Defending Minds, Not Just Machines: An Agentic Approach to Cognitive Security*

Roberto González, Jaime Fúster, Michele Gucciardo, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco†

NEC Labs Europe, Heidelberg, Germany
{roberto.gonzalez, jaime.fuster, michele.gucciardo, davide.sanvito, giuseppe.siracusano, roberto.bifulco}@neclab.eu

Abstract

We define *cognitive security* as the protection of human perception, judgment, and decision-making from manipulation, overload, and engineered uncertainty across digital channels and organizational contexts. Rather than replacing traditional technical controls, cognitive security complements them by focusing on human outcomes—helping people notice what matters, reason under pressure, and act in ways that are explainable and reversible.

Delivering this protection is a complex socio-technical challenge that requires multi-disciplinary work spanning security engineering, human-computer interaction (HCI), cyber threat intelligence (CTI), psychology, organizational behavior, governance, and law. Within this broader effort, we argue that generative-AI (GenAI) agents are powerful enablers: they can continuously watch information flows, synthesize and contextualize evidence, and support decision-making with timely, human-aligned assistance when their roles are clearly scoped.

Cognitive-security failures often stem from a tempo mismatch: machines demand continuous vigilance while humans excel at careful, contextual judgment—leading to overload, missed cues, and brittle responses. To address this, we propose a role-based taxonomy that separates speed from care and makes cognitive delegation explicit. **Sentinels** are always-on monitors that surface calibrated cues with minimal friction. **Advisors** are on-demand sensemakers that assemble and explain the most relevant evidence. **Executors** apply narrowly scoped, reversible actions under explicit bounds. **Keepers** maintain posture over time by tracking configuration, drift, and hygiene. **Stewards** provide overall governance—ensuring policy compliance, privacy handling, provenance, auditability, and human override. This composition aligns technical capability with human limitations and institutional constraints, reducing overload while preserving accountability, and provides a practical blueprint for deploying GenAI to protect societies, organizations, and citizens.

1 Introduction

Security has long focused on protecting machines and data, yet in today's threat landscape the decisive target is increasingly the human mind. Adversaries shape what people perceive, believe, and decide—through coordinated disinformation, realistic voice and video deepfakes, spear-phishing, manipulative interface patterns, and the chronic overload that blunts judgment. This phenomenon is often described as cognitive hacking [1]. We use the term *cognitive security* to denote the broader objective: protecting human perception, judgment, and decision-making

*Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. S5, December 16-18, 2025, Sapporo , Japan.

†Corresponding author

from manipulation and overload across digital channels and contexts. This applies at population, community/organizational, and individual scales.

Russia's Operation Overload represents a very good example of a large-scale, multi-platform Foreign Information Manipulation and Interference (FIMI) campaign targeting global audience. A recent study[1] has shown how hundreds of Telegram channels are used as a central hub for internal propaganda, targeting Russian speakers, while the same anti-Ukraine narratives are then reposted by more than 10 thousand fake accounts on X, Bluesky, and, most recently, TikTok, targeting an international audience. The key tactic consists of overwhelming organizations and individuals–particularly in the media and research sectors–with false narratives, draining resources from fact checking and favoring the spreading of misleading information. The focus of the operation are geopolitical targets including, and not limited to, France, Germany, Moldova, Poland, Ukraine, and USA.

In this context, NATO has expressed concern over Cognitive Warfare as the ultimate battlefield, one that targets human rationality through disinformation campaigns aimed at undermining the capacity to distinguish fact from fiction[2]. The European Union has identified protection against hybrid threats to the Member States, including FIMI, as one of the pillars of its strategy plan for internal security, ProtectEU[3].

On the technical side, designing effective cognitive-security solutions represents a significant challenge and it is inherently multidisciplinary. Beyond computer science and security engineering, it requires expertise from psychology and cognitive science (biases, attention, persuasion), human–computer interaction (explanations, choice architecture, dark patterns), communication and media studies (narratives, provenance), law and policy (privacy, consent, platform rules), and ethics. Our aim is not to replace those disciplines but to integrate their insights into operational defenses that respect human limits and institutional constraints.

Position. Cognitive delegation is the systematic offloading of routine detection and preliminary sensemaking to background processes, which then return *calibrated summaries, options, and reversible actions* at the human decision point. We need it because modern defense is dominated by tempo mismatches and overload: machines require continuous vigilance while humans excel at contextual, accountable judgment. Delegation reduces missed cues and fatigue by letting automation scan broadly and continuously, while reserving scarce human attention for high-stakes choices—*with uncertainty made explicit and control preserved.*

We therefore argue for an *agentic* solution that cleanly separates roles for vigilance, judgment, action, upkeep, and governance so systems become fast where they must be fast and careful where they must be careful. Generative-AI–based agents are a strong enabler because they can mediate natural language, compose tools and data sources, and coordinate workflows end-to-end. In practice, a GenAI agent can plan and call specialized detectors (e.g., phishing, deepfake, anomaly), CTI and provenance services, retrieval/verification pipelines, rule engines, and domain APIs—then summarize evidence, quantify uncertainty, and propose safe, auditable interventions. Other implementations (e.g., lighter local models or symbolic planners) may be appropriate in some settings, but our focus is on GenAI-enabled composition because it integrates the heterogeneous technologies defenders already rely on.

Why now. The cost of generating persuasive, targeted, multi-modal content has collapsed, while the volume and velocity of signals saturate attention. Citizens face feeds optimized for en-

[1] https://checkfirst.network/wp-content/uploads/2025/06/Overload%C2%A02_%20Main%20Draft%20Report_compressed.pdf
[2] https://www.act.nato.int/activities/cognitive-warfare/
[3] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52025DC0148

gagement; organizations contend with coordinated influence and rushed approvals; individuals receive "urgent" prompts that exploit bias and fatigue. A monolithic assistant is ill-suited; what is needed is a composition of roles that reduces overload, exposes uncertainty, and translates intent into safe, reversible action.

Agent roles. We propose a compact taxonomy. *Sentinels* run continuously to perform thousands of simple checks and surface low-friction signals without interrupting users. *Advisors* are invoked on demand to synthesize evidence, weigh trade-offs, and recommend next steps with clear explanations and uncertainty. *Executors* perform bounded, policy-conformant actions when triggers or approvals are met, shrinking time-to-mitigate while preserving reversibility and traceability. *Keepers* monitor posture and configuration (not human interactions); when they detect vulnerabilities or risky drift, they inform the operator and may propose a ready-to-run Executor. *Stewards* act as the "police" of the ecosystem, enforcing privacy, licensing, provenance, and audit across all agents so every decision and data flow remains compliant and reviewable.

Contributions. This paper makes four contributions toward operationalizing cognitive security with GenAI-enabled agents:

- **C1 Problem framing & requirements.** We formalize *cognitive security* as protecting human perception, judgment, and decision-making from manipulation and overload at population, organizational, and individual scales, and derive requirements for *cognitive delegation* with a *human-aligned interface* (low-friction cues, explicit uncertainty, reversible action).
- **C2 Role-based agent taxonomy.** We introduce a compact, operational taxonomy—*Sentinels* (continuous checks and calibrated signals), *Advisors* (on-demand sensemaking and explanation), *Executors* (bounded, auditable actions), *Keepers* (posture, configuration, and drift monitoring), and *Stewards* (policy, privacy, provenance, and audit). We specify each role's inputs/outputs, triggers, and interaction points so speed is applied where needed and care where required.
- **C3 Cognitive limitations to defense mapping.** We catalogue common cognitive limitations exploited in attacks (e.g., bounded attention, confirmation/authority/anchoring effects, habituation) and map them to role-specific mitigations (e.g., Sentinel calibration to reduce overload, Advisor sourcing to counter automation bias, Executor reversibility to prevent escalation).
- **C4 Application sketches & case narrative.** Using representative scenarios—including disinformation, fraud/phishing, deepfakes, dark patterns, and operational triage—and a compact narrative of the XZ Utils incident, we illustrate how the roles intervene along a *cognitive kill chain* to surface anomalies earlier, compress time-to-sense, and enable safe, reversible responses within existing workflows.

2 State of the art

In this section, we present a summary of relevant literature that can be used to support the use of GenAI agents as a defensive layer against cognitive security attacks. We first evaluate relevant works on cognitive hacking, then we analyze the current state of the art of LLM assistants.

2.1 Cognitive hacking

Cognitive hacking – i.e., the manipulation of a person's perception or mental state to influence their actions, beliefs, or decisions to achieve a specific outcome – is not a novel concept in cybersecurity. The term was first defined in 2002 by the authors of [1], who distinguished between two models: *covert* cognitive hacking, which aims to conceal the attack, and *overt* cognitive hacking, which openly alters or forges legitimate communications (e.g., defacement or spoofing) to influence the user. Based on these models, recent studies map attack surfaces, evaluate impacts, and assess countermeasures.

Drawing on dual-process theories from psychology, the authors of [2] argue that, although most current information systems' security theories assume a rational actor making deliberate decisions, such deliberate thinking is not as common as is usually expected. They show that behavioral responses to security threats like phishing often emerge from instantaneous pattern-matching and person-context specific heuristics (System 1 cognition), rather than conscious deliberation (System 2) [3, 4]. Hence, many security theories misread user behavior by assuming slow, deliberate decisions, while many security-relevant actions are, in fact, fast and automatic.

Recent work shows that users most vulnerable to phishing often limit cognitive effort and rely on superficial cues (e.g., treating "www" as a safety signal), while low-salience browser indicators continue to fail [5, 6, 7]. Vulnerability also appears domain-general: susceptibility to phishing, scam texts, and fake-news headlines is correlated across individuals [8]. Meanwhile, common defenses have limited success against internet-based social-engineering attacks [9], and LLM-based attack pipelines now automate and sharpen techniques such as spear-phishing [10, 11, 12]. Together, these findings call for reevaluation of current methods.

Promising defenses center on user interaction and assistance. Training systems can teach URL cue recognition and improve behavior in user studies [13], while interstitial warnings shown before a click guide users toward credible alternatives and reliably change click behavior [14]. Dataset driven evaluations also show that LLMs can assist during chat based social engineering by flagging risky exchanges and suggesting safer replies, providing help in the moment [15].

2.2 LLM assistants

LLM-based assistants have been rapidly adopted across domains since the advent of LLMs and Machine Learning as a Service (MLaaS), spanning healthcare and wellness [16], education [17], and cybersecurity [18]. This widespread use has prompted studies on their usability and impact on users. A recent large-scale usability evaluation analyzed over 11,000 app store reviews for popular generative AI apps (ChatGPT, Bing AI, etc.), finding significant differences in user satisfaction and effectiveness [19].

Studies in healthcare suggest that generative AI assistants can improve access to information and support decision-making. An experiment with older adults found that large language model tools such as Google Gemini (formerly Bard) and OpenAI GPT provided far more accurate and detailed health information than traditional voice assistants, with LLM-based systems producing errors only in a small fraction of responses compared to much higher error rates from conventional tools [20]. These systems also consolidated supplemental information that supported decision-making, potentially reducing cognitive burden; however, risks remain around residual errors, fluctuating behavior over time, and overly complex responses that require careful prompt and UI design [20].

In education, a 51-study meta-analysis reported overall positive effects of ChatGPT on learning performance, with additional gains in perceived learning experience and higher-order thinking when integration was guided (e.g., problem-based learning) [17]. Complementary usability

research in higher education found students rated generative tools favorably on usefulness and information quality; perceived usefulness was the strongest predictor of continued use, followed by trust and design appeal [21]. Together, these results highlight the importance of aligning assistant capabilities with instructional goals, transparency, and well-scaffolded workflows.

In cybersecurity, assistant-style applications increasingly support analysts by triaging alerts, explaining vulnerabilities, and proposing mitigations. This trend is enabled by broader AI advances: deep learning models in intrusion detection and malware classification often outperform signature-based systems, detecting attacks that legacy defenses miss [22, 23]; AI-assisted vulnerability management reduces noise while surfacing more true issues [24]; and state-of-the-art deepfake detectors exceed 90% accuracy on standard benchmarks [25]. These capabilities underpin assistants that accelerate investigation, improve coverage, and help prioritize risk.

Across different domains, the evidence suggests that LLM assistants can enhance task effectiveness and user experience when they are accurate, trustworthy, and well-integrated into existing practices. At the same time, limitations—hallucinations, variability, privacy and security concerns, and uneven performance across contexts—necessitate governance and careful evaluation.

3 Understanding Cognitive Security through a Real Example

To illustrate how cognitive-security attacks unfold in practice, we examine the 2024 compromise attempt against XZ Utils. The episode is emblematic not only because a technical backdoor was inserted, but because human and organizational dynamics made it possible—adversaries exploited limited attention, trust relationships, and social pressure to weaken governance.

We highlight three aspects: (1) background conditions that created asymmetries of attention and fatigue, (2) the adversary's long-term exploitation of those conditions through trust-building and consensus manipulation, and (3) the resulting cognitive kill chain that turned subtle social engineering into a systemic threat.

XZ Utils is a compression suite for Unix-like systems whose core library, `liblzma`, is linked by many userland components and distributions. Because it sits on the critical path of packaging and system tooling, changes to `liblzma` can indirectly affect high-value services on most Linux machines.

In 2024, a sophisticated backdoor [26] (CVE-2024-3094 [27]) was introduced into versions 5.6.0/5.6.1 via tarball-only build logic and opaque test artifacts, enabling malicious interference with `sshd` on certain builds; it never appeared in the public VCS history. The compromise was averted when a developer noticed unusual CPU use and Valgrind noise around SSH, traced the issue to `liblzma`, and triggered rapid vendor rollbacks before widespread impact.

Postmortem analyses underline a crucial precursor: *XZ Utils* relied for years on a single volunteer maintainer working in spare time, without institutional support [28]. This "bus factor of one" concentrated cognitive load, triage, and release decisions in one overburdened person—creating both sustained fatigue and a structural asymmetry in trust and scrutiny. Under such conditions, routine help becomes disproportionately valuable, and the path of least resistance is to delegate more to whoever reliably reduces workload.

Within those conditions, the maintainer was effectively compelled to accept help; the "Jia Tan" persona [29] stepped in offering to shoulder the unglamorous, time-consuming chores—bug fixes, packaging minutiae, and release preparation—steadily accruing credibility and day-to-day influence. According to timeline reconstructions [28] and reporting, this long-horizon trust

building was complemented by a coordinated pressure campaign on public lists, where new or low-reputation accounts criticized project velocity and argued for handing over more responsibility, then later pushed rapid downstream adoption. By late 2023–early 2024, the combination of earned trust and manufactured consensus enabled a side channel—tarball-only build logic and opaque "test" artifacts—to carry the backdoor outside usual VCS review surfaces, culminating in a classic supply-chain compromise attempt.

A further asymmetry likely compounded this dynamic: while the original maintainer contributed unpaid in spare time, multiple reports and timelines note that the patience, coordination, and sustained workload behind the operation are characteristic of resourced teams; it is therefore widely supposed that the organizers of the attack were remunerated—whether through institutional backing or dedicated funding—to maintain a multi-year, coordinated effort across identities, infrastructure changes, and release processes. This does not prove attribution, but it underscores the cognitive imbalance between a single volunteer and an adversary able to finance continuity, redundancy, and social engineering at scale.

The operation systematically targeted human limits on attention and judgment. Authority/halo effects were cultivated through years of "boring, helpful" work, so subsequent packaging oddities and tarball-only changes drew less scrutiny. Reciprocity and commitment nudged the maintainer to keep trusting the one person reliably reducing workload, while social proof—via new or low-reputation voices on mailing lists—manufactured a sense that "the community" wanted faster releases and a handover. Under chronic load, bounded attention and fatigue encouraged satisficing and delegation, and framing the updates as routine hygiene lowered perceived risk. Together these dynamics made elevated access and procedural shortcuts feel normal.

The *XZ Utils* incident exemplifies a cognitive kill chain[4]: (1) long-term trust cultivation; (2) consensus hijacking; (3) authority consolidation; (4) camouflage within low-salience process surfaces; (5) acceleration via anchoring and hygiene narratives. Its containment shows that anomaly-driven human sensemaking, paired with minimal governance hygiene, can still break such chains. For the broader agenda of cognitive security, the episode argues that protecting human decision processes must be a first-class design goal—one that blends socio-technical governance with GenAI agents capable of orchestrating detectors *and* reasoning about situations to produce faithful, accountable guidance.

4 Cognitive limitations relevant to security

This section summarizes cognitive limitations that attackers routinely exploit in security-relevant settings. It is *not* a comprehensive review; our goal is to make the problem surface visible, provide anchor references, and motivate design choices for the agent roles introduced later. We adopt the common view that many day-to-day judgments are driven by fast, heuristic processes rather than reflective analysis; this helps explain why simple manipulations (repetition, framing, social proof) systematically tilt perception and action in adversarial contexts. See Cybenko et al.[1] for the early articulation of *cognitive hacking* in security, and subsequent work connecting security behavior to heuristic, pattern-matching responses.

Authority and halo effects. Signals of status (titles, logos, verified badges) bleed into judgments about truth or risk [30]. Voice deepfakes of executives and spoofed "security team"

[4] A kill chain is a staged model describing how an attack progresses from preparation to impact—typically including steps like reconnaissance, tooling, delivery, exploitation, persistence/command-and-control, and actions on objectives. It is used to reason about where and how to detect, disrupt, or deter an adversary.

notices leverage authority to override normal verification.

Illusory truth via repetition. Repeated claims are judged truer, even when labeled uncertain [31, 32]. Coordinated reposts, duplicate images, and quote-tweets create familiarity that masquerades as credibility.

Social proof and bandwagon signals. Perceived consensus ("many people liked/shared this") increases acceptance [30]. Bot swarms and inauthentic coordination manufacture popularity cues that humans treat as validity.

Information overload, habituation, and decision fatigue. Too many alerts reduce sensitivity and increase error rates [33]. Repeated prompts drive habituation; operators satisfice, deferring to defaults or recent patterns.

Dual-process reasoning and bounded attention. Dual-process accounts distinguish fast, heuristic judgments from slower, reflective reasoning [3, 4]. Under time pressure, overload, or fatigue, humans default to the former, increasing susceptibility to persuasive but unreliable cues. In security, this manifests as click-through on urgent prompts, acceptance of plausible but false claims, and over-reliance on dashboard highlights. Empirical work links everyday security choices to person- and context-specific heuristics rather than deliberation.

Automation bias and over-trust. People over-rely on automated outputs, especially when interfaces look confident [34]. In security, poorly calibrated scores and persuasive explanations can induce passivity ("the system says it's fine").

Anchoring and default effects. Initial numbers or defaults pull subsequent estimates and choices [4, 35]. In security UIs, sticky thresholds and permissive defaults quietly set organizational risk posture.

Selective exposure and echo chambers. People preferentially consume like-minded sources, narrowing evidence diversity and reinforcing priors [36]. Coordinated operations exploit platform algorithms to maintain captive audiences.

Confirmation bias. People preferentially seek, notice, and recall information that fits prior beliefs [37]. Adversaries amplify congenial narratives and suppress disconfirming evidence (e.g., curated screenshots, selective statistics). In operations, this bias sustains incorrect working hypotheses and delays corrective action.

Sunk cost and escalation of commitment. Past investment biases continued commitment even as evidence degrades [38]. In incident response, teams keep pursuing a favored hypothesis while contradictory indicators accumulate.

Overconfidence and miscalibration. Humans are often more confident than correct, particularly in complex domains [39]. In security, unwarranted certainty delays escalation and reduces openness to contradictory signals.

Availability and recency. Vivid, recent, or widely shared items feel more likely [40]. Rumors resurface as "new" events; fresh spikes in noisy data look meaningful. Attackers exploit this with high-arousal content and timed bursts that crowd out base-rate context.

Framing and wording effects. Logically equivalent statements elicit different choices depending on wording [41]. Phishing and dark patterns use loss framing ("avoid penalty now"), default-accept flows, or pre-checked boxes to bias action.

Negativity/affect heuristic. High-arousal or negative content captures attention and is judged as more important [42]. Adversaries weaponize outrage and fear to accelerate unvetted sharing and rash approvals.

Change blindness and inattentional blindness. Low-salience but critical changes go unnoticed when attention is elsewhere [43]. Adversaries hide payloads in rarely reviewed surfaces (metadata, build glue, "test" artifacts).

Table 1 presents concrete examples of how each cognitive limitation can be exploited in security contexts.

5 GenAI agents as a solution

GenAI enables defenders to *reduce overload*, *make uncertainty explicit*, and *translate intent into safe, auditable outcomes* across tools and platforms. A single, all-purpose assistant is neither feasible nor desirable in practice: the work spans continuous vigilance, on-demand sensemaking, rapid but bounded execution, and cross-cutting governance—modes with different tempos, risks, and accountability needs. We therefore propose a framework of *specialized agent roles* that operationalize cognitive security by *composing* heterogeneous capabilities—detectors and verifiers (e.g., phishing, deepfake, provenance), retrieval and CTI pipelines, rule engines and policy checkers, cryptographic provenance services, domain APIs, and organization-specific workflows. Some agents run in the background to reduce noise, some engage on demand to structure decisions, some execute well-defined actions quickly, and some ensure that everything remains lawful, explainable, and accountable. (Other agentic stacks—lighter local models, symbolic planners—may fit specific settings; our focus is on GenAI-enabled composition because it most readily integrates what defenders already use.)

Our approach separates concerns along two axes—continuous versus on-demand, and assistive versus autonomous—to capture distinct performance and governance requirements. *Sentinels* operate continuously to handle thousands of simple checks with minimal friction, surfacing only concise signals when something merits attention. *Advisors* are invoked when judgment is required: they synthesize evidence, expose trade-offs, quantify uncertainty, and recommend next steps in a form people can immediately act on. *Executors* perform bounded actions under pre-approved playbooks, shrinking time-to-mitigate while preserving reversibility and traceability. *Keepers* monitor posture and configuration (not human interactions) and, upon detecting risky drift or vulnerabilities, brief the operator and may propose a ready-to-run Executor. *Stewards* wrap the whole system with policy, privacy, licensing, provenance, and audit, ensuring that data use is proportionate and every material step leaves a reviewable trail. Figure 1 presents a summary of the difference among the roles depending on their interaction mode (continuous [always running] vs. on-demand [started by a human]) and their autonomy (assistive [interacting with the human] vs autonomous [doing their actions on themselves]).

Cognitive limitation	**Typical exploitation in security-relevant contexts (examples)**
Authority / halo	CEO-voice deepfakes ordering urgent wire transfers; spoofed "security team" messages with logos/badges that suppress ordinary verification.
Illusory truth (repetition)	Coordinated reposts/quote-tweets of the same claim; recycling the same photo/video across different contexts to manufacture familiarity as "truth."
Social proof / bandwagon	Bot/troll swarms to fake consensus; artificially inflated like/share counts to make a dubious post appear widely endorsed.
Overload, habituation & fatigue	Alert storms that push analysts to click-through or ignore rare but critical signals; repeated MFA/consent prompts leading to automatic approval.
Dual-process reasoning & bounded attention	Urgency- and time-pressure scams that trigger fast, heuristic "click/approve now" responses; cluttered dashboards that nudge satisficing over verification.
Automation bias / overtrust	Operators deferring to a confidently styled "green" dashboard despite contradictory logs; overreliance on a single detector's score.
Anchoring & defaults	Over-permissive default thresholds in consoles that become sticky anchors; initial "suggested" numbers biasing incident impact estimates.
Selective exposure / echo chambers	Algorithmic curation that filters dissenting sources, letting coordinated narratives persist unchallenged in a community or workspace.
Confirmation bias	Curated screenshots/statistics that fit a group's prior beliefs; selective quoting of "expert" lines to sustain a favored hypothesis during incident response.
Sunk cost & escalation of commitment	Continuing to pursue an initial incident hypothesis after contradictory indicators accumulate, because significant effort has already been invested.
Overconfidence / miscalibration	Declaring an investigation "clean" with high confidence on weak evidence; bypassing two-channel verification due to perceived expertise or past success.
Availability & recency	Resurfacing an old event as "breaking" to drive sharing; timed rumor bursts right before a decision window so vivid, recent items feel more likely than base rates.
Framing / wording effects	Loss-framed phishing ("avoid penalty now") and dark-pattern flows; "urgent security update" pages that pre-check risky options.
Negativity / affect heuristic	Outrage-bait headlines that accelerate unvetted resharing; fear-framed prompts ("account will be deleted") that trigger rash approvals.
Change/ inattentional blindness	Critical but low-salience config changes (e.g., CI/CD secrets, sharing rules) go unnoticed while attention is on a flashy incident.

Table 1: Examples of exploitations of the different cognition limits.

This composition yields a division of labor that is hard to achieve in a monolith. Speed and coverage dominate in the Sentinel tier; clarity and usefulness dominate in the Advisor tier; decisive containment dominates in the Executor tier; and compliance and accountability dominate in the Steward tier. By allowing each archetype to optimize for its primary objective—and by linking them through simple, auditable contracts—we obtain a system that is fast where it must be fast and careful where it must be careful. Sentinels generate compact signals that Advisors can explain; Advisor recommendations can authorize or parameterize Executor actions; Keepers surface system issues and propose safe fixes; Stewards evaluate and log these flows so decisions remain defensible to operators, auditors, and external stakeholders.

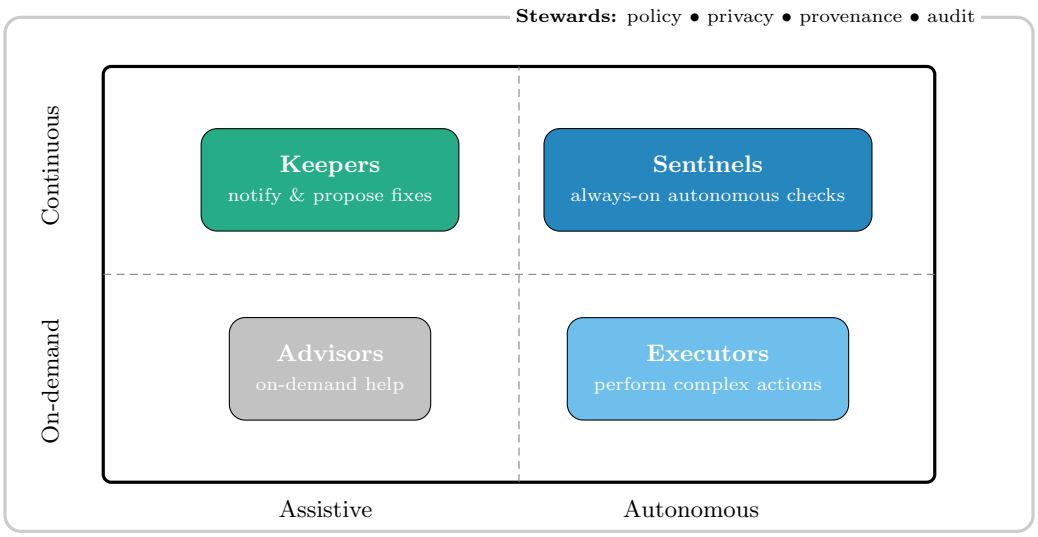

Figure 1: Agent roles by interaction mode (continuous vs. on-demand) and autonomy (assistive vs. autonomous). Stewards provide system-wide governance around all roles.

The architecture is intentionally human-centered. Users see subtle cues rather than interruptions, can request structured explanations when needed, and retain control over actions with clear approval gates for higher-impact steps. Organizations express risk appetite and legal constraints as policies that Stewards enforce uniformly, while still benefiting from the reach and responsiveness of automation. Because roles are explicit, evaluation becomes straightforward: Sentinels by latency, coverage, and calibration; Advisors by time-to-insight and explanation quality; Executors by action precision and rollback safety; Keepers by time-to-notice and fix clarity; Stewards by policy adherence and audit completeness.

Finally, the model is adaptable across domains. In consumer contexts it can quietly annotate feeds and messages, turning chaotic streams into comprehensible, safer experiences. In enterprise settings it plugs into case management, CTI pipelines, and collaboration tools to accelerate investigation without sacrificing due process. In public-sector and platform settings it supports consistent labeling, moderation, evidence handling, and reporting across teams and jurisdictions. The subsections that follow detail each archetype—Sentinels, Advisors, Executors, Keepers, and Stewards—and show how their coordinated operation delivers trustworthy cognitive security in practice.

5.1 Sentinels

Sentinels are always-running agents that take care of countless small checks in the background without asking for attention. They prioritize speed and coverage over perfect accuracy, using lightweight heuristics and small local models when possible, and only escalating when something looks truly unusual. Think of them as the quiet metabolism of the system: continuously looking at new links, profiles, images, indicators, and files, and turning them into quick "safe/iffy/needs a look" signals.

They are completely unattended. A Sentinel can invoke simple external tools—expand a URL, read basic metadata, query a reputation list, run a fast authenticity probe—and stitch

the results into a compact verdict with a short explanation. They are designed to live close to where the data appears (on-device or at the edge when feasible) for privacy and latency, and to hand off only the few items that merit human attention or a higher-level agent.

In cognitive security settings, Sentinels can triage news claims and images as they appear in a feed, attach a subtle "reposted from 2019" or "likely AI-generated" note, and spot coordinated reply swarms or brand-new accounts behaving like bots. They can unshorten links, detect bait-and-switch redirects, surface original sources, and flag manipulative UI patterns in suspicious pages. The user doesn't ask for any of this; the cues simply materialize in context, keeping the experience smooth while nudging safer judgment.

In cybersecurity and CTI pipelines, Sentinels quietly normalize indicators, enrich domains and IPs with basic lookups, deduplicate near-identical artifacts, and assign quick confidence tags to reports. For crypto-related cases, they can auto-tag known services or mixers, identify obvious scam patterns, and prepare a minimal evidence note that an Advisor can expand on demand. They also catch low-hanging hygiene issues—expired certificates on internal tools, typo-squatted domains mentioned in chats, or attachments that match known-bad hashes—so specialists focus on harder problems.

Implementations can be tailored to the domain: a privacy-first layer on personal devices that annotates the web and messaging; an organizational layer in SOCs and LEAs that feeds clean signals into case management; and platform-level Sentinels that run at ingestion for social networks, newsrooms, marketplaces, or civic platforms. Whatever the setting, their job is the same: keep the simple things simple, surface only what matters, and supply Advisors and Executors with just enough signal to be effective without burdening the user.

5.2 Advisors

Advisors are on-demand, human-facing agents that you call when you need judgment, not just signals. They synthesize evidence gathered by Sentinels and other sources, weigh trade-offs, and present clear options with pros, cons, and estimated risk. Instead of running constantly, they wake up for a question or a task, focus attention on the relevant context, and structure the path to a decision. They are built to be transparent: every recommendation comes with an explanation, cited evidence, and an explicit statement of uncertainty.

While Sentinels optimize for speed, Advisors optimize for clarity and usefulness. They can invoke tools—search, retrieval, graph queries, quick simulations, or domain APIs—and, when needed, spin up larger models for deeper reasoning, keeping privacy constraints and policies in view. The interaction is conversational but outcome-oriented: they propose next actions, draft artifacts, and adapt as the human asks "what if?" or provides new constraints. The human stays in control; the Advisor keeps the reasoning organized and the path forward concrete.

In cognitive security scenarios, an Advisor can take a messy rumor and turn it into a structured brief: what's claimed, what's corroborated, what conflicts, what likely happened, and how confident we are. It can compare narratives across outlets, highlight recycled imagery, and suggest counter-messaging that is factual and non-amplifying. For teams facing information storms, the Advisor summarizes the landscape, flags the few items worth responding to, and drafts a response or guidance note that a human can approve.

In cybersecurity and CTI work, an Advisor acts as an investigator's copartner. Given an alert or an indicator list, it correlates related events, enriches with threat intel, groups artifacts by modus operandi, and proposes a playbook: which hosts to check, which queries to run, what to contain first. For crypto-related reports, it classifies the case, links addresses to known services where possible, outlines plausible flows, and prepares a clean evidence bundle for review.

Across all of these, it keeps track of confidence, assumptions, and open questions so escalation is deliberate rather than reactive.

Implementations vary by domain and risk tolerance. A personal Advisor sits in the browser or messenger as a "ask me when needed" panel that turns confusing posts into understandable briefs. An enterprise Advisor is embedded in SOC consoles or case-management tools, tying together logs, CTI, and tickets to accelerate triage and reporting. A public-sector Advisor supports LEAs and policymakers with standardized dossiers and options papers that respect jurisdictional and privacy constraints. In every setting, the core promise is the same: when you ask for help, the Advisor delivers a reasoned, explainable recommendation you can act on.

5.3 Executors

Executors are autonomous agents that take bounded actions when certain conditions are met. Unlike Sentinels, which watch, and Advisors, which explain, Executors do. They wake on a trigger—a thresholded risk score, an explicit human request, or a time-critical event—and apply pre-approved playbooks. The emphasis is on speed and containment rather than analysis: act quickly, within strict guardrails, and leave a clear trace of what was done and why. Actions are reversible where possible, and higher-impact steps require lightweight approval gates or multi-party consent.

They operate under policy from the first step. An Executor only touches resources it is authorized to touch, chooses the least-invasive remedy that achieves the objective, and records evidence and reasoning so the action can be audited or rolled back. In practice this looks like short, well-scoped runs: isolate, rate-limit, revoke, file, notify. The goal is to shrink time-to-mitigate without turning autonomy into overreach.

In cognitive security, Executors can rate-limit an emerging bot swarm, temporarily lock replies on a post while provenance is checked, or hide obviously recycled or manipulated media pending review. They can prefill and submit a takedown request to a platform with linked evidence, or quietly add suspicious accounts to a watchlist while notifying a moderator. For individuals, they can auto-mute low-reputation replies or warn and detour a click away from a bait-and-switch link, restoring normal behavior once conditions clear.

In cybersecurity and CTI workflows, Executors handle the mechanical but urgent steps that follow a confident signal. They can open or update tickets with a completed checklist, quarantine a malicious attachment in shared storage, revoke a leaked token, block an egress domain on a temporary rule, or push a cleaned STIX/MISP event to collaborators. For crypto-related cases, they can snapshot relevant pages, archive transactions, notify an exchange's abuse desk with a structured report, and set timed reminders for follow-up—all without waiting on a human to push buttons.

Implementations vary with context and risk tolerance. A personal setup keeps Executors close to the user—inside the browser or mail client—to intercept risky clicks, quarantine downloads, or manage mute/allow lists. An enterprise SOC binds Executors to infrastructure and identity systems so they can enforce containment quickly but within change-control policy. Platforms and public agencies run Executors at ingestion or triage layers to de-amplify harmful patterns at scale while preserving due process. In every domain the principle is the same: act fast, act within bounds, and leave the system safer than it was a moment before.

5.4 Keepers

Keepers are continuous, assistive agents dedicated not to moderating human interactions but to watching the health of the system itself and informing humans when something is wrong. They

operate quietly in the background, scanning configurations, dependencies, credentials, quotas, certificates, sharing rules, and policy settings for signs of vulnerability or drift. When they detect a problem, they do not act unilaterally; instead, they raise a focused, human-readable brief that explains what is broken, why it matters, and what the safest next steps could be.

The interaction model is simple and respectful of control. A Keeper wakes on a clear signal—an expiring TLS certificate, a vulnerable library version, an over-permissive token, a data-sharing rule that violates policy—and presents a short recommendation with impact, confidence, and one or two reversible remedies. Crucially, Keepers may also propose the execution of an *Executor*: the brief includes a ready-to-run action that the user can approve with a single confirmation. The human can accept the proposed Executor, schedule it, or ask an Advisor for a deeper explanation. Once approved, the Keeper dispatches the precisely scoped instruction to the appropriate Executor and then verifies the outcome, leaving a traceable record for Stewards to audit.

Keepers shine wherever posture and configuration change frequently. In cybersecurity and CTI environments they surface least-privilege violations, stale API keys, or ingestion rules that accidentally capture PII, and they propose compliant corrections—often as an Executor that rotates a key, tightens a role, or updates a filter. In cognitive-security platforms they flag labeling thresholds or jurisdictional settings that no longer match policy and suggest adjustments that keep the system within acceptable bounds, again offering an Executor to apply the change. In crypto-related workflows they notice when heuristics drift toward false positives and offer calibrated threshold updates with an estimated impact, paired with an Executor to implement and monitor the adjustment.

The value is in timing and clarity rather than automation for its own sake. By notifying operators the moment a concrete, remediable issue is detected—and by proposing minimal, reversible actions that can be executed via an Executor under policy—Keepers reduce mean time to insight without eroding human oversight. They complement Sentinels, which monitor content and actors; Advisors, which provide deeper analysis on request; Executors, which apply approved changes; and Stewards, which ensure every notification, decision, and fix remains within policy and is fully explainable.

5.5 Stewards

Stewards are the guardians of the system's rules and memory. Where Sentinels watch, Advisors guide, Executors act, and Keepers inform operators of system issues, Stewards ensure that everything happens within policy, with privacy respected, provenance preserved, and accountability guaranteed. They surround the ecosystem, interpreting organizational and legal constraints, translating them into enforceable checks, and maintaining a durable, explainable record of what data was used, what decisions were made, and on what basis.

They operate quietly but decisively. A Steward evaluates whether a requested action or data flow is permitted, applies redaction or minimization when needed, verifies content licenses and terms of use, and attaches provenance so outputs are traceable. Consent, jurisdiction, retention, age-appropriateness, and sharing boundaries are treated as first-class concerns; if a rule is ambiguous, Stewards default to caution and request explicit human approval. They also tend the long-lived parts of the system—trusted source reputations, model and tool version histories, and immutable audit trails—so future reviews can reconstruct events without guesswork.

In cognitive security contexts, Stewards prevent well-intended defense from becoming overreach. If an Advisor drafts counter-messaging, the Steward checks that it does not amplify harmful narratives or disclose sensitive data. If Sentinels flag likely AI-generated media, the

Steward ensures the label and any subsequent moderation honor platform policies and local law. When a user requests evidence behind a claim score, the Steward produces a transparent, privacy-preserving explanation with citations and timestamps rather than opaque model outputs.

In cybersecurity and CTI workflows, Stewards keep handling of indicators, logs, and reports clean and lawful. They ensure that STIX/MISP exports respect sharing groups, scrub PII where unnecessary, and attach chain-of-custody details to forensic artifacts. When Keepers alert on a misconfiguration or vulnerability and an Executor is proposed to apply a fix, Stewards enforce change-control rules, record who authorized what, and set retention timers so temporary measures do not become permanent by accident. For crypto-related cases, Stewards govern what can be shared with exchanges or LEAs, enforce notification templates, and document precisely which sources support each allegation.

Implementations vary by setting. On personal devices, Stewards are the privacy layer that keeps data local by default and explains why a request needs broader access. In organizations, they live as policy engines and audit services embedded in SOC and case-management tools, aligning actions with internal guidelines and regional regulations. On platforms and in the public sector, Stewards provide consistent labeling, moderation, evidence handling, and reporting across teams and jurisdictions. In every domain, their role is to make the whole agent ecosystem trustworthy: decisions are explainable, data use is proportionate, and every important step leaves a clear, reviewable trail.

6 Conclusions

Cognitive security is the protection of human perception, judgment, and decision-making from manipulation, overload, and engineered uncertainty across digital channels and organizational contexts. It complements (not replaces) technical controls by focusing on the human outcomes those controls are meant to safeguard—surfacing signals at the right moment, exposing uncertainty, and keeping actions explainable and reversible.

Delivering this protection is a complex, socio-technical task that demands multidisciplinary research—spanning security engineering, HCI, psychology, cognitive science, governance, and law. Within that broad effort, we argue that generative-AI (GenAI) agents are a strong enabling substrate: they can watch wide information streams, synthesize evidence, support human sensemaking, and perform tightly bounded, auditable actions—provided their autonomy is calibrated and their provenance, privacy, and policy constraints are enforced.

We therefore propose a role-based taxonomy of agents. **Sentinels** are always-on detectors that surface low-friction, calibrated alerts. **Advisors** are on-demand sensemakers that assemble, cite, and explain evidence. **Executors** carry out narrowly scoped, reversible actions under explicit guardrails. **Keepers** maintain posture by monitoring configuration, drift, and hygiene over time. **Stewards** govern the system—enforcing policies, privacy, provenance, auditability, and human-override. This division of labor aligns speed with care and maps directly to human cognitive limits and institutional constraints.

Much remains to be done. We need shared benchmarks tied to human outcomes, robust defenses against prompt injection and tool abuse, reproducible pipelines and two-person integrity for high-impact steps, and deployment patterns that respect legal and ethical boundaries. Multiple solutions—technical, organizational, and educational—should be developed in parallel to protect societies, organizations, and citizens. Our contribution is a practical framing and an agentic blueprint to guide that next wave of research and implementation.

7 Acknowledgements

This work has been funded in part by the Horizon Europe research and innovation programme of the European Union, under grant agreement no 101136024, project EMPYREAN.

References

[1] G. Cybenko, A. Giani, and P. Thompson. Cognitive hacking: a battle for the mind. *Computer*, 35(8):50–56, 2002.

[2] Alan R. Dennis and Randall K. Minas. Security on autopilot: Why current security theories hijack our thinking and lead us astray. *SIGMIS Database*, 49(SI):15–38, April 2018.

[3] Daniel Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York, 2011.

[4] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[5] N. Ramkumar, Vijay H. Kothari, Caitlin Mills, R. Koppel, J. Blythe, Sean W. Smith, and A. Kun. Eyes on urls: Relating visual behavior to safety decisions, 2020.

[6] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian S. Fischer. The emperor's new security indicators, 2007.

[7] Rachna Dhamija, J. D. Tygar, and Marti A. Hearst. Why phishing works, 2006.

[8] Dawn M. Sarno and J. Black. Who gets caught in the web of lies?: Understanding susceptibility to phishing emails, fake news headlines, and scam text messages, 2023.

[9] Theodore Longtchi and Shouhuai Xu. Characterizing the evolution of psychological tactics and techniques exploited by malicious emails, 2024.

[10] Fred Heiding, Simon Lermen, Andrew Kao, Bruce Schneier, and Arun Vishwanath. Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. *arXiv preprint arXiv:2412.00586*, 2024.

[11] Qinglin Qi, Yun Luo, Yijia Xu, Wenbo Guo, and Yong Fang. Spearbot: Leveraging large language models in a generative-critique framework for spear-phishing email generation, 2024.

[12] S. Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. From chatbots to phishbots? - preventing phishing scams created using chatgpt, google bard and claude, 2023.

[13] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Trans. Internet Technol.*, 10(2), June 2010.

[14] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer. Adapting security warnings to counter online disinformation. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1163–1180. USENIX Association, August 2021.

[15] Lin Ai, Tharindu Sandaruwan Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael S. Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu, and Julia Hirschberg. Defending against social engineering attacks in the age of LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12880–12902, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[16] T. R. Campellone, M. Flom, R. M. Montgomery, L. Bullard, M. C. Pirner, A. Pavez, M. Morales, D. Harper, C. Oddy, T. O'Connor, J. Daniels, S. Eaneff, V. L. Forman-Hoffman, C. Sackett, and A. Darcy. Safety and user experience of a generative artificial intelligence digital mental health intervention: Exploratory randomized controlled trial. *Journal of Medical Internet Research*, 27:e67365, 2025.

[17] J. Wang and W. Fan. The effect of chatgpt on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12:621, 2025.

[18] Xihuan Lin, Jie Zhang, Gelei Deng, Tianzhe Liu, Xiaolong Liu, Changcai Yang, Tianwei Zhang, Qing Guo, and Riqing Chen. Ircopilot: Automated incident response with large language models, 2025.

[19] R. Alabduljabbar. User-centric ai: evaluating the usability of generative ai applications through user reviews on app stores. *PeerJ Computer Science*, 10:e2421, 2024.

[20] {Emily M.} Langston, Varitnan Hattakitjamroen, Mario Hernandez, {Hye Soo} Lee, Hannah Mason, Willencia Louis-Charles, Neil Charness, {Sara J.} Czaja, {Wendy A.} Rogers, Joseph Sharit, and {Walter R.} Boot. Exploring artificial intelligence-powered virtual assistants to understand their potential to support older adults' search needs. *Human Factors in Healthcare*, 7, June 2025. This research was supported in part by the National Institutes of Health (National Institute on Aging) grant [P01AG073090] under the auspices of the Center for Research and Education on Aging and Technology Enhancement (CREATE ; www.create-center.org).

[21] Goran Bubaš, Antonela Čižmešija, and Andreja Kovačić. Development of an assessment scale for measurement of usability and user experience characteristics of bing chat conversational ai. *Future Internet*, 16(1), 2024.

[22] Vincent Zibi Mohale and Ibidun Christiana Obagbuwa. Evaluating machine learning-based intrusion detection systems with explainable ai: enhancing transparency and interpretability. *Frontiers in Computer Science*, Volume 7 - 2025, 2025.

[23] M. Farhan, H. Waheed ud din, S. Ullah, et al. Network-based intrusion detection using deep learning technique. *Scientific Reports*, 15:25550, 2025.

[24] Shengye Wan, Joshua Saxe, Craig Gomes, Sahana Chennabasappa, Avilash Rath, Kun Sun, and Xinda Wang. Bridging the gap: A study of ai-based vulnerability management between industry and academia, 2024.

[25] Gourav Gupta, Kiran Raja, Manish Gupta, Tony Jan, Scott Thompson Whiteside, and Mukesh Prasad. A comprehensive review of deepfake detection using advanced machine learning and fusion methods. *Electronics*, 13(1), 2024.

[26] Wikipedia contributors. Xz utils backdoor. https://en.wikipedia.org/wiki/XZ_Utils_backdoor, 2024. Accessed 2025-09-30.

[27] Cve-2024-3094 detail. https://nvd.nist.gov/vuln/detail/CVE-2024-3094, 2024. Accessed 2025-09-30.

[28] Russ Cox. The xz backdoor timeline. https://research.swtch.com/xz-timeline, 2024. Accessed 2025-09-30.

[29] Matt Burgess and Andy Greenberg. Who is jia tan? inside the hunt for the xz backdoor's mystery author. https://www.wired.com/story/jia-tan-xz-backdoor/, 2024. Accessed 2025-09-30.

[30] Robert B. Cialdini. *Influence: The Psychology of Persuasion*. Harper Business, New York, revised edition, 2007.

[31] Lynn Hasher, David Goldstein, and Thomas Toppino. Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1):107–112, 1977.

[32] Lisa K. Fazio, Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5):993–1002, 2015.

[33] Maria Cvach. Monitor alarm fatigue: An integrative review. *Biomedical Instrumentation & Technology*, 46(4):268–277, 2012.

[34] Kathleen L. Mosier, Linda J. Skitka, Susan Heers, and Michael Burdick. Automation bias: Decision-making and performance in high-tech cockpits. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 40, pages 204–208. SAGE, 1996.

[35] Eric J. Johnson and Daniel Goldstein. Do defaults save lives? *Science*, 302(5649):1338–1339, 2003.

[36] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[37] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

[38] Barry M. Staw. Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational Behavior and Human Performance*, 16(1):27–44, 1976.

[39] Richard P. Larrick. Debiasing. In Derek J. Koehler and Nigel Harvey, editors, *The Blackwell Handbook of Judgment and Decision Making*, pages 316–338. Blackwell, Oxford, 2004.

[40] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973.

[41] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[42] Paul Slovic, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor. The affect heuristic. *European Journal of Operational Research*, 177(3):1333–1352, 2007.

[43] Ronald A. Rensink. Seeing, sensing, and scrutinizing. *Vision Research*, 40(10–12):1469–1487, 2000.