Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1921–1930
https://doi.org/10.54941/ahfe1008420

Psychological Foundations of Epistemic Profiles for Misinformation Resilience: Implications for Defence-Oriented Agent-Based Modelling

Petr Zámečník[1], David Černý[2], Vojtěch Juřík[3], and Darina Zámečníková Havlíčková[1]

[1]School of Arts, Humanities, and Social Sciences, Anglo-American University, Prague, Czech Republic
[2]Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic
[3]Department of Psychology, Faculty of Arts, Masaryk University, Brno, Czech Republic

ABSTRACT

This conceptual and methodological paper proposes a psychological and epistemological basis for agent-based modelling of misinformation resilience in defence-relevant information environments. We integrate bounded rationality, attention, threat appraisal, social identity and relational trust with intellectual humility, actively open-minded thinking and epistemic vigilance. Rather than classifying people into fixed susceptibility types, the proposed epistemic profiles distinguish enduring dispositions, context-dependent states and source-specific social relations. We specify separable behavioural stages—exposure, evaluation, belief updating, verification, sharing and correction—and offer a candidate mapping from theoretical constructs to agent parameters. Drawing cautiously on the Gateway Belief Model, we also propose credibility, relevance, efficacy and legitimacy as hypotheses concerning conditions for protective information behaviour, without assuming that climate-domain effects transfer to defence settings. Potential applications include testing defensive communication and verification procedures under time pressure, incomplete information and military organisational constraints. A staged validation agenda is outlined; no simulation results or validated psychological profiles are claimed. The proposed framework is developed in the context of Czech Technology Agency project TQ23000183.

Keywords: Misinformation, Cognitive resilience, Virtue epistemology, Gateway beliefs, Agent-based modelling, Defence

INTRODUCTION

The circulation of misinformation and coordinated information manipulation is a challenge not only for public debate but also for security organisations whose effectiveness depends on reliable information, confidence in legitimate institutions and collective decision-making. NATO defines information threats as intentional, harmful, manipulative and coordinated activities that can affect the Alliance and its members, and emphasises information-environment assessment as a basis for informed responses (NATO, 2024).

Received October 21, 2026; Revised November 26, 2026; Accepted December 4, 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/

The rapid amplification of content through digital networks, including AI-enabled synthetic media, increases the importance of understanding how people select, interpret and retransmit claims in uncertain or contested environments. The defence relevance therefore extends beyond the identification of false statements to their possible consequences for trust, coordination and resilience.

The psychology of misinformation cannot be reduced to a general deficit in intelligence or knowledge. Belief formation involves limited attention, source-based heuristics, prior beliefs, emotional salience and social relationships; moreover, misinformation may continue to influence reasoning even after correction (Ecker et al., 2022). Sharing and believing also need to be separated: users may propagate inaccurate information while failing to consider its accuracy at the moment of sharing (Pennycook et al., 2021). These differences are particularly relevant for simulation models, where treating all agents as interchangeable recipients or allowing a single parameter to determine both belief and transmission risks masking important behavioural mechanisms.

Computational approaches already incorporate network topology, homophily, topic-sensitive communication and adaptive sharing decisions. For example, Puri et al. (2024) developed a language-sensitive agent-based simulation grounded in an observed social network, illustrating the promise of empirically informed behavioural rules. The present paper does not claim that psychology has been absent from earlier models. Instead, it addresses the narrower methodological problem of translating psychological processes together with normatively relevant epistemic dispositions into explicit, distinguishable agent characteristics that could later be calibrated and tested. The philosophical work on epistemic virtues adds a dimension not identical to cognitive capacity: how individuals responsibly handle uncertainty, revise their views and evaluate reasons.

The proposed framework draws on Czech Technology Agency project TQ23000183, which integrates social epistemology, philosophy, psychology and agent-based modelling. The project's 2025 report documents conceptual work and exploratory network analyses; it does not establish validated epistemic profiles or effects of those profiles on simulation outcomes (TQ23000183 Project Consortium, 2025). Accordingly, the contribution offered here is a transparent and testable modelling specification, not an empirical evaluation.

PSYCHOLOGICAL FOUNDATIONS OF MISINFORMATION SUSCEPTIBILITY AND RESILIENCE

Bounded Rationality, Selective Attention and Epistemic Evaluation

Human information processing operates under limits of time, attention and accessible evidence. Bounded rationality implies that individuals rarely evaluate all available alternatives exhaustively; they often select responses that are sufficiently satisfactory within their cognitive and situational constraints. This becomes consequential in fast-moving information environments, where source familiarity, fluency, repetition and salient social cues may substitute for effortful verification. A useful account should avoid treating

heuristic processing as intrinsically irrational: it is frequently efficient, but its reliability depends on the informational environment and the cues available. The project's psychology presentations identify limited attention, satisficing and information overload as transferable mechanisms from environmental decision-making to misinformation contexts.

Evidence that analytic reflection predicts greater discrimination between true and false political news underlines the relevance of individual differences (Pennycook and Rand, 2019). At the same time, illusory truth effects and prior-belief congruence can make repeated or familiar claims appear credible without adequate verification (Ecker et al., 2022). For agent-based modelling, cognitive reflection, verification propensity, attentional constraints and sensitivity to familiarity should therefore be considered separately rather than collapsed into one global "rationality" score. Contextual factors such as time pressure or cognitive load may temporarily change how these dispositions are expressed.

Emotion, Threat Appraisal and Motivated Information Processing

Threat-related information competes successfully for attention because it is behaviourally consequential. Fear, anger and moral concern can increase salience, prompt rapid judgement and shape whether a message is passed on. Research on online moral discourse indicates that moral-emotional language may facilitate diffusion, especially within ideologically aligned networks (Brady et al., 2017). Yet emotional intensity should not be equated with false belief: emotionally relevant true warnings may also deserve rapid circulation. A defensible model must therefore distinguish emotional activation from accuracy and allow message characteristics to interact with individual dispositions.

Motivated cognition and cognitive dissonance offer complementary explanations of why corrective information is not uniformly effective. A correction may provide additional evidence while simultaneously threatening a valued identity or existing sense of coherence. Protective processing may then preserve prior positions, discount the source or postpone updating. This is especially plausible during crises, when the pressure for immediate and coherent explanations increases. Havlík and Horáček (2026) highlight cognitive rigidity and the need for closure among traits potentially relevant to weaponised information in security contexts. Their strategic analysis motivates candidate variables but should not be interpreted as proof that those characteristics deterministically predict susceptibility in military personnel.

Social Identity, Epistemic Trust and Group-Level Influences

People depend on others for knowledge about technically complex or distant phenomena. Institutional expertise, peer testimony and perceived source reliability are therefore not peripheral to reasoning; they are part of ordinary epistemic dependence. The meaning of a claim can change depending on who communicates it and whether the source is perceived as an in-group member, legitimate authority or adversary. In a study of scientific evidence evaluation, Rosman and Grösser (2024) found that trust in science was associated

with stronger belief updating and better discrimination of evidence quality, challenging the idea that trust necessarily entails uncritical acceptance.

Social environments can further regulate exposure and credibility. Nguyen (2020) distinguishes epistemic bubbles, in which certain voices are absent, from echo chambers, in which external voices are actively discredited. The distinction matters for modelling: adding an information link may alleviate missing exposure, but may have limited effect when the receiving agent regards the source as fundamentally untrustworthy. Accordingly, network position, relationship-specific trust, group identity and message content should form interacting rather than interchangeable components of an agent's informational environment.

Evolutionary Perspectives: Social Learning, Reputation and Coalition Cues

Evolutionary approaches provide candidate explanations for the salience of immediate threats, reputation, social belonging and cues from experienced or respected peers. Human reliance on social learning is not merely an individual shortcoming; it is an adaptive means of navigating complex environments. In digital settings, however, reputational markers and apparent popularity may be artificially amplified, and source cues may not reliably track expertise. The implications are best framed as testable hypotheses about attention and social transmission, rather than as universal biological rules about who believes misinformation. Potential computational proxies include sensitivity to peer endorsement, perceived authority, coalition congruence and reputational costs of public correction.

Gateway Beliefs: From Foundational Judgements to Downstream Action

The Gateway Belief Model (GBM) was developed to explain how perceptions of scientific consensus on climate change may influence more specific beliefs and, indirectly, support for collective action (van der Linden et al., 2015; van der Linden, 2021). Subsequent evidence includes a replication across 27 countries and a meta-analytic structural equation analysis; these studies support the importance of perceived consensus but also show that downstream changes are usually substantially smaller than corrections of consensus misperceptions (Većkalov et al., 2024; Rode et al., 2025). More recent work also suggests that misinformation can undermine gateway perceptions and subsequently affect attitudes (Logemann et al., 2026). The model should consequently not be reduced to the general claim that "the right frame persuades people": it specifies a proposed causal ordering that requires measurement and testing.

We propose an exploratory defence-oriented extension of this logic, not a transfer of the established climate effect sizes. Four candidate gateways deserve investigation: (G1) epistemic credibility, including confidence that the source and evidence are competent and accountable; (G2) relevance, including perceived plausibility and personal or collective significance of

the information threat; (G3) efficacy, including belief that verification or protective action can make a difference; and (G4) legitimacy and fairness, including whether the recommended institutional response is considered justified and proportionate. We hypothesise that these judgements shape downstream acceptance, willingness to verify, corrective updating and appropriate response behaviour, while their ordering may differ across contexts. In defence settings, legitimacy must never be operationalised as mere obedience, and scepticism towards an institution must not itself be classified as misinformation susceptibility.

The proposed chain can be expressed as: trusted and intelligible evidence → gateway judgements (credibility, threat relevance, efficacy, legitimacy) → issue-specific beliefs and emotional appraisals → information behaviour and protective action. This sequence is an analytical hypothesis that may contain feedback loops and be moderated by identity, group norms, uncertainty and mission context. It offers a useful complement to agent models based solely on exposure, belief thresholds or social influence, provided that the gateways can be operationalised independently and their incremental explanatory value established.

EPISTEMIC VIRTUES, VICES AND EVIDENCE-SENSITIVE JUDGEMENT

Virtue epistemology considers not only the accuracy of particular beliefs but also the intellectual dispositions through which individuals seek, evaluate and revise evidence (Zagzebski, 1996; Baehr, 2011). This perspective complements cognitive psychology without equating responsible inquiry with intelligence, educational attainment or mere conformity to expert opinion. In the present framework, virtues and vices are proposed as candidate dispositions influencing epistemic practices, not labels identifying morally superior or inferior groups.

Intellectual humility refers to recognising the fallibility of one's own beliefs, while actively open-minded thinking includes considering alternatives, attending to disconfirming evidence and delaying premature closure (Leary et al., 2017; Stanovich and Toplak, 2023). Neither entails indiscriminate credulity. Epistemic vigilance, conversely, captures the need to scrutinise communicated claims and their sources (Sperber et al., 2010). A defender who accepts every correction without weighing the evidence may be no more resilient than one who rejects all corrections. Intellectual conscientiousness can therefore be represented, tentatively, as a propensity to inspect provenance and warrant when circumstances permit.

Potential vices include dogmatic closure, selective dismissal of counterevidence and epistemic negligence (Cassam, 2019). These must be distinguished from justified scepticism, especially where real deception, restricted evidence or conflicting authoritative communications exist. Empirical evidence is qualified: in two studies, intellectual humility predicted greater exploration of counter-information, but not a consistent direct improvement in false-belief revision (Gollwitzer et al., 2024). Recent experiments suggest that prebunking which promotes actively open-minded thinking may improve discernment

indirectly (Biddlestone et al., 2025). Thus, epistemic characteristics should moderate specific behaviours and be tested against alternative explanations; their causal influence cannot simply be stipulated.

INTEGRATING PSYCHOLOGICAL AND EPISTEMIC AGENT PROFILES

We define an epistemic agent profile as a multidimensional, adjustable specification of how a simulated individual attends to, evaluates and acts on information. The proposal distinguishes: (a) relatively durable dispositions (e.g., cognitive reflection, openness to counterevidence, tolerance of uncertainty); (b) dynamic states (e.g., current threat appraisal, cognitive load, identity salience); and (c) relational and environmental conditions (e.g., trust in a particular sender, network position, perceived group norms). Profiles need not be fixed categories of persons. Values should be allowed to vary continuously and to interact across situations; the same agent may respond differently to different topics, sources and operational conditions.

A minimally interpretable process separates exposure from attention, credibility judgement, belief updating, verification, sharing and subsequent correction. Each stage can be parameterised independently. For instance, source-specific trust may alter the evidential weight of a claim, whereas time pressure may alter whether its provenance is checked. Likewise, apparent social endorsement may increase forwarding without increasing confidence in accuracy (Pennycook et al., 2021). This distinction avoids treating all diffusion as belief adoption. Table 1 describes proposed relationships; its entries are hypotheses and design requirements, not estimated or empirically validated effects.

We further distinguish the original climate-science Gateway Belief Model—which examines how perceived scientific consensus may shape downstream beliefs—from our proposed, unvalidated extension to defence-related information. Candidate gateway judgements are: (G1) credibility of available evidence and sources; (G2) perceived relevance of a threat; (G3) efficacy of feasible protective actions; and (G4) perceived legitimacy and procedural fairness of those actions. These judgements are not a necessary linear chain. They may act as moderators, mediators or independent inputs, and they may produce distinct results under routine, uncertain and crisis conditions (van der Linden et al., 2015; Rode et al., 2025).

A parsimonious implementation would estimate, at each time step, separate probabilities of verification, belief revision and forwarding conditional on the agent's disposition vector, current context, message attributes and relation to its source. Parameters should be calibrated using independent behavioural observations or established measures where possible, with plausible ranges explicitly labelled when such evidence is unavailable. Normative terms (e.g., "virtuous", "resilient") must not be equated mechanically with a chosen political viewpoint or uncritical compliance with official statements. Accuracy, justified confidence and proportionate source scrutiny provide preferable outcome criteria.

Table 1: Proposed operationalisation of psychological and epistemic mechanisms. All effects are hypotheses for future validation.

Construct / Level	Candidate Agent Variable	Testable Behavioural Role
Reflective processing / disposition	Verification propensity; evaluation cost	Likelihood of checking claim accuracy under available time.
Intellectual humility / disposition	Fallibility acknowledgement; confidence flexibility	Likelihood of examining credible counterevidence; no assumed direct effect on accuracy.
Open-mindedness / disposition	Counterevidence consideration; closure threshold	Delay of premature belief closure; responsiveness to diagnostic evidence.
Epistemic vigilance / disposition	Source scrutiny; provenance-check threshold	Differentiates justified scepticism from indiscriminate distrust.
Threat appraisal / state	Current perceived threat; urgency	Allocation of attention and urgency of response.
Identity and social norms / context	Message–identity fit; normative sharing signal	Potential moderation of belief updating and forwarding.
Source trust / relation	Trust by agent–source pair	Weights source testimony independently of network exposure.
Gateway judgements / context	Credibility, relevance, efficacy, legitimacy	Competing pathways from information exposure to protective action.
Military workload / context	Time pressure; access restrictions	Constrains verification despite otherwise favourable dispositions.

DEFENCE-ORIENTED APPLICATIONS AND SECTOR-SPECIFIC CONSTRAINTS

Defence information environments differ from ordinary social-media settings. Information can affect readiness, coordination, trust in lawful command and communication with civilian communities. Decision-makers may face compressed timelines, incomplete or classified evidence, disrupted channels and asymmetric knowledge across military units, civilian employees, allied organisations and the public. The proposed framework is therefore directed at protective resilience, not at identifying supposedly disloyal persons or optimising influence operations. NATO defines information threats in terms of coordinated, manipulative, intentionally harmful behaviour and emphasises understanding, prevention, mitigation and recovery while protecting freedom of expression (NATO, 2024).

In military settings, trust has several distinct referents: an authenticated operational message, a commander's professional credibility, an institutional spokesperson, an allied partner and informal peer networks should not share a single scalar "authority" value. A high-trust chain of command may facilitate timely coordination, yet perceived source legitimacy does not make an unverified claim factually correct. Equally, restricted access to evidence can make ordinary civilian verification practices impossible. The model should

therefore distinguish evidence provenance, perceived authority, authorised communication channels and individual access to verification resources. Trust calibration, rather than unconditional obedience or unconditional scepticism, is the defensible target.

Safe, fictional defensive scenarios could examine reactions to conflicting public and internal advisories during an emergency, delayed correction of a false administrative alert, or doubts about the authenticity of a communication attributed to a defence institution. Each scenario would vary only high-level contextual conditions—time pressure, information availability, cross-unit connectivity, source authentication and group trust—to compare resilient and fragile communication pathways. Outcomes might include time to recognise a false claim, reach of unverified forwarding, reliance on authenticated channels, appropriate escalation for verification and recovery of justified confidence after correction. No real units, personnel records, classified communications or current operational vulnerabilities are required.

Candidate countermeasures include clear provenance of official messages, rapid and accountable correction, training in recognising manipulative tactics, accuracy prompts and pre-emptive communication about anticipated uncertainty. Such interventions can be compared in silico, but a simulation cannot establish their operational effectiveness without independent validation. The framework should also record possible adverse effects, including overconfidence, indiscriminate suspicion of legitimate communications and disruption of timely mission-critical information. Defence applications require explicit human oversight, scenario transparency, data minimisation and restrictions against profiling identifiable individuals for political persuasion (Havlík and Horáček, 2026; NATO, 2024).

RESEARCH PROPOSITIONS AND VALIDATION ROADMAP

No completed simulation, empirically estimated epistemic profiles or effect estimates are reported here. The framework yields three falsifiable research propositions: P1, adding psychological heterogeneity changes diffusion outcomes relative to a homogeneous baseline; P2, incorporating epistemic dispositions improves explanatory fit beyond cognitive and network parameters alone; P3, the effect of a defensive communication intervention varies with source trust, gateway judgements and situational workload. These claims may be unsupported or falsified; the framework is designed to allow that possibility.

A staged evaluation would first specify baseline models and preregister parameter definitions. Second, behavioural tasks and available validated measures (e.g., intellectual humility and actively open-minded thinking) would inform parameter ranges; network data would inform topology rather than stand in for psychological measurements. Third, sensitivity analyses and held-out comparison data would test robustness, with particular attention to correlations among dispositions, measurement error and contextual change. Finally, benign military-relevant scenarios could be reviewed by

domain experts for construct validity while protecting personnel privacy and operational security. Until these steps are completed, policy and training implications remain provisional.

DISCUSSION AND CONCLUSION

The proposed contribution is a bridge between misinformation psychology, virtue epistemology and defence-oriented computational modelling. Its methodological value lies in decomposing apparently unitary "susceptibility" into distinguishable stages and conditions, and in identifying variables whose inclusion can be contested and empirically tested. The proposed gateway-belief extension is especially exploratory: relevance, efficacy, legitimacy and source credibility may matter without forming a single causal sequence, and transfer from climate communication cannot be assumed.

Important limitations include incomplete operationalisation of normative virtues, the risk of circular calibration, potential confounding between trust and identity, and uncertainty about generalisation from civilian digital environments to military organisations. Future work should estimate parameters, compare nested models and test resilience outcomes using transparent, non-sensitive scenarios. The appropriate endpoint is not a model that ranks people as virtuous or credulous but a validated instrument for examining how reliable communication and evidence-sensitive judgement can be supported under defensive constraints.

ACKNOWLEDGMENT

This research was conducted as part of project No. TQ23000183, "Modeling the Spread of Mis- and Disinformation: An Agent-Based Model", co-financed from the state budget by the Technology Agency of the Czech Republic under the SIGMA Programme.

REFERENCES

Baehr, J. (2011) The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology. Oxford: Oxford University Press.

Biddlestone, M., Ziemer, C.-T., Maertens, R., Roozenbeek, J. and van der Linden, S. (2025) "Norm-enhanced prebunking for actively open-minded thinking indirectly improves misinformation discernment and reduces conspiracy beliefs", Journal of Experimental Social Psychology, 121, 104818. doi: 10.1016/j.jesp.2025.104818.

Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A. and Van Bavel, J.J. (2017) "Emotion shapes the diffusion of moralized content in social networks", Proceedings of the National Academy of Sciences, 114(28), pp. 7313–7318. doi: 10.1073/pnas.1618923114.

Cassam, Q. (2019) Vices of the Mind: From the Intellectual to the Political. Oxford: Oxford University Press.

Ecker, U.K.H., Lewandowsky, S., Cook, J. et al. (2022) "The psychological drivers of misinformation belief and its resistance to correction", Nature Reviews Psychology, 1, pp. 13–29. doi: 10.1038/s44159-021-00006-y.

Gollwitzer, A., Bao, E. and Oettingen, G. (2024) "Intellectual humility as a tool to combat false beliefs: An individual-based approach to belief revision", British Journal of Social Psychology, 63(3), pp. 1450–1478. doi: 10.1111/bjso.12732.

Havlík, M. and Horáček, J. (2026) War is a Mind Game: Countering Weaponised Information. NDC Outlook 2026-01. Rome: NATO Defense College. Available at: https://www.ndc.nato.int/war-is-a-mind-game-countering-weaponised-information/.

Leary, M.R., Diebels, K.J., Davisson, E.K. et al. (2017) "Cognitive and interpersonal features of intellectual humility", Personality and Social Psychology Bulletin, 43(6), pp. 793–813. doi: 10.1177/0146167217697695.

NATO (2024) NATO's Approach to Counter Information Threats. Endorsed by Allied Defence Ministers, 18 October. Available at: https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/10/18/natos-approach-to-counter-information-threats.

Nguyen, C.T. (2020) "Echo chambers and epistemic bubbles", Episteme, 17(2), pp. 141–161. doi: 10.1017/epi.2018.32.

Pennycook, G. and Rand, D.G. (2019) "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning", Cognition, 188, pp. 39–50. doi: 10.1016/j.cognition.2018.06.011.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D. and Rand, D.G. (2021) "Shifting attention to accuracy can reduce misinformation online", Nature, 592, pp. 590–595. doi: 10.1038/s41586-021-03344-2.

Puri, P., Hassler, G., Katragadda, S. and Shenk, A. (2024) "Digital cloning of online social networks for language-sensitive agent-based modeling of misinformation spread", PLOS ONE, 19(6), e0304889. doi: 10.1371/journal.pone.0304889.

Rode, J.B., Remshard, M., Groot, L.J. and van der Linden, S. (2025) "A meta-analytic structural equation analysis of the Gateway Belief Model: Highlighting scientific consensus increases support for public action on climate change", Current Opinion in Behavioral Sciences, 63, 101521. doi:10.1016/j.cobeha.2025.101521.

Rosman, T. and Grösser, S. (2024) "Belief updating when confronted with scientific evidence: Examining the role of trust in science", Public Understanding of Science, 33(3), pp. 308–324. doi: 10.1177/09636625231203538.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. and Wilson, D. (2010) "Epistemic vigilance", Mind & Language, 25(4), pp. 359–393. doi: 10.1111/j.1468-0017.2010.01394.x.

Stanovich, K.E. and Toplak, M.E. (2023) "Actively open-minded thinking and its measurement", Journal of Intelligence, 11(2), 27. doi:10.3390/jintelligence11020027.

TQ23000183 Project Consortium (2025) Modelování šíření mis- a dezinformací: Odborná zpráva o řešení projektu za rok 2025. Internal project report [unpublished].

van der Linden, S. (2021) "The Gateway Belief Model (GBM): A review and research agenda for communicating the scientific consensus on climate change", Current Opinion in Psychology, 42, pp. 7–12. doi: 10.1016/j.copsyc.2021.01.005.

van der Linden, S., Leiserowitz, A.A., Feinberg, G.D. and Maibach, E.W. (2015) "The scientific consensus on climate change as a gateway belief: Experimental evidence", PLOS ONE, 10(2), e0118489. doi: 10.1371/journal.pone.0118489.

Zagzebski, L.T. (1996) Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge. Cambridge: Cambridge University Press.