From Individual Trust to Collective Knowledge: Modeling Epistemic Vigilance

Mélinda Pozzi

Sciences Normes Démocratie (SND), Sorbonne University, Paris, France
Department of Logic and Philosophy of Science, UC Irvine, USA

1 Introduction

Human communication relies on cognitive mechanisms for *epistemic vigilance* toward misinformation [5]. While the cognitive underpinnings of this process are well-studied, its collective consequences in social networks are less understood. It is known that network topology critically shapes collective outcomes [4, 6], but the interplay with individual cognitive strategies remains an open question. This paper applies a formal agent-based model of trust reasoning to investigate how a key vigilance strategy—adapting trust by tracking sources' past accuracy [2, 3]—scales up to shape group-level knowledge. We hypothesize that vigilance improves collective performance, but that it is less effective in large or dynamic networks.

Adapting the self-assembling network model [1], we introduce agents with binary beliefs that face 100 empirical problems, start each in a state of ignorance (inaccurate belief = 0), and have 100 rounds to solve the problem (by acquiring an accurate belief = 1). Each agent has a fixed random reliability between 0 and 1 (i.e., their success rate in observing nature). Decisions follow a reinforcement-learning "urn" model, which simulates the cognitive process of updating trust: each agent's urn starts with one ball for nature, one for each other agent, and one for itself; balls are added after reinforcement.

During each round, agents took turns in a random order. Observing nature yields a correct belief with probability equal to reliability and is reinforced upon success. Consulting oneself maintains the current belief and is reinforced if accurate. When consulting other agents, their belief is adopted and reinforcement depends on vigilance (V), a parameter representing the strength of this cognitive filter, ranging from 0.0 to 1.0 (the same for all agents) under two scenarios. In the *credulous* scenario, agents always reinforce accurate social sources, and vigilance reduces their credulity toward inaccurate ones: at $V = 0.0$ all inaccurate agents are reinforced, at $V = 1.0$ none are. In the *skeptical* scenario, agents never reinforce inaccurate social sources, and vigilance reduces their skepticism toward accurate ones: at $V = 0.0$ no accurate agents are reinforced, at $V = 1.0$ all are. In both scenarios, fully vigilant agents reinforce only accurate social sources. To capture network dynamics, agents are replaced with a 0.05 probability each round, testing stable vs. dynamic networks of $N \in \{5, 50, 100\}$ agents. Collective knowledge is the average proportion of agents holding an accurate belief during the last $T_{\mathrm{w}} = 20$ rounds across $S = 1000$ simulations:

$$\bar{K} = \frac{1}{S} \sum_{s=1}^{S} \left(\frac{1}{T_{\mathrm{w}}} \sum_{t=T-T_{\mathrm{w}}+1}^{T} \frac{A_{s,t}}{N} \right) \tag{1}$$

where $\bar{K}$ is the mean collective knowledge, S the number of simulations, T the last round, T_{w} the time window considered, $A_{s,t}$ the number of agents with an accurate belief (i.e., belief $= 1$) in simulation s at round t, and N the total number of agents.

2 Results

Implementing full epistemic vigilance as a cognitive strategy reshaped learning dynamics, forming a functional network where reliable agents primarily engaged in *independent learning* (observing nature) and exerted stronger *social influence* (were consulted more), while less reliable agents learned from them. This emergent structure contrasted sharply with credulous networks (where all agents exercised the same social influence) and skeptical ones (where they favored self-consultation or independent learning). In the credulous scenario, small stable groups ($N = 5$) achieved an accuracy ($\bar{K}$) of 0.64, which dropped to 0.24 in large groups ($N = 100$) when vigilance was absent ($V = 0.0$, Fig. 1). Notably, performance degraded with increasing size as reliable agents had fewer chances to observe nature. Reducing credulity reversed this trend, improving outcomes to 0.85 in large stable networks ($N = 100$) at $V = 1.0$. In dynamic networks, vigilance still improved credulous performance, but less so, as agent turnover hampered the ability to track reliable agents. Further analysis showed that vigilance improved the individual performance of all credulous agents.

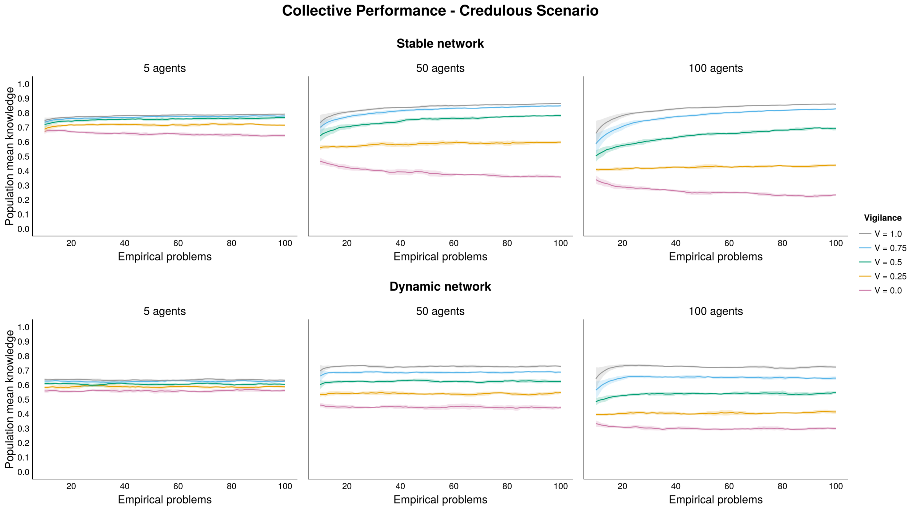

Fig. 1. Collective knowledge ($\bar{K}$) over successive empirical problems as a function of epistemic vigilance (V) in the *credulous* scenario.

In contrast, skeptical agents in small groups ($N = 5$) reached an accuracy ($\bar{K}$) of only 0.52, but scaled up to 0.61 in large groups ($N = 100$) without vigilance ($V = 0.0$, Fig. 2). Skeptical performance scaled positively with size, as more reliable agents reinforced themselves and nature, also giving more chances for unreliable agents to encounter other accurate agents and self-consult afterwards. Despite an initial performance lag in large groups, reducing skepticism ultimately improved outcomes in stable networks, yielding a performance of 0.86 at $V = 1.0$ ($N = 100$). In dynamic skeptical networks, this initial lag stabilized quickly, making full vigilance slightly less efficient than partial skepticism, although not significantly (0.74 at $V = 1.0$ vs. 0.79 at $V = 0.5$, $N = 100$). Further analyses highlighted a key tension: while vigilance benefited only less reliable agents in stable networks, the same strategy harmed experts (as they observed nature less and relied more on other agents) in large and dynamic networks, and eventually the group as a whole in dynamic settings.

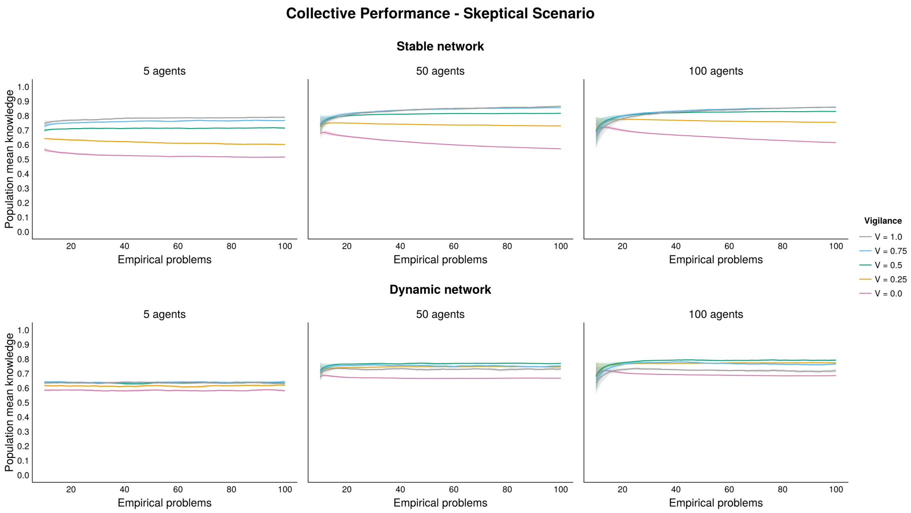

Fig. 2. Collective knowledge ($\bar{K}$) over successive empirical problems as a function of epistemic vigilance (V) in the *skeptical* scenario.

Surprisingly, full vigilance was more effective in large groups, suggesting the benefit of a larger expert pool outweighed the challenge of tracking more sources (e.g., in stable skeptical networks, accuracy reached 0.86 for $N = 100$ vs. 0.78 for $N = 5$). However, performance was ultimately bounded by the agents' fixed and imperfect reliability. We tested a broader range of network sizes and replacement probabilities to assess robustness, and the qualitative patterns of how vigilance influences collective knowledge remained consistent across these variations.

Future work should address simplifying assumptions in the current model. First, while agents vary in reliability, they all share identical levels of vigilance. Although homogeneity allows us to isolate the effects of an idealized strategy, modeling heterogeneous communities with mixed levels of vigilance would be more realistic. Second, agents exhibit different levels of vigilance, thus capturing realistic errors, but they retain perfect long-term memory; introducing memory decay could reveal how forgetting sources' past accuracy might affect vigilance. Third, the model assumes a fully emergent network where all agents are equally likely to interact, whereas real-world networks are often imposed and centralized. Comparing emergent versus predefined structures could clarify how network constraints affect vigilance. Finally, vigilance is currently cost-free; adding a cognitive cost that scales with the level of vigilance would allow us to determine when the effort of being vigilant outweighs its benefits.

While we cannot infer direct real-world behavior from this idealized model, the findings provide insights for managing collective knowledge and reducing misinformation. Epistemic vigilance can be seen as an epistemic virtue, reducing the spread of misinformation—especially for less reliable agents who depend more on others. In contrast, the most reliable sources of information may exercise virtue by maintaining skepticism. However, in real social communication, tracking the accuracy of sources is not always possible, agents may not always have perfect vigilance, networks can be unstable, and individuals may have limited choice over whom they consult; all of which constrain the effectiveness of vigilance.

Summary. This formal model of trust reasoning shows that when individuals must divide attention across multiple problems under time constraints and cannot improve their intrinsic reliability, the implementation of a simple heuristic for trust (i.e., adapting trust based on sources' past accuracy) is essential for collective knowledge. However, reducing skepticism reveals a key tension: a strategy that benefited less reliable agents in stable networks harmed the group's experts in large and dynamic networks, and ultimately collective performance in dynamic settings. This highlights the deep interplay between network structure and members' epistemic strategies, when all agents share the same level of vigilance.

References

1. Barrett, J.A., Skyrms, B., Mohseni, A.: Self-Assembling Networks. The British Journal for the Philosophy of Science **70**(1), 301–325 (Mar 2019). https://doi.org/10.1093/bjps/axx039
2. Douven, I., Schurz, G.: Integrating individual and social learning: Accuracy and evolutionary viability. Computational and Mathematical Organization Theory **30**(1), 32–74 (Mar 2024). https://doi.org/10.1007/s10588-022-09372-1
3. Koenig, M.A., Clément, F., Harris, P.L.: Trust in Testimony: Children's Use of True and False Statements. Psychological Science **15**(10), 694–698 (Oct 2004). https://doi.org/10.1111/j.0956-7976.2004.00742.x
4. O'Connor, C., Weatherall, J.O.: Scientific polarization. European Journal for Philosophy of Science **8**(3), 855–875 (Oct 2018). https://doi.org/10.1007/s13194-018-0213-9
5. Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., Wilson, D.: Epistemic Vigilance. Mind & Language **25**(4), 359–393 (Aug 2010). https://doi.org/10.1111/j.1468-0017.2010.01394.x
6. Zollman, K.J.S.: The Epistemic Benefit of Transient Diversity. Erkenntnis **72**(1), 17–35 (Jan 2010). https://doi.org/10.1007/s10670-009-9194-6