Alister, M. & Perfors, A.(to appear). Stochastic search algorithms can tell us who to trust (and why). *Proc. of the 47th Annual Conference of the Cognitive Science Society.*

Stochastic search algorithms can tell us who to trust (and why)

Manikya Alister (alisterm@student.unimelb.edu.au)
Andrew Perfors (andrew.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

Abstract

Relying on information from other people (social testimony) is essential for efficiently learning and reasoning about the world. However, determining who to trust is often challenging. In this paper, we argue that trust in social agents (i.e., those providing testimony) can be evaluated by assessing how optimally they have acquired their knowledge. Building on theories that describe knowledge acquisition as a stochastic search through a space of hypotheses, we present a framework which yields predictions about which agents will provide better testimony (because they are more likely to have uncovered higher-probability hypotheses) in different contexts. This approach allows us to jointly predict how the quality of testimony is affected by 1) features of the agents themselves, like their expertise; 2) consensus among multiple agents; and 3) features of the topic and hypothesis space, like its knowability. We present initial simulations demonstrating how even a basic implementation of our framework yields insight into which types of agents and topics are more likely to result in accurate testimony (and why). We conclude by discussing how this preliminary research might be extended to address more complicated social reasoning scenarios. **Keywords**: social reasoning, trust, exploration, persuasion, search

Introduction

When deciding what to believe, we often look to other people. Our reliance on social testimony is largely adaptive because it allows us to learn efficiently in the face of intractable amounts of data which we often do not have the expertise or ability to access firsthand. However, effective social learning requires knowing how to trust: trusting the wrong people can lead to disastrous outcomes, and lack of societal consensus about who to trust are one of the roots of problems such as extremist political movements, denigration of the value of expertise, and ideological polarisation.

In this paper, we propose that social reasoning can be productively viewed as reasoning about the *search* processes of other agents. Drawing from theories that conceptualise learning as the process of searching for knowledge, we can understand an agent's chances of arriving at the "ground truth" in terms of how effectively it is able to search the space of possible hypotheses. This conceptual lens permits us to map features of social informants and reasoning topics to features of stochastic search algorithms and hypothesis spaces – and, in so doing, enables us to apply insights from the latter to the former. In that vein, we present simulations that demonstrate how this approach can help us understand which agents should be trusted most, in which contexts, and why.

Background

We focus here on three factors that previous experimental research has shown are highly influential in social reasoning. First, **people care about the credibility of other agents**: we are more persuaded by experts as well as those without obvious conflicts of interest or bias (Orchinik et al., 2024; Simmonds et al., 2023). Second, **people care about consensus** when reasoning about specific claims: we find it more persuasive when multiple others endorse a claim than when a single person does so or there is disagreement (for reviews, see Mercier & Morin, 2019; Oktar & Lombrozo, 2025). Third, **people care about features of the reasoning topic itself**, independent of expertise or topic familiarity (Richardson & Keil, 2022): we are less likely to be persuaded by others when the claim is less likely to have a ground truth in the first place, as in the case of opinions (Alister et al., 2025).

Although these factors are important to social reasoning, our theoretical understanding of *why* they are important (as well as how they interact with each other) is limited. This is partly because existing computational models have yet to incorporate all of these aspects into a unified framework. Some normative models explain why a consensus should be persuasive (e.g., Montgomery et al., 2024; Oktar & Lombrozo, 2025; Romeijn & Atkinson, 2011), but they primarily focus on how opinions are aggregated (e.g., the proportion of responses for vs. against, or the distribution of responses on a continuous scale). There has been less consideration of how features of the agents who contribute to the consensus affect its persuasiveness (or should do so, or why). Although some models consider some features like the reliability of agents or the independence of their testimony (Harris et al., 2016; Madsen et al., 2020; Pilditch et al., 2020; Whalen et al., 2018), there is less consideration of other influential factors like the biases of the agents or the diversity of ideas considered by the agents.The common thread is that these models overlook how social reasoning is shaped by the *knowledge acquisition processes* of the agents who are providing testimony – a key theoretical consideration for understanding social influence.

Another shortcoming of existing models is that they focus more on features of the agents than the reasoning topic itself. For instance, we know that the persuasiveness of social testimony varies substantially based on whether the topic is likely to have a ground truth, regardless of the reasoners' prior beliefs (Alister et al., 2025). Although you may not personally know whether there was an armed robbery at the local shopping centre, there *is* a ground truth about what exactly happened and it is highly probable that someone knows it. Conversely, subjective claims like "cats are better pets than dogs" are generally not viewed as having an objective truth; they are largely a matter of individual preference. The importance of knowability is evident in the fact that one disinformation tactic is to reduce the extent to which a topic is perceived as knowable; for example, to discredit the fact that there is a scientific consensus on climate change (Readfearn, 2022). However, current models of social reasoning neither incorporate knowability nor explain how (and why) it should matter.

Social reasoning as reasoning about search

In the previous section, we highlighted three key limitations of existing computational models of social reasoning. They do not jointly incorporate, or provide an explanation of, 1) the features of individual agents that indicate their ability to know the truth; 2) how people integrate testimony from multi-

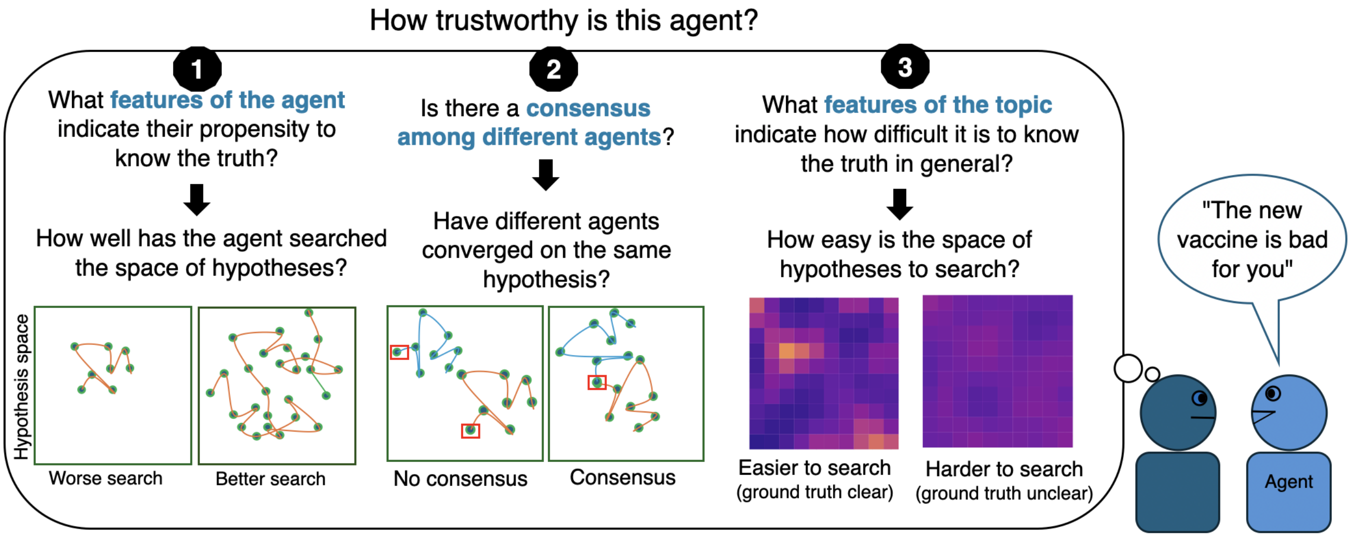
How trustworthy is this agent?
1
What features of the agent indicate their propensity to know the truth?
How well has the agent searched the space of hypotheses?
Hypothesis space
Worse search
Better search
2
Is there a consensus among different agents?
Have different agents converged on the same hypothesis?
No consensus
Consensus
3
What features of the topic indicate how difficult it is to know the truth in general?
How easy is the space of hypotheses to search?
Easier to search (ground truth clear)
Harder to search (ground truth unclear)
"The new vaccine is bad for you"
Agent

Figure 1: **Framework overview.** People decide how much they can trust other agents by taking into account 1) features that indicate how the agents acquired their knowledge (their search process); 2) how many other agents agree, and what the features of those agents are; and 3) indications that reflect how knowable the topic is or how easy the space is to search.

ple agents who might vary on these features; and 3) the nature of the topic being reasoned about, like its knowability (the extent to which a ground truth exists). Our framework addresses these limitations by considering how agents *acquire their knowledge*.

Our work emerges out of a popular theoretical perspective in cognitive science (e.g., Gopnik et al., 2017; Ullman et al., 2012) and philosophy of science (e.g., Alexander et al., 2015) which suggests that people acquire knowledge by exploring over the space of possible ideas or explanations. In other words, these theories describe knowledge acquisition as a stochastic search through a set of hypotheses, which people encounter and evaluate over time. This perspective has much in common with computational search algorithms, which resemble aspects of how people explore and test hypotheses during the course of learning (Bonawitz et al., 2014; Bramley et al., 2017; Giron et al., 2023). These algorithms move within hypothesis spaces much like people explore ideas, discarding less promising ones and retaining better ones.

Viewing knowledge acquisition as a search process suggests that both the social informant's search strategy and the structure of the reasoning topic should shape how effectively an informant can arrive at the truth. Just as stochastic search algorithms vary in their efficiency and convergence depending on features of the algorithm and the structure of the hypothesis space, we suggest that agents vary in their ability to uncover the truth depending on factors like their cognitive biases, resources, and the nature of the topic. For instance, knowability could reflect how the hypotheses in a space are distributed, with certain distributions being harder to search and therefore the truth of that topic more difficult to know (see also, e.g., Alexander et al., 2015).

To summarise, our framework suggests that people should decide how much to trust other agents by attending to features that reflect their search process. Therefore, people should be more persuaded by agents whose search is more likely to have uncovered the truth.

Specifically, as outlined in Figure 1, people need to consider 1) what features of that agent indicate their *individual* ability to search the space (e.g., their expertise, bias); 2) whether and how *multiple* agents have converged on the same hypothesis (e.g., consensus) and 3) what features of the reasoning topic affect how easy it would be to search the space of hypotheses (e.g., knowability). Table 1 lists some specific examples of these three aspects of social reasoning and describes how they may map onto features of search algorithms. Because our work is still in its infancy, this space of possibilities is (ironically) not yet fully explored. The table demonstrates the potential of our framework in formalising and predicting whose dynamics can be understood through modelling and then empirically tested on real-world phenomena. In the next section, we conduct some basic simulations that formalise some of these features and show how they can be used to make predictions about social reasoning.

Simulations

In this section, we demonstrate how our framework can be used to make normative predictions about which features of agents and hypothesis spaces result in a higher likelihood of uncovering better hypotheses and therefore, how much we should trust agents in different contexts. These simulations are intended as a proof of concept, showing how even a relatively simple hypothesis space and search algorithm make novel predictions and add conceptual clarity about the different factors that may matter in social reasoning (as well as how they might interact). In the general discussion, we will consider how our preliminary work might be extended to model more sophisticated reasoning and more complex scenarios.

Table 1: **Mapping features of social reasoning onto features of search**. This is not an exhaustive list of all possible features, nor even necessarily the most accurate mappings; however, the range of features and the potential insight derived from considering mappings like these demonstrate the utility of conceptualising social reasoning in this way. Each feature is grouped by how it maps onto the broad aspects of social reasoning our framework is trying to integrate (see Figure 1).

	Feature of social reasoning	**Parameter**	**Description**
Individual Agent	1 Expertise	Number of iterations or samples	All else being equal, agents who search longer end up considering more different hypotheses. In this way, these agents "know" more, i.e., have more expertise.
	2 Bias	Probability of accepting samples in different areas of space	Some people may be more likely to accept or reject a hypothesis based on some factor other than objective quality (e.g., similarity to current hypothesis or performance according to some other metric). Unbiased agents who choose hypotheses based solely on relative probability can search the space better.
	3 Diversity of ideas that are considered	Step size	Some people evaluate a wide variety of different ideas, and some do not. Smaller steps mean that at each iteration, an agent only sees hypotheses that are similar to what they already believe. Larger steps allow for greater exploration but may miss more fine-grained changes in hypothesis quality.
	4 Exploration/exploitation over time	Cooling parameter	People vary in their willingness to accept *any* hypotheses (explore), compared to only accepting good ones (exploit). Good search often involves more exploration in the early stages (i.e., childhood) and more exploitation later on (i.e., adulthood).
	5 Scepticism	Acceptance rule (i.e., probability of accepting new hypotheses)	Some people are very sceptical and will reject nearly every new idea. Others are much less sceptical and are highly likely to accept new hypotheses, even if lower in quality. Trustworthy reasoning is balanced, neither accepting nor rejecting too much.
Consensus	6 Consensus from multiple sources	Convergence of multiple chains	People sometimes agree with each other about which idea is right. When multiple agents settle on the same hypothesis, this agreement can be more indicative of ground truth.
	7 Independence among multiple sources	Starting point of chains and diversity of search paths	People vary in their initial beliefs as well as the directions they investigate, which influences their search path. If multiple agents with different paths arrive at similar conclusions, that can be taken as evidence in support of that conclusion.
Topic	8 Knowability: extent to which ground truth exists	Variation in hypothesis probability	For some topics, some hypotheses are obviously more correct than others: the sky is blue, not red. In these, the ground truth is more apparent and agents will find it easier to converge to the correct hypothesis. In topics where all hypotheses have similar likelihood, agents will not be as likely to converge to the correct conclusion.
	9 Knowability: similarity of best hypotheses	Spatial correlations of hypotheses	For some topics, better hypotheses may cluster together because they are close in the underlying representational space. In other topics, there are many local minima, and good hypotheses may be far from each other. It is easier for agents to identify the best hypothesis when the correlational structure assists search.

Method

Hypothesis spaces. To model each hypothesis space, we used a 10×10 grid where each of the 100 cells represented a unique hypothesis. While this simplifies the complex, multidimensional hypothesis spaces that real-world reasoning likely involves, it resembles how they have been operationalized in similar studies examining how people acquire knowledge through search (e.g., spatially correlated multi-armed bandit tasks; see Giron et al., 2023; Witt et al., 2024) and allows us to modify theoretically interesting features of the space while maintaining reasonable tractability.

The likelihoods associated with each hypothesis were generated using a Gaussian process with a squared exponential kernel defined by two key parameters, as in Figure 2 (rows 8 and 9 of Table 1 illustrate how these parameters map onto real-world hypotheses). The first, spatial correlation, controls the smoothness of the hypothesis space: how similar nearby hypotheses are to each other in likelihood. The second, variance, governs the relative difference in likelihoods across the space: how much better good hypotheses are than bad hypotheses. The likelihood of each hypothesis was normalised such that the sum of all likelihoods was one.

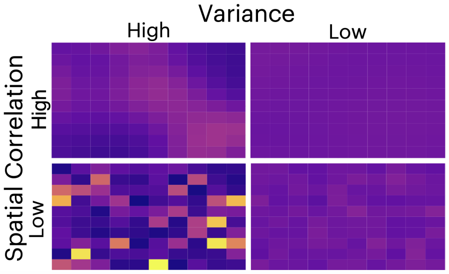

Figure 2: **Example hypothesis spaces** demonstrating how variance and spatial correlation affect the distribution of likelihoods in the hypothesis space. Spatial correlation (rows) captures the smoothness of correlations between nearby hypotheses, while variance (columns) captures how different the best hypotheses are from the worst.

Search algorithm. Consistent with previous research (e.g., Ullman et al., 2012), agent search behaviour was modelled using Metropolis-Hastings Markov Chain Monte Carlo. This algorithm was chosen for its flexibility in capturing the features described in Table 1, its adaptability to various discrete search scenarios, and its straightforward implementation.

Starting from a randomly selected initial hypothesis, the MH-MCMC algorithm proposes new hypotheses by sampling a position within 1 to 5 cells in any direction from the current one. At each step, the algorithm accepts proposed hypotheses with higher likelihoods unconditionally, while those with lower likelihoods are accepted probabilistically based on the difference in likelihood between the proposed and current hypothesis. This rejection rule means that agents are less likely to accept a lower-quality hypothesis as the difference in likelihood between it and the current hypothesis increases. This stochastic acceptance mechanism captures the noisy evaluation of hypotheses observed in real-world reasoning, where individuals cannot perfectly determine the quality of hypotheses. The process is repeated for a fixed number of iterations, simulating each agent's search over time.

For ease of interpretation, we assume that the currently accepted hypothesis at any iteration represents the agent's belief about the true hypothesis at that time, even if that agent had previously accepted hypotheses with higher likelihoods. The movement of the algorithm from one hypothesis to another thus captures the agent updating their beliefs about the true hypothesis. Rejecting a proposed hypothesis corresponds to the agent considering (but ultimately disbelieving) an alternative hypothesis. We also assume that each agent's testimony reflects the likelihood of the hypothesis they have accepted.

Simulation Procedure We conducted 2,000 simulations, each with a unique hypothesis space which was generated based on parameters controlling spatial correlation and variance. The values of these parameters were selected using Latin hypercube sampling with fixed ranges (both between 0.1 & 3).[1]) Each landscape was searched by 20 "agents" (chains in the search algorithm) for 100 iterations. This resulted in 4,000 simulated agents and 4,000,000 observations. This procedure therefore allowed us to see simulate how the average likelihood of an agent's testimony varied as a function of the hypothesis space (knowability), how long the agent has searched for (expertise), and how many other agents agreed (consensus).

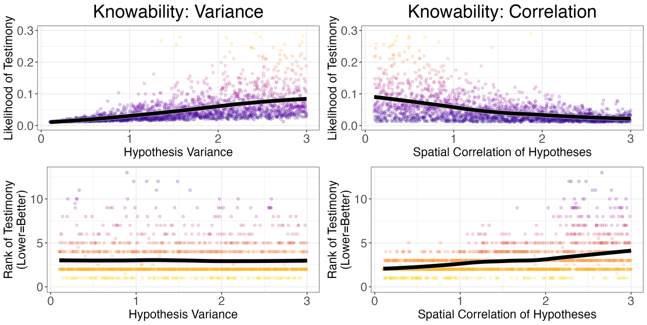

Figure 3: **Role of hypothesis space in agent testimony quality**. We ran 2000 simulations, each with a different hypothesis landscape; each dot shows the median likelihood of all agents in a given landscape, collapsed across every level of expertise (iteration) . The colour of the dots reflects the quality of that agent's testimony (the likelihood of its final hypothesis at the end of 100 iterations). On the *y* axis, the top row shows the median testimony quality and the bottom row shows the rank of the testimony (a rank of 1 indicates that the final hypothesis of that agent has the highest likelihood in the space, so lower rank is better). As the variance of hypotheses in a landscape increases and the correlation of hypotheses decreases (*x* axis), the median quality of testimony increases (top). However, only spatial correlation affects the ability to uncover the best hypothesis (better performance with lower spatial correlation (bottom).

Results & Discussion

Previous research indicates that social testimony may be more persuasive for topics that are more inherently knowable, in which the ground truth is easier to uncover (Alister et al., 2025; Richardson & Keil, 2022; Yousif et al., 2019). However, it is still unclear precisely what knowability means in the context of social reasoning, as well as *why* knowability should matter. Our approach suggests two possible dimensions of knowability as it relates to the structure of hypothesis spaces (see also Alexander et al., 2015). First, it could pertain to the variance of hypotheses in a space, with higher variance indicating a greater difference in quality (likelihood) between the best and worst hypotheses. Second, knowability could also pertain to the extent to which good hypotheses cluster together (i.e., their spatial correlation). The top row of plots in Figure 3 suggest that both variance (left column) and spatial correlation (right column) affect the likelihood of the final hypothesis of each agent (i.e., their testimony, *y* axis), collapsing across how long each agent has searched for. The effects are in opposite directions: as variance increases, the testimony has higher likelihood; but as spacial correlation increases, the testimony has lower likelihood.

Is testimony more convincing in knowable contexts because it's *easier to find* the best hypotheses, or simply because the best hypotheses have a higher average likelihood? We can examine this by analysing the ordinal quality of hypotheses; ranking each hypothesis so that those with higher likelihood have lower rank yields a measure of relative hy-

[1] These ranges were chosen because they produced reasonable variation in the distribution of hypothesis quality across simulations.

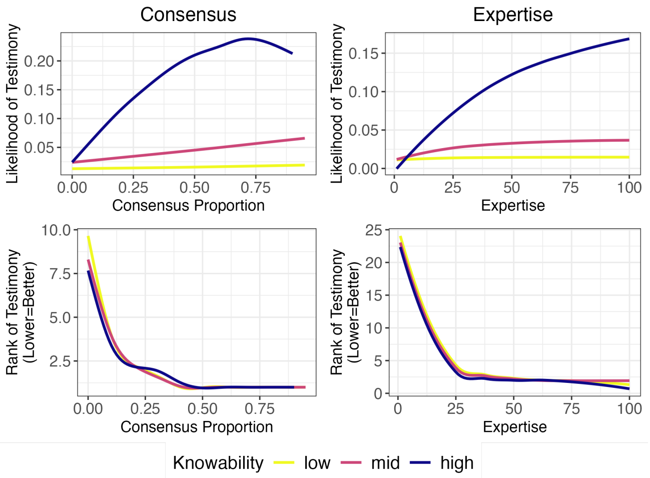

Figure 4: **Role of expertise, consensus, and knowability on agent testimony quality**. On the *y* axis, the top row shows the median testimony quality and the bottom row shows the rank of the testimony (a rank of 1 means that the final hypothesis of that agent has the highest likelihood in the space; lower rank is better). Expertise is defined by the number of iterations they have searched (*x* axis) and consensus as what percent of agents agreed on the best hypothesis.

pothesis quality (rank 1 is best, rank 100 is worst). As shown in the bottom half of Figure 3, agents are more likely to uncover the best hypotheses when spatial correlation is lower. However, hypothesis variance does not appear to matter.

Intuitively, one might think that having *higher* spatial correlation would increase the probability of finding better hypotheses, since it would provide a gradient that the search algorithm can follow. Our simulations suggest that this intuition may not be the case, at least when search processes resemble those of our simulated agents and the hypothesis spaces have the character of ours. One reason we observed this is that higher spatial correlation reduces the difference in quality between near hypotheses, making the algorithm more likely to accept new hypotheses that were worse than the first. This suggests that high spatial correlation might be more beneficial for agents who can only take small steps away from their current position or who are less willing to accept lower-quality hypotheses. It might also have different effects in a much larger or more structured hypothesis space – a possibility we consider in more detail in the discussion.

What about the features of the agents themselves? We explored this in our simulation by analysing two possible features. First, we consider the expertise of the agent, defined as the number of iterations they searched the hypothesis space for. Second, we evaluate the consensus of testimony, defined as the proportion of other agents within the same hypothesis space that converged to the same hypothesis on any given iteration. Perhaps unsurprisingly, our simulations indicate that testimony is higher quality when agents have more expertise and there is more consensus between different agents (Figure 4). While these are not novel insights, it is reassuring that our approach yields findings that are consistent with existing research (Alister et al., 2022; Connor Desai et al., 2024; Schulz et al., 2023; Simmonds et al., 2024). Moreover, while the predictions themselves may not be surprising, our framework offers a novel explanation for *why* these factors should matter: expertise is important because it reflects a longer search of the hypothesis space, and consensus is important because different agents are more likely to have converged on the same hypothesis when the likelihood of that hypothesis is high.

Our simulations also allow us to predict how these factors trade off against each other, which would be difficult without our framework. How is the value of expertise and consensus affected by knowability? We addressed this by conducting simulations in which we collapsed our two dimensions of knowability into a single scale, where lower correlation and higher variance corresponded to higher knowability. As the top half of figure 4 shows, the effect of both expertise and consensus depends substantially on knowability; both matter most when the topic is more knowable. These findings are consistent with Alister et al. (2025), who found that people are less convinced by a consensus when the topic is less knowable. Our framework suggests that this may happen because in less knowable contexts, a consensus is not as strong an indication of high quality hypotheses. Given that informants in Alister et al. (2025) were always experts, our simulations also suggest that their results may have been less pronounced if the informants had lower expertise: a prediction to be tested in future work. The bottom half of Figure 4 suggests that knowability does not matter when considering only the relative rank of hypotheses, rather than their raw likelihood. These results suggest that social testimony may be less persuasive in less knowable environments because all hypotheses are weighted similarly, so even if an agent has technically uncovered a *better* hypothesis than the learner (in terms of rank), it is unlikely to be very different in quality to what the learner already believes.

General Discussion

In this paper, we suggest that social reasoning can be conceptualised as reasoning about how other agents acquired their knowledge. Consistent with the idea that knowledge acquisition can be described as a stochastic search through a hypothesis space, we propose that modelling the search processes of other agents using search algorithms yields insight into the factors that affect the quality of their testimony.

As a proof of concept, we conducted a simulation study to demonstrate how this approach generates predictions about the role of: 1) features of agents that indicate their credibility; 2) the presence of agreement between multiple agents (consensus); and 3) the knowability of the topic. Although existing models capture some of these factors (e.g., Madsen et al., 2020; Oktar et al., 2024), our framework is, to our knowledge, the first to incorporate all three.

In future research we plan to extend these basic simulations to more sophisticated situations. For example, we modelled informants as completely unbiased and rational in the sense that they always accepted better hypotheses and evaluated hypotheses solely based on their objective likelihood. In reality, however, people have biases that influence which hypotheses they consider or accept. For instance, if the hypothesis space were conceptualised as representing different political dimensions, someone on the far left might be less likely to accept a hypothesis aligned with the far right, even if it had a high likelihood (and vice versa). This kind of bias could be integrated into stochastic search algorithms by incorporating a prior distribution over the hypothesis space.

Another important consideration is that we modelled agent's acceptance rate as constant over time. However, this does not align with how human knowledge acquisition is typically described in developmental research: children often have a tendency to explore poorer hypotheses rather than exploit known good ones, a tendency that decreases with age (Bonawitz et al., 2014; Gopnik et al., 2017). This exploration is thought to be evolutionarily adaptive, as being too biased towards good hypotheses could result in being stuck in a local maximum, thus preventing the discovery of even better ones. Notably, this behaviour parallels simulated annealing algorithms (Kirkpatrick et al., 1983), which use a temperature parameter that controls the balance between exploration and exploitation and has been used to model developmental changes in search behaviour (Giron et al., 2023).

Another way that our simulations were unrealistic is that they assumed informants do not have any memory or capacity to form a representation of the hypothesis space. Future research could address this limitation by integrating search algorithms with more advanced approaches like Gaussian process models (e.g., Giron et al., 2023; Witt et al., 2024). These models form a structured representation of the hypothesis space based on past samples, and would enable us to better capture the complexity of human reasoning.

Most stochastic search algorithms, including those in our simulations, assume that each agent searches the hypothesis space independently. Of course, in the real world, people often communicate with one another, which can have different effects depending on the nature of the coordination. For instance, the quality of the final hypotheses might increase if agents coordinate to search different areas of the space or identify parts of the space that are unlikely to pay off. Conversely, quality might decrease if agents systematically ignore some parts of the space because they think others have already searched it or their ingroup tells them to do so. People do pay attention to the independence of sources when updating beliefs, but exactly *how* and *when* is a complicated story (Richardson & Keil, 2022). Incorporating communication into our framework is one way to better understand the complex mix of factors that shape when communication is useful and when it is not. Such communication mechanisms already exist in some search algorithms, such as in particle swarm optimisation (Kennedy & Eberhart, 1995).

In our simulations, we demonstrated how the quality of agent testimony was affected by two features of a small, 2D hypothesis space: spatial correlation and hypothesis variance. Of course, hypothesis spaces have many more features that might affect how easy it is to uncover good hypotheses. For example, larger hypothesis spaces with a long distance between local maxima – or more structured hypothesis spaces where there are long sequences or clumps of high quality hypotheses – might result in qualitatively different patterns (e.g., agents being more likely to get stuck in local maxima; multiple agents converging on local maxima; or agents splitting between multiple distant local maxima, as in belief polarisation). Here, we have only considered discrete hypothesis spaces, but the same principles could still apply in a continuous space; for instance, where hypotheses are represented as vectors in a high dimensional space (Piantadosi et al., 2024).

Another avenue to consider is contexts where the hypothesis space is *dynamic*. What if the distribution of hypotheses in a space change over time, such that hypotheses that previously had high likelihood no longer do? Or what if new maxima emerge in parts of the space that previously contained only low-likelihood hypotheses? Dynamic hypothesis spaces may reflect the fact that utility of different hypotheses change over time as the environment changes or the context evolves. If a reasoner assumes that a topic corresponds to a dynamic hypothesis space, we predict that the reasoner should have less faith in experts who have spent a long time searching the space: after all, areas that the expert searched and rejected previously may now have high value.

A key assumption of our framework is that the agent's testimony is always what they believe to be true (and the reasoner knows this). It therefore does not currently account for situations where a reasoner believes an agent is deliberately misleading or outright lying. This sort of reasoning is an important factor in the real world, and has been the focus of modelling work (e.g., Alister et al., 2023; Goodman & Frank, 2016; Shafto et al., 2014). While in principle the framework could be extended in this way, it fills a vital gap even without this extension; honesty alone is not sufficient to persuade someone unless there is some reason to believe that the agent actually knows the truth in the first place.

Our framework is normative as it offers an explanation for what kinds of informants *should* be more persuasive given features that indicate the quality of their knowledge acquisition (search through a hypothesis space). Although we qualitatively replicate and explain some existing findings, more work needs to be done to better understand how well this framework applies to real human behaviour. Regardless, we suggest that it provides a useful new way of formalising constructs of social reasoning like expertise, bias, or knowability. In doing so, we not only are able to more precisely define what these concepts might mean but also explore more fully how they trade off against each other and generate falsifiable new predictions. This includes features in Table 1 as well as others yet to be explored. Future experimental work will be useful for identifying where human social reasoning diverges from our model predictions and in such cases, could also motivate interventions to promote greater belief in sources who are more likely to know the truth.

Acknowledgments

We would like to thank the following people who helped consolidate our ideas, including (and in no particular order): Keith Ransom, Rahul Bhui, Josh Tenenbaum, David Lagnado, Bill Thompson, Frank Mollica, & Saoirse Connor-Desai. Manikya Alister was supported by an Australian Government Research Training Scholarship.

References

Alexander, J. M., Himmelreich, J., & Thompson, C. (2015). Epistemic Landscapes, Optimal Search, and the Division of Cognitive Labor. *Philosophy of Science*, *82*(3), 424–453. https://doi.org/10.1086/681766

Alister, M., Ransom, K. J., Connor Desai, S., Soh, E. V., Hayes, B. K., & Perfors, A. (2025). How convincing is a crowd? Quantifying the persuasiveness of a consensus for different individuals and types of claims. *Psychological Science*.

Alister, M., Ransom, K. J., & Perfors, A. (2022). Source independence affects argument persuasiveness when the relevance is clear. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). Retrieved February 1, 2023, from https://escholarship.org/uc/item/5hg4p8cm

Alister, M., Ransom, K. J., & Perfors, A. (2023). Inferring the truth from deception: What can people learn from helpful and unhelpful information providers? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). Retrieved November 2, 2023, from https://escholarship.org/uc/item/8vt661bv

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497–500. https://doi.org/10.1016/j.tics.2014.06.006

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338. https://doi.org/10.1037/rev0000061

Connor Desai, S., Lee, J., & Hayes, B. (2024, October). Explaining Away the Illusion of Consensus. https://doi.org/10.31234/osf.io/9hnm7

Giron, A. P., Ciranka, S., Schulz, E., van den Bos, W., Ruggeri, A., Meder, B., & Wu, C. M. (2023). Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour*, *7*(11), 1955–1967. https://doi.org/10.1038/s41562-023-01662-1

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899. https://doi.org/10.1073/pnas.1700811114

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cognitive Science*, *40*(6), 1496–1533. https://doi.org/10.1111/cogs.12276

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of ICNN'95 - International Conference on Neural Networks*, *4*, 1942–1948 vol.4. https://doi.org/10.1109/ICNN.1995.488968

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science*, *220*(4598), 671–680. Retrieved September 30, 2024, from https://www.jstor.org/stable/1690046

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2020). The impact of partial source dependence on belief and reliability revision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(9), 1795–1805. https://doi.org/10.1037/xlm0000846

Mercier, H., & Morin, O. (2019). Majority rules: How good are we at aggregating convergent opinions? *Evolutionary Human Sciences*, *1*, e6. https://doi.org/10.1017/ehs.2019.6

Montgomery, L. E., Bradford, N., & Lee, M. D. (2024). The wisdom of the crowd with partial rankings: A Bayesian approach implementing the Thurstone model in JAGS. *Behavior Research Methods*, *56*(7), 8091–8104. https://doi.org/10.3758/s13428-024-02479-0

Oktar, K., & Lombrozo, T. (2025). How aggregated opinions shape beliefs. *Nature Reviews Psychology*, 1–15. https://doi.org/10.1038/s44159-024-00398-7

Oktar, K., Lombrozo, T., & Griffiths, T. L. (2024). Learning From Aggregated Opinion. *Psychological Science*, 09567976241251741. https://doi.org/10.1177/09567976241251741

Orchinik, R., Dubey, R., Gershman, S. J., Powell, D., & Bhui, R. (2024). Learning from and about climate scientists. *PNAS Nexus*. https://doi.org/10.31234/osf.io/ezua5

Piantadosi, S. T., Muller, D. C. Y., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, *28*(9), 844–856. https://doi.org/10.1016/j.tics.2024.06.011

Pilditch, T. D., Hahn, U., Fenton, N., & Lagnado, D. (2020). Dependencies in evidential reports: The case for informational advantages. *Cognition*, *204*, 104343. https://doi.org/10.1016/j.cognition.2020.104343

Readfearn, G. (2022). 'Word salad of nonsense': Scientists denounce Jordan Peterson's comments on climate models. *The Guardian*. Retrieved February 16, 2024, from https://www.theguardian.com/environment/2022/jan/27/word-salad-of-nonsense-scientists-denounce-jordan-petersons-comments-on-climate-models

Richardson, E., & Keil, F. C. (2022). The potential for effective reasoning guides children's preference for small group discussion over crowdsourcing. *Scientific Reports*, *12*(1), 1193. https://doi.org/10.1038/s41598-021-04680-z

Romeijn, J.-W., & Atkinson, D. (2011). Learning juror competence: A generalized Condorcet Jury Theorem. *Politics, Philosophy & Economics*, *10*(3), 237–262. https://doi.org/10.1177/1470594X10372317

Schulz, L., Schulz, E., Bhui, R., & Dayan, P. (2023, October). Mechanisms of Mistrust: A Bayesian Account of Misinformation Learning. https://doi.org/10.31234/osf.io/8egxh

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. https://doi.org/10.1016/j.cogpsych.2013.12.004

Simmonds, B. P., Ransom, K. J., & Stephens, R. (2024). Navigating Health Claims on Social Media: Reasoning from Consensus Quantity and Expertise. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0). Retrieved October 28, 2024, from https://escholarship.org/uc/item/76j8m25k

Simmonds, B. P., Stephens, R., Searston, R. A., Asad, N., & Ransom, K. J. (2023). The Influence of Cues to Consensus Quantity and Quality on Belief in Health Claims. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). Retrieved January 10, 2024, from https://escholarship.org/uc/item/73t0j4tc

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480. https://doi.org/10.1016/j.cogdev.2012.07.005

Whalen, A., Griffiths, T. L., & Buchsbaum, D. (2018). Sensitivity to Shared Information in Social Learning. *Cognitive Science*, *42*(1), 168–187. https://doi.org/10.1111/cogs.12485

Witt, A., Toyokawa, W., Lala, K. N., Gaissmaier, W., & Wu, C. M. (2024). Humans flexibly integrate social information despite interindividual differences in reward. *Proceedings of the National Academy of Sciences*, *121*(39), e2404928121. https://doi.org/10.1073/pnas.2404928121

Yousif, S. R., Aboody, R., & Keil, F. C. (2019). The Illusion of Consensus: A Failure to Distinguish Between True and False Consensus. *Psychological Science*, *30*(8), 1195–1204. https://doi.org/10.1177/0956797619856844