagement by escalating prior public commitment.

3.2 Social-Influence Heuristics

Social Proof / Bandwagon Effect. Asch's [5] line-judgement experiments produced 36.8% conformity to obviously incorrect majorities; Cialdini's [26] *Influence* operationalised social proof as a core compliance principle. *Exploitation:* The Internet Research Agency (2014–2018) used coordinated personas—@TEN_GOP (150 000 followers), Blacktivist (1.2 million followers, 11.2 million Facebook engagements)—seeded by botnets to manufacture grassroots consensus illusions [36]. Visible like counts and "trending" labels function as Cialdini-style social-proof cues at platform scale.

Authority Bias. Milgram's [88] obedience studies recorded 65% of participants administering apparent 450-volt shocks under Yale-affiliated authority. Hofling et al. [62] showed 21 of 22 nurses prepared to administer unauthorised overdoses on phone instruction from a "doctor." *Exploitation:* The **Arup deepfake CFO fraud** (January 2024)—HK$200 million across 15 transactions after a multi-person video conference reconstructing the CFO from public footage [29, 30]—is the highest-value documented case. LLM-generated expert accounts with fabricated credentials launder disinformation in election cycles [123].

Halo Effect. Thorndike [126] found commanding officers' ratings of aviation cadets on logically independent traits implausibly correlated. Nisbett and Wilson [94] showed identical lecturer features rated appealing or irritating depending on warm vs. cold behaviour. *Exploitation:* Deepfake Tesla giveaway livestreams and Edelson et al.'s [**?**] finding that X/Twitter Blue check status raises perceived credibility independent of content exemplify *halo prosthetics*—synthetic authority cues that bleed into specific judgements.

Mere-Exposure Effect. Zajonc's [137] monograph showed monotonic liking increases for repeated nonsense words and ideographs; Bornstein's [16] meta-analysis (208 experiments) gave $r = 0.26$. *Exploitation:* OpenAI's May 2024 Threat Intelligence Report documented coordinated networks (Doppelganger, Spamouflage) flooding platforms with thousands of LLM-generated repeats of identical talking points—exposure-saturation tuned for normalisation rather than persuasion.

In-Group / Out-Group Bias (Social Identity Theory). Tajfel, Billig, Bundy and Flament's [125] minimal group paradigm assigned schoolboys to "Klee" vs. "Kandinsky" groups on trivial criteria; participants sacrificed absolute in-group gain to maximise differential advantage. *Exploitation:* Thirty of 81 IRA Facebook

pages targeted African Americans; the IRA staged simultaneous pro- and anti-Muslim rallies in Houston (May 2016) [122]. Rathje, Van Bavel and van der Linden [103] found posts about the political out-group are shared roughly twice as often as in-group posts.

3.3 Attentional and Memorial Biases

Affect Heuristic / Emotional Reactivity. Finucane et al. [45] and Slovic et al. [119] demonstrated that judgements of risk and benefit derive from fast affective evaluation, producing an inverse correlation between perceived risk and benefit that *strengthens* under time pressure. *Exploitation:* Brady et al. [18] found each moral-emotional word in a tweet increased retweets by ≈20%; the 2025 large-scale replication [19] confirmed an incidence rate ratio ≈1.17. Vosoughi, Roy and Aral [132] found false news spread six times faster than truth on Twitter, attributable in part to greater fear, disgust, and surprise responses.

Variable-Reward Seeking / Dopamine System. Skinner [117] demonstrated that variable-ratio reinforcement produces the highest, most extinction-resistant response rates; Schultz, Dayan and Montague [114] localised reward-prediction-error coding to midbrain dopamine neurons. *Exploitation:* TikTok's "For You" page operates a multi-armed bandit balancing exploitation and exploration. Internal documents disclosed in a 2024 Kentucky lawsuit revealed an internal threshold of ≈260 videos for habit formation and acknowledged that screen-time tools have "negligible impact."

Novelty Bias. Berlyne's [13] collative-properties framework and Bunzeck and Düzel's [20] fMRI evidence for substantia nigra/VTA activation to absolute novelty establish the phenomenon. *Exploitation:* Vosoughi, Roy and Aral [132] explicitly tested the novelty hypothesis: false tweets had higher information-theoretic novelty (KL divergence) than true ones, and this differential statistically explained the 70% higher retweet probability.

Inattentional and Change Blindness. Simons and Chabris [116] found ≈50% of observers counting basketball passes failed to notice a gorilla-suited person walking through the scene for nine seconds. Drew, Võ and Wolfe [38] replicated this with radiologists missing a gorilla artefact on lung CT scans 83% of the time. *Exploitation:* Mathur et al. [84] catalogued 1 818 dark-pattern instances across 11 000 shopping sites, including pre-checked consent boxes and low-contrast "Reject All" buttons that exploit attentional narrowing.

Negativity Bias. Rozin and Royzman [108] decomposed the effect into negative potency, steeper gradients, dominance, and differentiation. Baumeister et al.'s [10] "Bad is stronger than good" remains canonical. *Exploitation:* Internal Facebook documents (2021) revealed that "angry" reactions were weighted five times a "like" in the engagement-prediction model from 2017 to 2019, systematically amplifying outrage-inducing content.

Recency Bias. Murdock [90] established the U-shaped serial-position curve; Glanzer and Cunitz [51] dissociated recency from primacy via distractor-task interference. *Exploitation:* The Macron Leaks dump released 48 hours before the 2017 French presidential vote during the legal media-silence period; the Slovak Šimečka deepfake audio released two days before the September 2023 election. Coordinated last-hour drops weaponise the impossibility of correction within the recency window.

Availability Heuristic. Tversky and Kahneman [127] showed participants given lists with 19 famous women among 20 less-famous men judged the famous gender more numerous, regardless of actual frequency. *Exploitation:* The IRA flooded Facebook and Instagram with vivid, emotionally graphic posts about police violence and immigrant crime, inflating perceived prevalence [36, 122].

Hindsight Bias. Fischhoff [48] found participants told an outcome assigned higher prior probabilities to whichever outcome they were told had occurred ("creeping determinism"); meta-analysis [56] gave $d = 0.39$ across 95 studies. *Exploitation:* AI-generated "explainer" videos retro-fitting catastrophes as "predictable" exploit hindsight to seed conspiracy interpretations [91].

3.4 Decision-Theoretic Biases

Cognitive Miserliness / Need for Cognitive Closure. Fiske and Taylor [47] framed humans as economical processors; Webster and Kruglanski's [134] Need for Closure Scale and the "seizing and freezing" framework formalised motivated aversion to ambiguity. *Exploitation:* Gabielkov et al. [49] found 59% of links shared on Twitter are never clicked—users share on headlines alone. AI chatbots' fluent, decisive answers satisfy closure motivation while discouraging verification [12].

Anchoring Effect. Tversky and Kahneman's [128] wheel-of-fortune study produced UN-membership estimates of 25% vs. 45% depending on a rigged anchor of 10 vs. 65. Englich, Mussweiler and Strack [40] extended this to judicial sentencing anchored by dice rolls. *Exploitation:* E-commerce dynamic pricing displays struck-through "original prices" to inflate discount perception; AI-generated first offers in negotiation chatbots anchor consumer counter-offers upward.

Framing Effect. Tversky and Kahneman's [129] Asian Disease Problem produced a dramatic preference reversal between gain framing (72% chose certain Program A) and loss framing (78% chose risky Program D) for logically equivalent options. *Exploitation:* Cambridge Analytica leaked materials describe ad creatives tailored by OCEAN profile, with neuroticism-high voters receiving loss-framed messaging.

Loss Aversion (Prospect Theory). Kahneman and Tversky [70] established that losses loom roughly 2.25× larger than equivalent gains; Ruggeri et al. [110] replicated cross-culturally across 19 countries. *Exploitation:* The FTC complaint (2023) against Amazon's "Iliad Flow" Prime cancellation maze invokes endowment-framed loss aversion; ransomware notes invoke escalating losses to push payment.

Sunk-Cost Fallacy. Arkes and Blumer's [3] Ohio University season-ticket field study showed full-price purchasers attended significantly more performances over six months. *Exploitation:* Free-to-play games and gacha mechanics leverage time- and money-sunk progression [138]; pig-butchering scams escalate "investment" requests, exploiting commitment to prior outlays.

Illusion of Control. Langer's [78] lottery-ticket study found participants who *chose* their ticket demanded a mean resale price of $8.67 vs. $1.96 for those *assigned* tickets. *Exploitation:* Robinhood's confetti animations, customisable watchlists, and stop-loss choice architecture induce illusion of control in retail traders.

Overconfidence / Dunning–Kruger Effect. Kruger and Dunning [75] found bottom-quartile performers estimated themselves at the 62nd percentile; the metacognitive deficit producing poor performance also prevents accurate self-evaluation. *Exploitation:* Motta et al. [89] found low-knowledge individuals expressed highest confidence in fringe anti-vaccine positions. Stack Overflow banned ChatGPT answers in December 2022 because low-skill users uncritically submitted confidently incorrect AI outputs.

Optimism Bias. Weinstein [135] found students rated themselves above average on virtually all positive events; Sharot et al. [**?**] demonstrated *asymmetric belief updating*—people incorporate good news more than bad. *Exploitation:* The Verizon DBIR 2024 reports 68% of breaches involve a human element; FBI IC3 2023 reports $1.14 billion in romance-scam losses. Optimism bias drove crypto FOMO into the 2022 Terra/Luna collapse (≈$40 billion in retail losses).

Cognitive Fatigue and System 1/2 Dynamics. Kahneman's [68] dual-process framework remains widely supported [98]. The specific ego-depletion construct is among the most replication-fragile in psychology: the Hagger et al. [57] Registered Replication Report (23 labs, $N = 2\,141$) and Vohs et al. [131] RRR (36 labs, $N = 3\,531$; $d \approx 0.06$) failed to detect the original effect. We retain dual-process theory and use "cognitive fatigue" operationally. *Exploitation:* Infinite-scroll platforms engineer extended sessions during which deliberative System 2 disengages; MFA-bombing fatigue attacks (2022 Uber breach, Lapsus$) deliberately exhaust users with repeated push prompts until an approval click occurs.

3.5 Self-Attribution Biases

Egocentric Bias. Ross and Sicoly [107] found married couples' self-claimed responsibility for household activities summed to over 100%, traceable to differential self-availability in memory. *Exploitation:* Character.AI and Replika companion chatbots generate affirmation tailored to user framings [115]; Salvi et al.'s [112] result that personalised GPT-4 wins 81.2% more often in debates exploits the egocentric tendency to credit congenial framings.

Fundamental Attribution Error. Jones and Harris [67] found undergraduates attributed pro- or anti-Castro attitudes to essay writers even when told the position had been *assigned* with no choice. *Exploitation:* The 2022 Zelensky surrender deepfake and the July 2024 Olena Zelenska "Bugatti" deepfake exploit the *liar's dividend* [25]—viewers infer disposition from depicted behaviour despite known situational fakery.

4 Collective-Layer (Emergent) Vulnerabilities

At the collective layer, cognitive vulnerabilities take the form of emergent group-level phenomena in which the macro-state cannot be read off individual states even when each agent is fully characterised. Six are foundational.

Pluralistic Ignorance. Coined by Allport and Katz [1] and formalised by Prentice and Miller [102], pluralistic ignorance is the second-order misperception in which a majority privately rejects a norm but publicly conforms, each falsely believing others endorse it. *Exploitation:* Authoritarian regimes exploit visible compliance to make dissent feel uniquely deviant (Kuran's [77] preference falsification). Viral pile-ons infer broad approval from visible likes and shares, suppressing dissent.

Bystander Effect / Diffusion of Responsibility. Darley and Latané's [33] smoke-filled-room study found 75% alone, 38% in groups of three, and only 10% paired with two passive confederates reported the apparent fire. Fischer et al.'s [46] meta-analysis confirmed the effect across $> 7\,700$ participants. *Exploitation:* The 2019 Christchurch mosque attack (viewed live by ≈200, re-shared 1.5 million times in 24 hours) reproduces the dynamic at platform scale. Users defer reporting assuming the algorithm or others will act; platforms diffuse responsibility to users and third-party moderators.

Authority Bias (Collective Dimension). While treated individually in Section 3, authority bias also operates as a collective dynamic: when authority cues are synthetic and widely visible, they coordinate group compliance at scale. The Milgram [88] result—65% obedience to voiced authority—acquires new force when the voiced authority is deepfaked in group video calls [29].

Spiral of Silence. Noelle-Neumann [95] developed the theory from German federal-election panel data showing last-minute swings of 3–4% toward the perceived winner. The mechanisms are a "quasi-statistical sense" of the climate of opinion, fear of isolation, and a "hard core" of resisters. *Exploitation:* Hampton et al.'s [59] Pew study ($N = 1\,801$) on the Snowden revelations found 86% willing to discuss in person but only 42% willing to post on Facebook/Twitter; online silencing *spilled over* into offline settings. China's Social Credit System operationalises the spiral by making the costs of dissent legible and quantifiable.

Groupthink and Herd Mentality. Janis [66] defined groupthink as a mode in which striving for unanimity overrides realistic appraisal, manifesting in eight symptoms including illusion of invulnerability, collective rationalisation, self-censorship, and self-appointed mindguards. *Exploitation:* QAnon's evolution (2017–2021) is a textbook groupthink fiasco: an initially loose community developed in-group invulnerability, demonised out-groups (the "cabal"), generated mindguards, and culminated in the January 6, 2021 Capitol breach.

False Consensus Effect. Ross, Greene and House [106] found Stanford undergraduates who consented to wear an "EAT AT JOE'S" sandwich board estimated 62% of peers would also consent; those who refused estimated 67% would refuse—and both groups rated opposite choosers as more diagnostic of personality. *Exploitation:* Bail et al. [7] and the Duke Polarisation Lab document Twitter's skewed activity distribution (≈10% of users produce ≈80% of political content), inflating estimates of in-party agreement and out-party disagreement.

5 Cultural-Layer Vulnerabilities

The cultural layer operates over generational and civilisational timescales, embedding individual and emergent biases in institutional architectures.

Just-World Hypothesis. Lerner's [79] synthesis of two decades of work originating in Lerner and Simmons [80] showed observers rated a confederate apparently receiving electric shocks *less* favourably the more she suffered and the less they could help. Hafer and Bègue's [58] *Psychological Bulletin* review confirmed core mechanisms. *Exploitation:* Russian state-media coverage of MH17, the Skripal poisoning, and the 2022 Bucha massacre repeatedly invoked just-world framings to position victims as deserving [101]. Sandy Hook "crisis actor" claims similarly preserve the conviction that misfortune cannot strike the virtuous [37].

Hypernormalisation. Yurchak's [136] *Everything Was Forever, Until It Was No More* (Princeton UP; winner of the Wayne Vucinich Prize) argues that the late Soviet system collapsed in a way "simultaneously completely unexpected and completely unsurprising" because its participants had long inhabited a hypernormalised reality—official discourse so formulaic that everyone recognised failure but could not imagine alternatives. *Exploitation:* Pomerantsev's [100] *Nothing Is True and Everything Is Possible* documents the simultaneous funding of liberal NGOs and ultranationalist movements—owning all sides of every argument so that no independent political reality could form. Generative AI accelerates the constative–performative decoupling Yurchak diagnosed.

Path Dependency and Status Quo Bias. David's [34] "Clio and the Economics of QWERTY" showed how a keyboard layout designed to prevent mechanical typebar jamming survived more than a century after that constraint became irrelevant. Arthur's [4] increasing-returns formalism and Samuelson and Zeckhauser's [113] status quo bias provide macro-foundations; Kahneman, Knetsch

and Thaler's [69] endowment-effect experiments (WTA exceeding WTP by 2–3×) provide micro-foundations. *Exploitation:* Network effects, data-portability barriers, and switching costs entrench dominant platforms despite documented harms. Engagement-ranking algorithms became defaults for a generation of attention infrastructure even after the Haugen disclosures demonstrated they amplify divisive content.

Goodhart's Law and Map–Territory Confusion. Goodhart [53] observed that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes"; Strathern's [124] reformulation—*when a measure becomes a target, it ceases to be a good measure*—is now canonical. Korzybski's [73] principle that **"the map is not the territory"** and the AI-alignment formalisations by Manheim and Garrabrant [83] and Skalse et al. [118] treat Goodhart as the canonical frame for proxy–target divergence. *Exploitation:* The Haugen disclosures showed Facebook's engagement-based ranking systematically amplified divisive, hateful, and misinformative content—the metric had captured the territory and inverted it. In LLMs the same pathology manifests as sycophancy and reward-model exploitation [35, 115].

6 AI-Specific Amplification Mechanisms

Four mechanisms convert latent cognitive vulnerabilities into operational exploits at unprecedented scale.

6.1 Engagement-Optimised Recommender Systems

Brady et al. [18, 19] confirmed that moral-emotional and out-group language drive disproportionate diffusion. Cinelli et al. [27] found clear homophilic clustering on algorithmically curated feeds. The empirical picture is nuanced: Hosseinmardi et al. [64] and Chen et al. [24] find limited *causal* effect of YouTube's recommender algorithm on radicalisation, with subscriptions and external links accounting for most extreme-content exposure. Nevertheless, Amnesty International's (2023, 2025) sock-puppet audits of TikTok's For You page found that within hours, approximately 50% of recommendations to simulated 13-year-old accounts were mental-health-related and potentially harmful. **The aggregate finding: the algorithm is a powerful selector for vulnerable subpopulations and for moral-emotional content.**

6.2 Generative AI Persuasion

Salvi, Horta Ribeiro, Gallotti and West [112] (preregistered, $N = 900$, *Nature Human Behaviour*) found GPT-4 with minimal sociodemographic personalisation was more persuasive than human debaters 64.4% of the time, with an 81.2% relative increase in odds of post-debate agreement ($p < 0.01$). Without personalisation, GPT-4 ≈ humans. Spitale, Biller-Andorno and Germani [121] found

GPT-3 produced both clearer accurate information *and* more compelling disinformation than humans. Goldstein et al. [52] found AI-generated propaganda judged equally persuasive as human-written. Conversely, Costello, Pennycook and Rand [32] show the dual-use property: GPT-4 personalised rebuttals reduced conspiracy belief by ≈20%, persisting at two-month follow-up.

6.3 Synthetic Media

Chesney and Citron [25] anticipated the harms; Vaccari and Chadwick [130] showed deepfakes primarily generate *uncertainty* rather than direct deception, which itself reduces trust in news. Documented incidents include the March 2022 Zelensky surrender deepfake, the September 2023 Slovak Šimečka audio, the January 2024 Biden New Hampshire robocall (FCC ruling: AI-voice robocalls illegal; Lingo Telecom $1 million civil penalty), the January 2024 Arup HK$200 million fraud, and Taylor Swift non-consensual deepfakes (one X post viewed $>$ 47 million times). DARPA's Semantic Forensics programme and the Coalition for Content Provenance and Authenticity (C2PA) anchor the detection arms race.

6.4 Personalised Targeting

Kosinski, Stillwell and Graepel [72] demonstrated that Facebook Likes predict sexual orientation at 88% accuracy. Matz et al. [85] found personality-matched ads produce up to 40% more clicks and 50% more purchases. GDPR Article 22 governs automated decision-making; the **EU Digital Services Act** (in force February 2024) bans targeted advertising to minors, uses special-category data, requires Very Large Online Platform risk assessments (Articles 34–35), and grants researcher data access (Article 40). The **EU AI Act** (Regulation 2024/1689) prohibits manipulative AI exploiting vulnerabilities (Article 5) and mandates deepfake disclosure (Article 50, in force August 2026).

7 Cognitive Security as a Policy Domain

The 2020–2025 cognitive-warfare literature crystallises around three claims. First, **cognition is a sixth operational domain** alongside land, sea, air, space, and cyber [28]. Second, stewardship of collective behaviour is a **crisis discipline** [8]. Third, **defence is possible but asymmetrically expensive**.

Inoculation / Prebunking. Roozenbeek and van der Linden's *Bad News* game and the YouTube field experiment [105] (5.4 million impressions) significantly improved manipulation-detection ($d \approx 0.16$–0.40). Effects decay over weeks, suggesting periodic re-inoculation analogous to seasonal vaccination. The DEPICT framework (Discrediting, Emotion, Polarisation, Impersonation, Conspiracy, Trolling) provides a teachable taxonomy.

Platform Architecture. Pennycook et al. [99] showed simple accuracy nudges reduced low-quality sharing in a Twitter field experiment. Lewandowsky, Ecker and Cook's [81] concept of *technocognition* calls for joint design of platforms and cognition-aware interventions. Engagement-ranking can be replaced or complemented by quality- or diversity-weighted ranking.

Regulatory Regimes. The EU DSA, AI Act, and Digital Markets Act constitute the world's most comprehensive cognitive-security infrastructure. The UK Online Safety Act (2023) and Australian eSafety basic online safety expectations provide complementary frameworks. **Critical caveats:** "cognitive warfare" as a doctrinal concept faces principled critique (Ordén [96]; Saari et al. [111]) for risking the securitisation of legitimate political discourse.

8 Relevance to Cybersecurity

Security is both a feeling and a reality, yet the two rarely align in the modern digital landscape. In the physical world, the feeling of security is based on psychological reactions to risk and countermeasures, while the reality is based on mathematical probabilities and the effectiveness of those countermeasures. For most of human history, these two facets were coupled by immediate, physical feedback. A rustle in the tall grass suggested a predator; the feeling of fear prompted the reality of flight. However, as technology has advanced, it has increasingly obscured the inner workings of our security systems, creating a fundamental divergence between how we perceive risk and the actual threats we face.

The fundamental problem is one of trade-offs. Every security decision involves a cost—whether in money, time, convenience, or freedom. Humans evolved to make these trade-offs intuitively and quickly, a skill that was ecologically rational in small-scale, face-to-face environments. However, those same intuitive mechanisms are now being exploited in an information environment structured by high-frequency algorithms and generative artificial intelligence. The result is a cognitive architecture that is increasingly mismatched with its surroundings.

For the cybersecurity professional, this means that the most consequential attack surface is no longer the silicon in our servers, but the neural pathways in our brains. While mathematical cryptography remains robust, the "inferential machinery" that processes communication is increasingly fragile. Attackers have realized that it is far easier to hack a person than to hack a firewall. This shift represents the emergence of cognitive warfare—a domain now recognized as a sixth operational theater alongside land, sea, air, space, and cyber.

To address this, we must adopt a "holistic security mindset." This involves thinking like an attacker, looking at systems not just to see how they work, but to understand how they can be made to fail–cogntively. In the age of AI, this mindset must be applied to human cognition across all three layers: Individual, Collective, and Cultural. We must recognize that our biases are not random errors, but stable features of the human operating system that can be systematically exploited.

9 Conclusion

The thirty-nine vulnerabilities catalogued here form a coherent picture once viewed through three lenses simultaneously. Evolutionarily, most are *mismatched ancestral adaptations*—heuristics that were ecologically rational in small-group Pleistocene ecologies and that misfire in algorithmically structured information environments. Philosophically, their collective consequences are *weakly emergent* properties of populations of biased agents on engineered networks. Operationally, they constitute the attack surface of a sixth security domain in which generative AI now provides the attacker with hyper-personalised persuasion, synthetic media, and engagement-optimised amplification at marginal costs approaching zero.

Three claims merit emphasis. First, **the biases themselves are not pathologies to be eliminated but features to be governed.** Confirmation bias is a feature of an adaptive interactionist reasoning system [87]; the failure mode is its deployment against curated, homophilic feeds. Defensive design should restore the ecological conditions that the architecture presupposes.

Second, **cultural-layer vulnerabilities deserve disproportionate attention** because they scaffold the others. Goodhart's Law guarantees that any optimisation target operationalised at platform scale will diverge from human flourishing; path dependency locks in the resulting architectures; hypernormalisation saps the imaginative capacity to demand alternatives; just-world reasoning naturalises the harms.

Third, **the empirical caveats matter.** Cambridge Analytica's specific electoral causal impact remains unproven [39]; ego depletion failed to replicate; YouTube's recommender algorithm appears less causally powerful than once claimed; deepfake election-flipping causal claims remain unsubstantiated. Honest cognitive security requires honest epistemics about its own evidence base.

We close with the framing of Bak-Coleman et al. [8]: the global ecology of human collective behaviour is now the object of a crisis discipline—like climate science, operating with incomplete data on systems that cannot be paused for study, where the costs of inaction may be catastrophic and irreversible. The thirty-nine vulnerabilities catalogued here are the *inventory of features that any cognitive-security architecture must respect.* The problem is the asymmetry between an attacker armed with generative AI and a defender armed with a Pleistocene mind. Closing that asymmetry—through inoculation, architectural reform, regulation, and cultural repair—is the central political-epistemic task of the next decade.

Bibliography

[1] Allport, F.H., Katz, D.: Students' attitudes. *Craftsman Press* (1931)

[2] Anderson, P.W.: More is different. *Science* **177**(4047), 393–396 (1972)

[3] Arkes, H.R., Blumer, C.: The psychology of sunk cost. *Organizational Behavior and Human Decision Processes* **35**(1), 124–140 (1985)

[4] Arthur, W.B.: Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal* **99**(394), 116–131 (1989)

[5] Asch, S.E.: Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs* **70**(9), 1–70 (1956)

[6] Atran, S.: *In Gods We Trust: The Evolutionary Landscape of Religion*. Oxford University Press, Oxford (2002)

[7] Bail, C.A., et al.: Exposure to opposing views on social media can increase political polarization. *PNAS* **115**(37), 9216–9221 (2018)

[8] Bak-Coleman, J.B., et al.: Stewardship of global collective behavior. *PNAS* **118**(27), e2025764118 (2021)

[9] Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**(6239), 1130–1132 (2015)

[10] Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D.: Bad is stronger than good. *Review of General Psychology* **5**(4), 323–370 (2001)

[11] Bedau, M.A.: Weak emergence. *Philosophical Perspectives* **11**, 375–399 (1997)

[12] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots. In: *FAccT 2021*, pp. 610–623. ACM (2021)

[13] Berlyne, D.E.: *Conflict, Arousal and Curiosity*. McGraw-Hill, New York (1960)

[14] Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* **100**(5), 992–1026 (1992)

[15] Bloom, M., Moskalenko, S.: *Pastels and Pedophiles: Inside the Mind of QAnon*. Stanford UP, Stanford (2021)

[16] Bornstein, R.F.: Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin* **106**(2), 265–289 (1989)

[17] Boyer, P.: *Religion Explained*. Basic Books, New York (2001)

[18] Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., Van Bavel, J.J.: Emotion shapes the diffusion of moralized content in social networks. *PNAS* **114**(28), 7313–7318 (2017)

[19] Brady, W.J., Rathje, S., Globig, L.K., Van Bavel, J.J.: Estimating the effect size of moral contagion in online networks. *PNAS Nexus* **4**(11), pgaf327 (2025)

[20] Bunzeck, N., Düzel, E.: Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron* **51**(3), 369–379 (2006)

[21] Buss, D.M., Haselton, M.G., Shackelford, T.K., Bleske, A.L., Wakefield, J.C.: Adaptations, exaptations, and spandrels. *American Psychologist* **53**(5), 533–548 (1998)

[22] Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. *Reviews of Modern Physics* **81**(2), 591–646 (2009)
[23] Chalmers, D.J.: Strong and weak emergence. In: Clayton, P., Davies, P. (eds.) *The Re-Emergence of Emergence*, pp. 244–256. Oxford UP, Oxford (2006)
[24] Chen, A., et al.: Subscriptions and external links help drive resentful users to radical content. *PNAS Nexus* **2**(1), pgac186 (2023)
[25] Chesney, R., Citron, D.: Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review* **107**(6), 1753–1819 (2019)
[26] Cialdini, R.B.: *Influence: The Psychology of Persuasion*. Morrow, New York (1984)
[27] Cinelli, M., et al.: The echo chamber effect on social media. *PNAS* **118**(9), e2023301118 (2021)
[28] Claverie, B., du Cluzel, F.: *Cognitive Warfare: The Future of Cognitive Dominance*. NATO STO, Brussels (2022)
[29] CNN: Finance worker pays out $25 million after video call with deepfake "chief financial officer." CNN Business (4 February 2024), https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk
[30] CFO Dive: Scammers siphon $25M from engineering firm Arup via AI deepfake 'CFO.' CFO Dive (2024), https://www.cfodive.com/news/scammers-siphon-25m-engineering-firm-arup-deepfake-cfo-ai/716501/
[31] Cosmides, L., Tooby, J.: Cognitive adaptations for social exchange. In: Barkow, J.H., Cosmides, L., Tooby, J. (eds.) *The Adapted Mind*, pp. 163–228. Oxford UP, New York (1992)
[32] Costello, T.H., Pennycook, G., Rand, D.G.: Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**(6714), eadq1814 (2024)
[33] Darley, J.M., Latané, B.: Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* **8**(4), 377–383 (1968)
[34] David, P.A.: Clio and the economics of QWERTY. *American Economic Review* **75**(2), 332–337 (1985)
[35] Denison, C., et al.: Sycophancy to subterfuge: Investigating reward tampering in language models. *arXiv preprint* arXiv:2406.10162 (2024)
[36] DiResta, R., et al.: The tactics and tropes of the Internet Research Agency. New Knowledge/Senate Select Committee on Intelligence (2018)
[37] Douglas, K.M., Sutton, R.M., Cichocka, A.: The psychology of conspiracy theories. *Current Directions in Psychological Science* **26**(6), 538–542 (2017)
[38] Drew, T., Võ, M.L.H., Wolfe, J.M.: The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science* **24**(9), 1848–1853 (2013)
[39] Eady, G., et al.: Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications* **14**(1), 62 (2023)
[40] Englich, B., Mussweiler, T., Strack, F.: Playing dice with criminal sentences. *Personality and Social Psychology Bulletin* **32**(2), 188–200 (2006)

[41] Evans, J.St.B.T., Barston, J.L., Pollard, P.: On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition* **11**(3), 295–306 (1983)
[42] Fazio, L.K., Brashier, N.M., Payne, B.K., Marsh, E.J.: Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General* **144**(5), 993–1002 (2015)
[43] Festinger, L., Riecken, H.W., Schachter, S.: *When Prophecy Fails*. University of Minnesota Press, Minneapolis (1956)
[44] Festinger, L.: *A Theory of Cognitive Dissonance*. Stanford UP, Stanford (1957)
[45] Finucane, M.L., Alhakami, A., Slovic, P., Johnson, S.M.: The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making* **13**(1), 1–17 (2000)
[46] Fischer, P., et al.: The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin* **137**(4), 517–537 (2011)
[47] Fiske, S.T., Taylor, S.E.: *Social Cognition*. Addison-Wesley, Reading (1984)
[48] Fischhoff, B.: Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance* **1**(3), 288–299 (1975)
[49] Gabielkov, M., Ramachandran, A., Chaintreau, A., Legout, A.: Social clicks: What and who gets read on Twitter? *ACM SIGMETRICS Performance Evaluation Review* **44**(1), 179–192 (2016)
[50] Gigerenzer, G., Todd, P.M., ABC Research Group: *Simple Heuristics That Make Us Smart*. Oxford UP, New York (1999)
[51] Glanzer, M., Cunitz, A.R.: Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior* **5**(4), 351–360 (1966)
[52] Goldstein, J.A., et al.: How persuasive is AI-generated propaganda? *PNAS Nexus* **3**(2), pgae034 (2024)
[53] Goodhart, C.A.E.: Problems of monetary management: The U.K. experience. In: *Papers in Monetary Economics*, vol. I. Reserve Bank of Australia (1975)
[54] Gould, S.J., Lewontin, R.C.: The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society B* **205**(1161), 581–598 (1979)
[55] Granovetter, M.: Threshold models of collective behavior. *American Journal of Sociology* **83**(6), 1420–1443 (1978)
[56] Guilbault, R.L., Bryant, F.B., Brockway, J.H., Posavac, E.J.: A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology* **26**(2–3), 103–117 (2004)
[57] Hagger, M.S., et al.: A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science* **11**(4), 546–573 (2016)
[58] Hafer, C.L., Bègue, L.: Experimental research on just-world theory. *Psychological Bulletin* **131**(1), 128–167 (2005)
[59] Hampton, K.N., Rainie, L., Lu, W., Dwyer, M., Shin, I., Purcell, K.: Social media and the spiral of silence. Pew Research Center (2014)

[60] Hasher, L., Goldstein, D., Toppino, T.: Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior* **16**(1), 107–112 (1977)
[61] Haselton, M.G., Nettle, D.: The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review* **10**(1), 47–66 (2006)
[62] Hofling, C.K., Brotzman, E., Dalrymple, S., Graves, N., Pierce, C.M.: An experimental study in nurse–physician relationships. *Journal of Nervous and Mental Disease* **143**(2), 171–180 (1966)
[63] Holland, J.H.: *Emergence: From Chaos to Order*. Addison-Wesley, Reading (1998)
[64] Hosseinmardi, H., et al.: Examining the consumption of radical content on YouTube. *PNAS* **118**(32), e2101967118 (2021)
[65] Hutchins, E.: *Cognition in the Wild*. MIT Press, Cambridge (1995)
[66] Janis, I.L.: *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*, 2nd edn. Houghton Mifflin, Boston (1982)
[67] Jones, E.E., Harris, V.A.: The attribution of attitudes. *Journal of Experimental Social Psychology* **3**(1), 1–24 (1967)
[68] Kahneman, D.: *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York (2011)
[69] Kahneman, D., Knetsch, J.L., Thaler, R.H.: Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* **5**(1), 193–206 (1991)
[70] Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. *Econometrica* **47**(2), 263–292 (1979)
[71] Kim, J.: Making sense of emergence. *Philosophical Studies* **95**(1–2), 3–36 (1999)
[72] Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. *PNAS* **110**(15), 5802–5805 (2013)
[73] Korzybski, A.: *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. International Non-Aristotelian Library Publishing, Lakeville (1933)
[74] Kreps, S., McCain, R.M., Brundage, M.: All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science* **9**(1), 104–117 (2022)
[75] Kruger, J., Dunning, D.: Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* **77**(6), 1121–1134 (1999)
[76] Kunda, Z.: The case for motivated reasoning. *Psychological Bulletin* **108**(3), 480–498 (1990)
[77] Kuran, T.: *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard UP, Cambridge (1995)
[78] Langer, E.J.: The illusion of control. *Journal of Personality and Social Psychology* **32**(2), 311–328 (1975)

[79] Lerner, M.J.: *The Belief in a Just World: A Fundamental Delusion*. Plenum Press, New York (1980)
[80] Lerner, M.J., Simmons, C.H.: Observer's reaction to the "innocent victim." *Journal of Personality and Social Psychology* **4**(2), 203–210 (1966)
[81] Lewandowsky, S., Ecker, U.K.H., Cook, J.: Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition* **6**(4), 353–369 (2017)
[82] Li, N.P., van Vugt, M., Colarelli, S.M.: The evolutionary mismatch hypothesis. *Current Directions in Psychological Science* **27**(1), 38–44 (2018)
[83] Manheim, D., Garrabrant, S.: Categorizing variants of Goodhart's law. *arXiv preprint* arXiv:1803.04585 (2018)
[84] Mathur, A., et al.: Dark patterns at scale: Findings from a crawl of 11K shopping websites. *ACM CSCW* **3**(CSCW), 1–32 (2019)
[85] Matz, S.C., Kosinski, M., Navé, G., Stillwell, D.J.: Psychological targeting as an effective approach to digital mass persuasion. *PNAS* **114**(48), 12714–12719 (2017)
[86] McGlone, M.S., Tofighbakhsh, J.: Birds of a feather flock conjointly (?). *Psychological Science* **11**(5), 424–428 (2000)
[87] Mercier, H., Sperber, D.: *The Enigma of Reason*. Harvard UP, Cambridge (2017)
[88] Milgram, S.: *Obedience to Authority*. Harper & Row, New York (1974)
[89] Motta, M., Callaghan, T., Sylvester, S.: Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Social Science & Medicine* **211**, 274–281 (2018)
[90] Murdock, B.B.: The serial position effect of free recall. *Journal of Experimental Psychology* **64**(5), 482–488 (1962)
[91] NewsGuard: Unreliable AI news websites grow 1,000% in 2023. NewsGuard Technologies (2024)
[92] Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* **2**(2), 175–220 (1998)
[93] Nightingale, S.J., Farid, H.: AI-synthesized faces are indistinguishable from real faces and more trustworthy. *PNAS* **119**(8), e2120481119 (2022)
[94] Nisbett, R.E., Wilson, T.D.: The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology* **35**(4), 250–256 (1977)
[95] Noelle-Neumann, E.: *The Spiral of Silence: Public Opinion—Our Social Skin*. University of Chicago Press, Chicago (1984)
[96] Ordén, H.: Cognitive warfare and the securitisation of knowledge. *Security Dialogue* (2024). https://doi.org/10.1177/09670106241233985
[97] Paul, C., Matthews, M.: The Russian "firehose of falsehood" propaganda model. RAND Corporation PE-198-OSD (2016)
[98] Pennycook, G., et al.: Cognitive style and the misattribution of conspiracy theories, fake news, and pseudoscience. *Cognition* **188**, 26–35 (2019)
[99] Pennycook, G., et al.: Shifting attention to accuracy can reduce misinformation online. *Nature* **592**(7855), 590–595 (2021)

[100] Pomerantsev, P.: *Nothing Is True and Everything Is Possible*. PublicAffairs, New York (2014)
[101] Pomerantsev, P.: *This Is Not Propaganda*. PublicAffairs, New York (2019)
[102] Prentice, D.A., Miller, D.T.: Pluralistic ignorance and alcohol use on campus. *Journal of Personality and Social Psychology* **64**(2), 243–256 (1993)
[103] Rathje, S., Van Bavel, J.J., van der Linden, S.: Out-group animosity drives engagement on social media. *PNAS* **118**(26), e2024292118 (2021)
[104] Reber, R., Schwarz, N.: Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition* **8**(3), 338–342 (1999)
[105] Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., Lewandowsky, S.: Susceptibility to misinformation is consistent across question format and response mode. *Science Advances* **8**(25), eabo6254 (2022)
[106] Ross, L., Greene, D., House, P.: The false consensus effect. *Journal of Experimental Social Psychology* **13**(3), 279–301 (1977)
[107] Ross, M., Sicoly, F.: Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology* **37**(3), 322–336 (1979)
[108] Rozin, P., Royzman, E.B.: Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review* **5**(4), 296–320 (2001)
[109] Rozenblit, L., Keil, F.: The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* **26**(5), 521–562 (2002)
[110] Ruggeri, K., et al.: Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour* **4**(6), 622–633 (2020)
[111] Saari, A., Häkkinen, T., Moilanen, T.: Cognitive warfare: Doctrinal tensions and democratic risks. *Journal of Strategic Studies* (2024). https://doi.org/10.1080/01402390.2024.2312123
[112] Salvi, C., Horta Ribeiro, M., Gallotti, R., West, R.: On the conversational persuasiveness of GPT-4. *Nature Human Behaviour* (2025). https://doi.org/10.1038/s41562-025-02107-z
[113] Samuelson, W., Zeckhauser, R.: Status quo bias in decision making. *Journal of Risk and Uncertainty* **1**(1), 7–59 (1988)
[114] Schultz, W., Dayan, P., Montague, P.R.: A neural substrate of prediction and reward. *Science* **275**(5306), 1593–1599 (1997)
[115] Sharma, M., et al.: Towards understanding sycophancy in language models. *arXiv preprint* arXiv:2310.13548 (2024)
[116] Simons, D.J., Chabris, C.F.: Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* **28**(9), 1059–1074 (1999)
[117] Skinner, B.F.: *Science and Human Behavior*. Macmillan, New York (1953)
[118] Skalse, J., Howe, N., Krasheninnikov, D., Krueger, D.: Defining and characterizing reward hacking. In: *NeurIPS 2022*. Curran Associates (2022)
[119] Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G.: The affect heuristic. *European Journal of Operational Research* **177**(3), 1333–1352 (2007)
[120] Sloman, S., Fernbach, P.: *The Knowledge Illusion*. Riverhead Books, New York (2017)
[121] Spitale, G., Biller-Andorno, N., Germani, F.: AI model GPT-3 (dis)informs us better than humans. *Science Advances* **9**(26), eadh1850 (2023)

[122] Senate Select Committee on Intelligence: Report on Russian active measures campaigns and interference in the 2016 US election, vol. 2. US Senate (2019)
[123] Stanford Internet Observatory: How AI bots and fake social media accounts are deceiving voters worldwide. SIO Technical Report (2023)
[124] Strathern, M.: "Improving ratings": Audit in the British university system. *European Review* **5**(3), 305–321 (1997)
[125] Tajfel, H., Billig, M.G., Bundy, R.P., Flament, C.: Social categorization and intergroup behaviour. *European Journal of Social Psychology* **1**(2), 149–178 (1971)
[126] Thorndike, E.L.: A constant error in psychological ratings. *Journal of Applied Psychology* **4**(1), 25–29 (1920)
[127] Tversky, A., Kahneman, D.: Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* **5**(2), 207–232 (1973)
[128] Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157), 1124–1131 (1974)
[129] Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. *Science* **211**(4481), 453–458 (1981)
[130] Vaccari, C., Chadwick, A.: Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* **6**(1) (2020)
[131] Vohs, K.D., et al.: A multi-site preregistered paradigmatic test of the ego-depletion effect. *Psychological Science* **32**(10), 1566–1581 (2021)
[132] Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**(6380), 1146–1151 (2018)
[133] Wason, P.C.: On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology* **12**(3), 129–140 (1960)
[134] Webster, D.M., Kruglanski, A.W.: Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology* **67**(6), 1049–1062 (1994)
[135] Weinstein, N.D.: Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* **39**(5), 806–820 (1980)
[136] Yurchak, A.: *Everything Was Forever, Until It Was No More: The Last Soviet Generation*. Princeton UP, Princeton (2005)
[137] Zajonc, R.B.: Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monograph Supplement* **9**(2, Pt.2), 1–27 (1968)
[138] Zendle, D., Cairns, P.: Video game loot boxes are linked to problem gambling. *PLoS ONE* **13**(11), e0206767 (2018)