Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2770–2779
https://doi.org/10.54941/ahfe1008501

Detection Without Speed Isn't Defense: Human-AI Coordination Against Cognitive Cyberattackers

Ayla Santos[1], Stephan Carney[2], Rob Groome[3], and Nikolos Gurney[3]

[1]California State Polytechnic University, Pomona, CA, USA
[2]Marshall School of Business, University of Southern California, Los Angeles, CA, USA
[3]Institute for Creative Technologies, University of Southern California, Playa Vista, CA, USA

ABSTRACT

Accurate attacker identification improves defense only if defenders can act in time. Because an attacker's cognitive profile can dictate their network maneuvering, we evaluated whether an AI defender that profiles attacker cognition outperforms a human-inspired defender, and if pairing them improves defense. Using an agent-based simulation of an enterprise network, we tested two cognitively distinct attackers (aggressive vs. cautious) against three defender configurations (human-analog heuristic, AI, and human-machine team) under partial observability in a 2×3 factorial design. Counterintuitively, the human-machine team underperformed both individual defenders. The AI defender performed best, yet its identification accuracy remained at chance, indicating its success stemmed from strong default actions rather than accurate profiling. Improving the AI's observability raised identification accuracy across three interventions but failed to improve defense. Only a structural intervention trapping attackers without identification reduced breaches, leveraging a compact topology with centralized hubs. These findings show that rapidly stopping a threat is distinct from identifying it. In many instances, effective human-AI cyber defense may rely less on profiling accuracy than on response speed and seamless coordination.

Keywords: Cyberpsychology, Human-AI teaming, Cognitive profiling, Cyber defense, Agent-based simulation, Decision-making

INTRODUCTION

Cyber defenders operate under a fundamental asymmetry: they must secure every path into a network, while an attacker needs only one. It is easy to overlook that the adversary is human, and human attackers often emit identifiable behavioral traces. For example, an operator who disregards detection risk moves very differently from one who advances cautiously. Prior work shows these differences are driven more by an attacker's cognitive profile than by experience or expertise (Kim et al., 2025). We argue that a defender who recognizes the attacker type they face is able to more effectively use their defense resources.

Judgment and decision-making (JDM) research establishes much of the vocabulary needed to understand the differences behind identifiable attacker types. Decision-makers are boundedly rational (Simon, 1955), meaning

Received September 25, 2026; Revised October 31, 2026; Accepted November 8, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

that they rely on heuristics that produce systematic departures from the predictions of rationality axioms (Tversky and Kahneman, 1974; Kahneman and Tversky, 1979). A small number of measurable dimensions are able to recover individuals' susceptibility to these departures (Stanovich and West, 2000; Appelt et al., 2011). Of particular interest to the present research are risk propensity, the general tendency to accept or seek risk (Zhang et al., 2019); cognitive reflection, the tendency to override an intuitive response in favor of deliberation (Frederick, 2005; Toplak et al., 2011); and ambiguity tolerance, the willingness to act when outcome probabilities are unknown rather than merely uncertain (Ellsberg, 1961; Fox and Tversky, 1995). Each of these cognitive dimensions maps onto a frequent attacker choice: trading detection risk for faster access, verifying before acting, or probing an unfamiliar service rather than a documented one. Prior art establishes that each mapping correlates with observed red-team behavior (Gutzwiller et al., 2019; Kim et al., 2025; Carney et al., 2025).

We use an agent-based simulation to ask whether a defender that profiles attacker cognition can convert that knowledge into better defense, and whether a human-inspired defender, an AI defender, or a paired human-machine team does so most effectively. The simulated agents are effectively points in the trait space created by the relevant JDM dimensions. Treating attackers as "types" in this space enables us to ground the simulation in validated constructs and formalize scoring rules. Alternative approaches, such as using ad hoc personas, would not allow us to formalize, for example, how a profile weighs value against visibility. We hypothesize that the paired defender outperforms either individual defender when the AI is uncertain (H1), and that the risk-seeking attacker is hardest to defend against regardless of defender type (H2).

BACKGROUND AND RELATED WORK

A growing body of work from the IARPA ReSCIND program establishes that attacker behavior is shaped by measurable cognitive traits. Beltz et al. (2026) introduced cognitive profiling of red-team operators, characterizing the decision-making tendencies each attacker brings to an operation. Building on that data, Kim et al. (2025) showed that these profiles predict which MITRE ATT&CK techniques (Strom, et al., 2020) an operator uses. Critically, they show that behavior can align more closely to JDM processes than to expertise or experimental conditions. Their use of the Cognitive Reflection Test (Frederick, 2005) and the General Risk Propensity Scale (Zhang, et al., 2019) motivates the attacker parameters modeled here. Carney et al. (2025) extended the profile space with an ambiguity-averse attacker that prefers known over unknown probabilities. On the defender side, Hans et al. (2025) demonstrated that Bayesian and LLM-based inference can learn attacker strategy and cognitive signals from network telemetry, motivating our Bayesian AI defender. Whereas Gurney et al. (2025) asked whether human predispositions impact their effectiveness in human-machine teams, motivating our human, AI, and paired defender conditions. We test whether cognitive profiling, once achieved, actually improves defensive outcomes,

and whether pairing a human and an AI defender achieves the benefit that teaming models predict in a simulated network attack.

METHOD: SIMULATION DESIGN

We created a fixed network topology for all trials to ensure results are comparable across runs. The network consisted of 30 nodes organized into three zones: the Demilitarized Zone (DMZ), the Internal Subnet (INT), and High-Value Targets (HVT). The DMZ represents the attacker's entry point and contains the highest-visibility nodes. The INT is the traversal space connecting the DMZ to the HVTs and contains the network's hub nodes, which have the highest connectivity used for lateral movement. The HVT represents the crown jewels of the network. These nodes have the highest value but lowest visibility and familiarity, making them both attractive and difficult to detect or reach. Nodes in this zone have a 50% compromise probability, while other zones are 100%. Each node was defined with four properties: value, familiarity, visibility, and connectivity. Value represents how attractive a node is to an attacker. Familiarity reflects how well-known or commonly used a service is; it was designed to drive an Ambiguity-Averse attacker's behavior specifically and is not used by the attacker types tested in this study. Visibility is the probability that an alert is triggered when a node is interacted with, representing the likelihood of detection. Connectivity refers to the number of nodes directly adjacent to a given node, representing its worth as a pivot point for lateral movement.

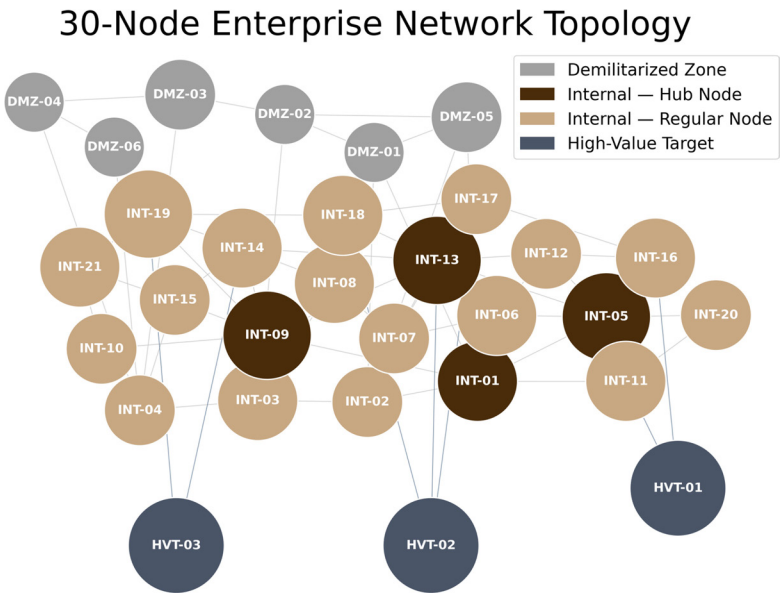


Figure 1: The fixed 30-node enterprise network, organized into DMZ, Internal Subnet, and High-Value Target zones.

The two attackers are given the values of each node property and are distinguished by how they score neighboring nodes each turn, moving to

the highest-scoring uncompromised neighbor. In addition to the base scores, we added a small random noise term (epsilon) to attacker decisions so their movement is less than perfectly predictable. The Rational attacker represents a baseline that maximizes expected value without cognitive bias; it is included for comparison and is not an experimental condition. Its scoring formula is score = value + connectivity – visibility, and it is expected to move deliberately, choosing broadly beneficial nodes. The high-GRiPS attacker—modeled as an operator scoring high on the General Risk Propensity Scale, a domain-general measure of risk-taking tendency (Zhang et al., 2019; applied to attacker behavior by Kim et al., 2025)—sits at the other end of the scale: it moves aggressively and focuses on high-value nodes, disregarding detection risk and readily triggering alarms. Its scoring formula is score = (value × $\text{risk}_{\text{multiplier}}$) – (visibility × $\text{risk}_{\text{discount}}$), where $\text{risk}_{\text{multiplier}} = 1.5$ and $\text{risk}_{\text{discount}} = 0.4$, so value dominates and visibility is largely disregarded.

We simulated a Human, an AI, and a Human-Machine defender. The heuristic defender applies fixed priority rules in order, with no memory or inference, representing a human operator relying on experience and intuition. During each turn, it has an action budget of three actions chosen from monitoring a node, hardening a node, and blocking an edge. Monitoring temporarily boosts detection for any attacker interaction on that node for that turn. When an attacker lands on a hardened node, the compromise probability falls to 50%; which expires after four turns. Blocking removes the highest-connectivity edge adjacent to a known compromise and is restored after one turn. The heuristic defender applies these rules sequentially, moving to the next if the current one does not apply: harden a node adjacent to known compromise; harden HVTs within two hops of compromise if an alert fired; monitor the highest-value uncompromised nodes; and block the highest-connectivity edge near compromise.

Our analytical defender maintains a probability distribution over attacker types, updated after each turn via Bayesian inference, and represents the AI cognitive-profiling component. It knows the attacker scoring equations and the exact value, visibility, and connectivity of every node, but the attacker type is unknown. Before each trial, every attacker type is equally likely, so the defender has no initial confidence in what it faces. During a trial it observes which nodes triggered alerts, compares the alert pattern to the expected behavior of each attacker type, updates the type probabilities, and allocates actions by the most probable type. Once confidence in a single type exceeds 45%, it performs predefined actions; for a dominant high-GRiPS attacker it sorts uncompromised nodes by value and hardens them. Because we introduced decision noise (epsilon) and partial observability, confidence does not rise easily; below the threshold, the default action is to monitor the highest-value uncompromised nodes.

Our Paired defender combines the heuristic and analytical agents to represent a Human-Machine team. Both agents share the same underlying state for hardening, blocking, and monitoring, so an effect applied by one persists (e.g., a harden node still expires after four turns) when control switches. Belief updating happens once per turn regardless of which agent is in control, so the analytical half keeps learning even while the heuristic directs.

Once the analytical agent's confidence crosses the 45% threshold it takes control; below the threshold, the paired defender falls back to the heuristic's fixed rules. This realizes the human-machine dynamic: the human heuristic acts as a reliable floor until the machine has gathered enough evidence, at which point the analytical component takes over.

Each turn proceeds in a fixed sequence: the attacker moves to its highest-scoring uncompromised neighbor and attempts compromise; an alert then fires probabilistically from that node's visibility plus any monitoring bonus; the defender observes only whether an alert fired, updates its beliefs (if analytical or paired), and spends its action budget according to the attacker type it believes it faces. A trial ends when the attacker compromises a HVT (attacker win) or after 50 turns (defender win).

We ran a fully balanced factorial design, testing both attacker types (Rational and high-GRiPS) against all three defender types (Heuristic, Analytical, Paired) for six conditions. Each condition ran 500 trials, for 3,000 trials total, with every cell equally represented. Each trial used its own seed, built from a master seed of 42 plus the trial's ID, which drove every random draw so any trial can be reproduced exactly from its logged seed. The graph was fixed and hard-coded across runs, and the attacker always started at the same DMZ entry node (DMZ-01). Every turn's state was logged; once a trial ended, we logged the outcome, turns elapsed, nodes compromised, total alerts, and the defender's mean inference confidence.

HYPOTHESES

Our first hypothesis (H1) states that the paired human-machine defender will outperform both individual defenders most significantly under moments of low confidence. We predict that when the analytical agent is uncertain about the attacker type, the heuristic agent applies its fixed rules to provide a reliable floor that prevents misallocation of defensive resources. Our second hypothesis (H2) states that the high-GRiPS attacker will be the hardest to defend against regardless of defender type. Given its disregard for risk, it willingly triggers alerts, making its behavior look like noise rather than signal, so any defender struggles to distinguish true attacker movement; it speeds toward high-value targets and accepts detection costs that other attacker types avoid.

RESULTS

Overall Attacker Win Rate

To determine the overall attacker win rate, we pooled results across all defenders. The high-GRiPS attacker won more trials than the rational attacker, 87.9% vs. 79.9% ($\chi^2(1) = 35.5$, $p < .001$), confirming that the risk-disregarding attacker is harder to stop. Pooling across attackers produced a result contrary to our expectation: the heuristic defender allowed the attacker to win 87.5% of the time, the analytical defender 75.4%, and the paired defender 88.8%. The analytical defender alone was the best-performing defender in the experiment, and the paired defender was the worst.

To confirm this statistically, we ran a logistic regression predicting attacker success from defender type, attacker type, and their interaction, using the heuristic defender against the rational attacker as the reference. The analytical defender significantly reduced attacker success relative to the heuristic (OR = 0.58, $p < .001$), while the paired defender showed no significant advantage (OR = 1.32, $p = .104$). The high-GRiPS attacker was significantly more likely to succeed overall (OR = 3.05, $p < .001$), and the analytical defender's advantage decreased against it (OR = 0.46, $p = .003$), while the paired defender showed no significant change ($p = .126$). The gap between best and worst defender was 13.4 percentage points against the rational attacker ($\chi^2 = 29.7$, $p < .001$) and 14.6 points against high-GRiPS ($\chi^2 = 61.9$, $p < .001$). Hence, defender type mattered for both attackers, but not in a way that benefited the paired defender.

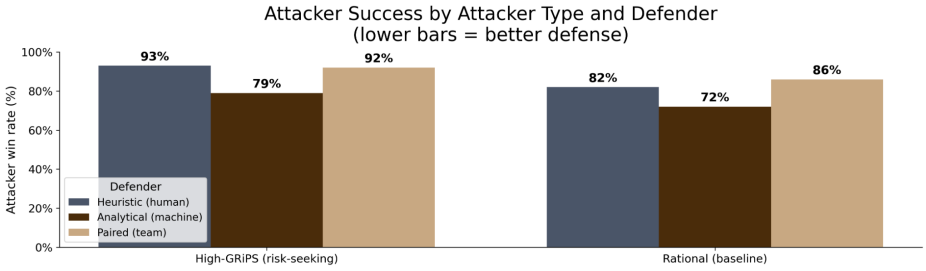


Figure 2: Attacker win rate by attacker type and defender (lower bars indicate better defense).

Table 1: Attacker win rate by attacker type and defender.

Attacker	Heuristic	Analytical	Paired
High-GRiPS	93.2%	78.6%	92.0%
Rational	81.8%	72.2%	85.6%

Inference Quality

Because the analytical defender outperformed the others, we examined how it was actually functioning. The probability assigned to the attacker's true type over a trial sat essentially at chance: 0.236 for high-GRiPS and 0.235 for rational trials. A Mann-Whitney test found this tiny difference statistically significant ($U = 354484$, $p < .001$), but only because of the large sample size; the effect itself is negligible. The analytical defender was effectively guessing for both attacker types yet still won more trials than the others, so its defensive advantage came from something other than correct profiling.

To test H1's prediction, we ran a second logistic regression on analytical and paired trials, predicting attacker success from whether the defender was paired, the analytical agent's confidence, and their interaction. Higher confidence strongly reduced attacker success (OR = 0.01, $p < .001$), and the paired × confidence interaction was positive and significant (OR = 12.42, $p < .001$): the paired defender's advantage over the analytical defender shrinks as the machine grows more confident, consistent with H1's

predicted direction. However, isolating low-confidence trials directly, the analytical defender still won 98.4% of trials and the paired defender 97.4%, a non-significant difference ($z = 0.88$, $p = .381$). The paired defender did not actually outperform the analytical defender when confidence was low; both were losing at nearly identical rates.

Interventions

Given the analytical defender's chance-level inference, we ran a follow-up in which it actively searches the network's most diagnostic nodes when uncertain, rather than waiting for alerts to trigger on their own. Identification of the high-GRiPS attacker rose from 31% to 43% accuracy, confirming that the earlier chance-level inference was an observability problem rather than a fault in the updating rule. The rational attacker stayed essentially unidentifiable regardless of sensing (0% and 1% accuracy with sensing off and on), meaning its scoring formula left almost no detectable signal. Although identification increased, attacker win rates barely moved (78% to 81% for high-GRiPS and 70% to 78% for rational), so better identification did not translate into better defense.

We ran three parameter sweeps to test whether the pattern held under different conditions. Making the HVT nodes progressively more visible raised high-GRiPS identification from 30% to 62% (sensing off) and 38% to 66% (sensing on), yet the attacker win rate stayed flat at 84% without sensing and 82% with it across the entire range—a large identification gain with no defensive benefit. Placing honeypot decoys on the network's hub nodes cut the attacker win rate sharply, from 86% with none, to 21% with two, and to 0% with three or more. Identification fell toward zero, likely because trials ended too quickly for meaningful profiling. Finally, turning a DMZ decoy from a silent lure into a loud canary moved the average turn of first confident detection from 41.8 to 4.6 and improved identification, but the attacker win rate still did not fall.

Across all three sweeps, identifying the attacker earlier or more accurately did not, on its own, improve defender outcomes. Only the honeypot sweep, which was a direct change to the network rather than an attempt to read the attacker, produced a large drop in attacker success.

DISCUSSION

Our results anchor on a central finding: better identification of the attacker was not sufficient to translate into better defense, meaning against fast attackers, defenders identify the threat too late to act on it. The initial finding that the analytical defenders outperformed both the heuristic and paired defenders led us to deepen our research into what was actually driving that advantage.

Directly testing Hypothesis 1, we compared the paired and analytical defenders head-to-head under low confidence, where the paired was predicted to have a clearer advantage. We found no significant difference (98.4% vs. 97.4% attacker win rate), driving us to reject H1. This likely traces back to

the same observability problem that led our other results: since the analytical component's confidence rarely rose above chance, the paired defender spent most trials using the heuristic agent's fixed rules, limiting whatever advantage pairing could have offered. Our near-chance inference results point to a related limitation in the Bayesian approach to attacker profiling proposed by Hans et al. (2025): when a network's most diagnostic nodes are also its quietest, the profiler simply has too little informative evidence to work with, regardless of how the inference itself is structured. The H1 result reflects a limitation of this specific confidence-threshold design rather than evidence against human-machine teaming as a whole. It does align with other evidence that human-machine teams can lead to underperformance (Gurney, et al., 2025), although those results were driven by human overreliance on an AI teammate.

That pattern of cognitive profiling not being sufficient only held up across our behavioral interventions, which tested whether sharper detection alone could close the gap. Once we moved on to testing changes in structural defense, we found much better results for defenders. Changing the game board changed outcomes more drastically than changing the actions or beliefs of the defenders.

Human threat actors continually modify their behavior to achieve success. Past research demonstrates that their behavior is often laced with identifiable cognitive artifacts. We demonstrate some of the applied and operational implications of such in cyber defense. The real-world challenges of stopping an attacker, and the work to support that end-goal, must hinge around the ability to understand the reasoning behind the attack—what is the attacker after. How well your network and infrastructure topology protects these assets is the true test of your security program. Structural defenses are critical to supporting the overall architecture of a security program and general technology infrastructure, but as seen here, a motivated attacker can still get through the defenses if not properly managed and maintained.

LIMITATIONS

The inference model was calibrated to distinguish four different attacker types, but we tested it on only two. It is possible that calibrating it to distinguish two attacker types would change our results. The network topology is also a simplification, which is arguably why the honeypot intervention performed as well as it did. This outcome may be specific to this topology and may not generalize to networks with more distributed connectivity. Finally, both the attacker's and defender's action sets in this simulation are limited. Attackers can only choose where to move and defenders can only choose to monitor, harden, and/or block a fixed number of times per turn. Real attackers and defenders have a much wider range of options, which restricts how directly these results apply to real world settings.

CONCLUSION AND FUTURE WORK

Across a simulated attacker-defender contest, identifying an attacker proved necessary but not sufficient for stopping one. Improving the AI's perception

through active sensing, greater visibility at the network's quiet high-signal nodes, or earlier alarms reliably sharpened identification. However, none of these translated into fewer breaches; only a structural change to the network reduced attacker success. Future work should recalibrate the inference model for the specific attacker types under study, test a wider range of profiles, and pair structural defenses with faster-responding defenders. Above all, these results argue for treating response speed as co-equal with detection accuracy. It is helpful for a defender to know who is coming, but detection without the speed to act on it is not defense.

ACKNOWLEDGMENT

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-25-2-0040. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

Appelt, K.C., Milch, K.F., Handgraaf, M.J.J. and Weber, E.U. (2011) 'The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research,' *Judgment and Decision Making*, 6(3), pp. 252–262.

Beltz, B., Doty, J., Fonken, Y., Gurney, N., Israelsen, B., Lau, N., Marsella, S., Thomas, R., Trent, S., Wu, P. and Yang, Y.T., 2026. 'Guarding against malicious biased threats (GAMBiT) datasets: Revealing cognitive bias in human-subjects red-team cyber range operations.' *Data in Brief*, 112476.

Carney, S., Hans, S., Hirschmann, S., Marsella, S., Fonken, Y., Wu, P. and Gurney, N. (2025) 'Detecting ambiguity aversion in cyberattack behavior to inform cognitive defense strategies,' *AHFE International Conference, Hawaii*.

Ellsberg, D. (1961) 'Risk, ambiguity, and the Savage axioms,' *Quarterly Journal of Economics*, 75(4), pp. 643–669.

Fox, C.R. and Tversky, A. (1995) 'Ambiguity aversion and comparative ignorance,' *Quarterly Journal of Economics*, 110(3), pp. 585–603.

Frederick, S. (2005) 'Cognitive reflection and decision making,' Journal of Economic Perspectives, 19(4), pp. 25–42.

Gurney, N., Pynadath, D.V. and Miller, J.H., 2025. Willingness to work as a predictor of human-agent team success. *Frontiers in Computer Science*, *7*, 1405436.

Gutzwiller, R.S., Ferguson-Walter, K.J. and Fugate, S.J. (2019) 'Are cyber attackers thinking fast and slow? Exploratory analysis reveals evidence of decision-making biases in red teamers,' *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), pp. 427–431.

Hans, S., Hirschmann, S., Marsella, S. and Gurney, N. (2025) 'Security logs to ATT&CK insights: Leveraging LLMs for high-level threat understanding and cognitive trait inference,' *AHFE International Conference, Hawaii*.

Kahneman, D. and Tversky, A. (1979) 'Prospect theory: An analysis of decision under risk,' *Econometrica*, 47(2), pp. 263–291.

Kim, R., Carney, S., Fonken, Y., Hans, S., Hirschmann, S., Marsella, S., Wu, P. and Gurney, N. (2025) 'Risk psychology & cyber-attack tactics,' *AHFE International Conference, Hawaii*.

Simon, H.A. (1955) 'A behavioral model of rational choice,' *Quarterly Journal of Economics*, 69(1), pp. 99–118.

Stanovich, K.E. and West, R.F. (2000) 'Individual differences in reasoning: Implications for the rationality debate?' *Behavioral and Brain Sciences*, 23(5), pp. 645–665.

Strom, B. E., Applebaum, A., Miller, D. P., Nickels, K. C., Pennington, A. G. and Thomas, C. B. (2020) MITRE ATT&CK: Design and philosophy. McLean, VA: MITRE Corporation.

Toplak, M.E., West, R.F. and Stanovich, K.E. (2011) 'The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks,' *Memory & Cognition*, 39(7), pp. 1275–1289.

Tversky, A. and Kahneman, D. (1974) 'Judgment under uncertainty: Heuristics and biases,' *Science*, 185(4157), pp. 1124–1131.

Zhang, D. C., Highhouse, S. and Nye, C. D. (2019) 'Development and validation of the General Risk Propensity Scale (GRiPS),' *Journal of Behavioral Decision Making*, 32(2), pp. 152–167.