Track Records: A Cautionary Tale

Alice C.W. Huang

Abstract

In the literature on expert trust, it is often assumed that track records are the gold standard for evaluating expertise, and the difficulty of expert identification arises from either the lack of access to track records, or the inability to assess them (Goldman, 2001; Nguyen, 2020; Schurz, 2012). I show, using a computational model, that even in an idealized environment where agents have a God's eye view on track records, they may fail to identify experts. Under plausible conditions, selecting testimony based on track records ends up reducing overall accuracy, and preventing the community from identifying the real experts.

Keywords: testimony, network epistemology, expertise, agent-based models

1 Introduction

One of the most important day-to-day epistemic problems we face is deciding who to trust. Most of my beliefs come from the testimony of others. For instance, I believe that the earth is round, even though the terrain I walk on is flat. I believe that quarks exist, without fully understanding the reason physicists believe they exist. Within expert communities, scientists also rely regularly on data and results established by their peers. Given that most of our beliefs are formed fully or partially based on testimony, and potential informants vary in reliability, it is important to decide who to solicit testimony from.

In the literature on testimony and trust in experts, it is sometimes suggested that challenges in identifying experts stem from either the unavailability of putative experts' past beliefs, or the inability to assess those past beliefs. In this paper, I show, however, that even if we have full access to correctly assessed track records, using them as guidance to seek testimony can be problematic under plausible conditions.

I make my case using computational models of scientific inquiry, where agents update their beliefs on both data and testimony. The main result compares two models. In the baseline model, agents identify experts based on track

records, without updating their own beliefs on expert opinion. In the testimony model, agents not only identify experts, but also update their own beliefs partially in light of the beliefs of those experts. I measure overall accuracy as well as "meta-expertise," the ability to recognize real experts. The testimony model does worse on both metrics. This result can be explained by premature convergence on the opinions of a small set of individuals. The phenomenon uncovered in this paper bears some similarity to the Zollman effect (Zollman, 2007), but is different in many important aspects.

There are different notions of meta-expertise, depending on who manifests it. Individual meta-expertise involves a person's ability to identify experts, whereas group meta-expertise concerns whether recognition within a community tracks expertise. We can also make the distinction between insiders' and outsiders' meta-expertise. For instance, one might think that a scientist is more likely to manifest individual meta-expertise than a layperson, or that a community of scientists is better at manifesting group meta-expertise than a community of laypeople.

The notion of meta-expertise I am primarily interested in is the group meta-expertise of the scientific community. In other words, I am interested in whether the most prestigious scientists tend to hold more accurate beliefs. This question is of course pertinent to outsiders. If the scientific community can confer recognition on scientists with accurate beliefs, then laypeople can identify experts easily, without having expertise themselves—they can simply trust the testimony of an acclaimed scientist. But my focus will be on the first question: the meta-expertise of the scientific community.

The model I develop may be applied to problems such as estimating the chance of a natural disaster, vaccine safety, stock market forecasts and so on.[1] And the community of experts I model is the scientific community broadly construed, as opposed to, say, art experts or film critics.[2]

There are various reasons why track records can be bad indicators of reliability. For instance, a top scientist may only be interested in extremely difficult questions, making their track record worse despite superior abilities. An expert might also choose to share beliefs that are more controversial rather than beliefs that have been established beyond reasonable doubt, since the latter receive less interest in the public sphere. This kind of systematic preferential sharing could skew a layperson's judgment of a putative expert's track record. But these are not the kinds of cases I am interested in here. In my model,

[1] I focus mainly on modelling specialized domains in which some level of cognitive ability, such as specific training or experience, is required to be potentially considered an expert. I am not concerned with the broader notion of expertise, where I am an expert of my private perspective, nor questions such as whether I should believe my absentminded roommate when she claims to have locked the door before leaving the apartment. However, the model can possibly be generalized to less specialized domains. Moreover, I focus on questions whose answers may be represented as a true proposition, setting aside practical questions, such as how to swim efficiently, without assuming that there is always a clear distinction between cognitive and practical domains.

[2] I assume that the main scientific methods include data collection, inferences based on collected data, and information exchange with other scientists. Other techniques of scientific inquiry are idealized away.

there is no preferential sharing, and all agents, regardless of expertise, solve the same problem.

What I want to show is that even in the highly idealized scenario where we assume a God's eye view on track records—the accuracy of past predictions are all publicly available—track records may fail to provide guidance on who to trust. Therefore, the problem of selecting expert testimony goes beyond the lack of access to, or assessment of, track records.

In the next section, I lay down the groundwork and review some of the literature on expertise and testimony. In section 3, I develop the model I use, and explain how meta-expertise and accuracy are operationalized. In section 4, I present the results. Section 5 compares this model with some previous models, and section 6 concludes.

2 Expertise and Testimony

At the beginning of the Covid-19 pandemic, some putative experts said that the use of face masks can slow the spread of the virus, whereas others said that the coronavirus is too small for face masks to be effective. With little knowledge in epidemiology, how should I adjudicate between contradicting advice from putative experts? In general, how can a layperson distinguish credible sources of information from unreliable ones in a specialized domain? This is what Goldman (2001) calls the "novice/2-expert problem."

One natural thought is that we can look at the track records of these putative experts. To decide which expert advice to take, I should discern which potential expert has a history of making true claims.[3]

There are two main difficulties with assessing expertise by track records, however. First, laypeople have little access to the cognitive history of putative experts. Although academic journals, popular articles or personal communication platforms such as Twitter sometimes provide some access, only a fraction of a scientist's beliefs makes their way into publication. And even then, it is rare that a layperson would research a putative expert's cognitive history.

Second, even with access to a putative expert's cognitive history, the layperson is not in a position to evaluate these past beliefs, due to exactly the lack of expertise that makes them a layperson. For example, even if a chemist's past beliefs were presented to me, I would not possess the necessary knowledge to evaluate their accuracy.

On the other hand, experts themselves are in a better position to both access and assess the track records of their peers. The way for laypeople to get around this problem, then, is to let the putative experts assess one another. A layperson can simply follow the advice of someone who is recognized by the scientific community. As Coady (2012) points out, "although expertise and meta-expertise are logically distinguishable, they overlap to a large extent. Because experts typically work closely with other experts on the same subject, we can usually assume that experts will be able to recognize other experts."

[3] See, e.g., Goldman (2001) and Nguyen (2020).

Many systems of expert ranking proceed under this assumption. For instance, *Lexpert* ranks lawyers based on who are most frequently recommended by other lawyers. The Philosophical Gourmet Report, which ranks graduate programs, also relies on philosophers' mutual assessment.

In practice, it is easier to know whether an individual is recognized by the scientific community than it is to evaluate their track record. Goldman (2001) suggests, for example, that "academic degrees, professional accreditation, work experience and so forth (all from specific institutions with distinct reputations) reflect certifications by other experts of...demonstrated training or competence." Information about institutional affiliation, appraisals and certification is more easily available and comprehensible to laypeople than track records in a specialized domain. If recognition reliably indicates accuracy, then the novice can identify real experts without the relevant expertise.

The implication of the results in this paper is, therefore, that even with fully accessible track records, a group of putative experts can fail to manifest meta-expertise. In turn, an individual relying on the mutual assessment of the expert community will also fail to manifest meta-expertise.

3 The Model

In this section, I develop the model used in the simulation.

3.1 The Mechanism

The goal of the agents in the model is to estimate the bias of a coin. They update their beliefs based on the coin tosses they observe, and they also solicit testimony from each other.

The coin has some fixed probability p of landing heads each time. The quantity p represents a proportion that scientists are trying to estimate. Agents use the outcomes of the coin flips to infer the value of p. The tosses model evidence, gathered from randomized controlled trials, for example. Real scientific investigations are more complex, of course. Here I only aim to track a few features of scientific inquiry to show that, even in a very simple idealized model where we would expect that identifying experts is straightforward, selecting testimony is non-trivial.

The hypothesis space consists of 6 mutually exclusive and jointly exhaustive hypotheses about the coin bias, ranging from 0 to 1, with increments of 0.2. At each time step, each agent tosses the coin once.

Agents also receive testimony from their informants—the subset of scientists whose opinions they pool in their updating. Scientists choose agents who have the best track records as their informants.

Testimony comes in the form of conclusions, as opposed to evidence. In other words, agents hear the credences of their informants, but not the coin tosses observed. They cannot infer the coin tosses that their neighbors have observed, since they do not know how others combine data and testimony.

3.2 Cognitive Agents

Agents in the model are cognitively diverse. Cognitive variation is important here: without it, the task of assessing others would be trivially equivalent to guessing who is luckier with the data they observe. Cognitive strategies are determined by three parameters—the evidential component, the social component, and how the two are combined.[4] These parameters capture only a fraction of the factors contributing to diversity in scientific processes. The best scientists are often also good at generating new hypotheses, building novel causal models, connecting the dots, and so on. But we abstract away from these aspects.

At each time step, an agent's credence is given by the linear combination of an evidential and a social component, each being a probability function. The weight given to data versus testimony is the parameter c, a random real number drawn from $[0, 1]$.[5] If a scientist has $c = 1$, she works in complete isolation, merely with her own evidence and completely unaffected by the opinion of other scientists. If a scientist has $c = 0$, she relies entirely on the research of other agents to form her credences, without directly engaging with any data. The combination of the social and evidential components proceeds as follows:[6]

$$P_t(H_j) = (1 - c)\ P_t^s(H_j) + c\ P_t^e(H_j). \tag{1}$$

The superscripts e and s stand for *evidential* and *social*. H_j is the hypothesis that the bias of the coin is $j \in [0, 0.2, ..., 1]$.

Let P_{t-1} be the agent's credence function at the previous time step, and P_t^e be their evidential component at the current time step, after updating on evidence E (the outcome of the coin toss at t). An ideal agent (*) updates following standard Bayesian conditionalization: $P_t^{e*}(H_j) = P_{t-1}(H_j \mid E)$. However, agents in the model are not ideal. Each scientist has a *noise* parameter b, randomly drawn from the interval [0, 0.2]. Instead of the posterior that an ideal Bayesian agent would have after the update, each agent's posterior $P_t^e(H_j)$ for each hypothesis is randomly drawn from a normal distribution with mean $P_t^{e*}(H_j)$ and standard deviation b, with $P_t^{e*}(H_j)$ being the posterior of an ideal Bayesian.[7] The entire distribution is then normalized so that the probabilities add up to 1.

The noise level b determines how much an agent deviates from the ideal Bayesian. The same level of noise persists across every round of updating. An

[4] Mohseni and Williams (2021) develop a model that also includes an evidential and a social component, but the social component involves the pressure to conform, rather than track-record-based testimony.

[5] One might take testimonial updates to be a kind of evidential update, in which case, the distinction is rather between direct and indirect evidential updates. I am neutral on the nature of testimony here.

[6] I owe my thanks to Hegselmann and Krause (2006) and Douven and Riegler (2009) for inspiration behind this model.

[7] The normal distribution is truncated to avoid negative values.

agent who updates with a lot of noise has less chance of forming accurate credences when working alone.[8]

The social component, on the other hand, is determined by the openmindedness parameter m, which specifies what percentage of the community the agent is willing to solicit testimony from. For instance, if I have $m = 0.2$ and there are 30 agents in total, then I will solicit testimony from the top $30 \cdot 0.2 = 6$ agents, ranked by track record. The parameter m is randomly drawn from the interval $[0.05, 0.5]$. In other words, the most openminded agents are willing to solicit testimony from anyone with an above-median track record, whereas the most demanding agents only trust the agents with the best track records.

After the set of informants are identified, the social component of the update is given by the average of the their credences. Formally, for each hypothesis H_j,

$$P_t^s(H_j) = \frac{\sum_{i=1}^{k} P_{i,t-1}(H_j)}{k}. \tag{2}$$

$P_{i,t-1}$ is the credence function of the i^{th} informant at time step $t-1$, and k is the number of informants.[9]

3.3 Output Evaluation

I now explain how I operationalize two important notions: accuracy of credences, and recognition by the scientific community.

I adopt a broadly veritistic approach in this paper. Epistemic agents in my model have the aim to form true beliefs, and their success is evaluated solely based on accuracy. The accuracy of credences is measured by the Brier score (Brier, 1950). It is used both for measuring the accuracy of each agent's credences about the coin's bias, and for assessing track records.

The Brier score is defined as

$$1 - \frac{1}{n}\sum_{j=1}^{n}(P(H_j) - I(H_j))^2, \tag{3}$$

where n is the total number of hypotheses, and H_j is the j^{th} hypothesis. $I(H_j) = 1$ if H_j is true, and $I(H_j) = 0$ otherwise.

For instance, suppose the coin bias $p = 0.8$ and an agent's credence distribution is as follows: $P(p = 0.4) = 0.2, P(p = 0.6) = 0.3, P(p = 0.8) = 0.5$, and

[8] The priors are either uniform or randomly generated. In the former case, scientists have no preconception about the scientific problem, whereas in the latter, scientists have various prior biases about the topic before gathering any data. The priors made no difference to the results presented in section 4.

[9] Arithmetic average may not be an ideally rational method of aggregating beliefs, and some have argued for different alternatives (Babic, Gaba, Tsetlin, & Winkler, 2021; Dietrich & List, 2016). We may also obtain the social component using weighted arithmetic mean, where the weights are informed by track records. Preliminary tests suggest no qualitative difference between the current approach and weighted arithmetic average. These other methods of aggregation are left for future work to explore.

credence 0 for the three other hypotheses. The Brier score for the agent's credences will be $1 - \frac{(0.2-0)^2+(0.3-0)^2+(0.5-1)^2}{6}$.[10] When measuring the accuracy of an agent's beliefs about the coin's bias, the Brier score will be 1 when the agent has credence 1 for the correct hypothesis, and 0 for all other incorrect hypotheses. Agents do not know the accuracy of anyone's credences, including themselves, since they do not know what the coin bias p is.

Track records, on the other hand, are available to all agents. After each coin toss, the agent is assessed by their prediction, based on their credences from the previous time step. For example, suppose at time t, I have credences $P(p = 0.2) = 0.5$ and $P(p = 0.4) = 0.5$. My expectation will be $0.5(0.2) + 0.5(0.4) = 0.3$. In other words, my prediction for the next coin flip is 30% heads and 70% tails. If at $t + 1$, the coin toss is heads, then my Brier score is $1 - (0.3 - 1)^2 = 0.51$, and if the coin toss is tails, then my Brier score is $1 - (0.3 - 0)^2 = 0.91$. The mean of the prediction Brier scores up until each time step is my track record at that time step. Track records are constantly updated, and fully transparent to all.

Coin toss predictions are a simple way to model track records. Given the tight connection between an agent's credences about the coin bias, and the accuracy of their predictions about coin tosses, the latter is as good as track records get, if our goal is to evaluate the former. The simplicity makes the conclusions stronger—even with such a simple understanding of track records in the model, problems can still arise.[11]

As for recognition, drawing from metrics in network analysis, I assume that the cluster of concepts such as prestige, reputation and recognition can be understood in terms of an agent's status in a network of epistemic trust relations. I understand a reputable individual in the scientific community to be someone deemed reliable, and whose work informs other scientists. In other words, a prestigious scientist is someone whose testimony is sought after. My operationalization of the concept of recognition is in line with Golub and Jackson (2010), Katz (1953) and Bonacich (1987)'s understanding of social status and power.

In network theory, centrality measures how important or influential a node is within a network, based on its position relative to other nodes. Different centrality metrics assess different aspects of a node's importance in a network,

[10] One problem with the Brier score is that it does not account for the distance to truth, and rewards flatter beliefs over false hypotheses. For instance, suppose the true coin bias is 0.4. If both Johnny and Mina have credence 0.7 that the coin bias is 0.4, but Johnny has credence 0.3 that the coin bias is 0.9 whereas Mina has credence 0.3 that the coin bias is 0.5, then presumably, we want to say that Mina has more accurate beliefs than Johnny, since her false belief is closer to the truth than that of Johnny. The Brier penalty is unable to distinguish between these two cases. To address this concern, I used an additional distance-sensitive scoring rule, the continuous ranked probability score (CRPS), to ensure that the results are robust. Within the parameter settings of this simulation, the Brier scores highly correlate with CRPS.

[11] This way of modelling track records does not capture all the different types of data that we sometimes use as track records. For instance, we might think that past success in manufacturing electric cars is evidence for future success in developing electric helicopters. A financial analyst's performance in predicting market movements in the energy sector might be indicative of her performance in predicting trends in the renewable energy market. Offering a general account of what track records consist of is complicated, and a broader construal of track records comes with its own problems.

and the appropriate metric depends on the context of application. In general, a central node tends to be well-connected, or in a position to influence information flow. For example, in a citation network, a paper that is widely cited is often considered central. In diplomacy, a liaison bridging two groups of global powers in conflict is central, even if it is not a global power itself.

There are lots of ways to measure centrality. Using several common metrics of centrality, I find that the results do not depend on which measure we choose. So in the remainder of the paper, the social status of scientists in the community is quantified using the *authority* metric, computed by the Hyperlink-Induced Topic Search (HITS) algorithm (Newman, 2010).[12]

The meta-expertise of a community is measured by how well recognition predicts accuracy. The correlation between centrality and accuracy is evaluated by the R-value. R-values, which fall in the range $[-1, 1]$, are also called correlation coefficients. The sign of the R-value indicates whether the correlation is positive or negative. The absolute value of the correlation coefficient indicates how strong the relation is between the variables—the closer the absolute value is to 1, the stronger the correlation is, whereas an R-value close to 0 indicates that there is little or no relation between the variables.

If recognition is a strong predictor of accuracy, then the community has meta-expertise, and laypeople can simply trust the recognized scientists. If this is not the case, then using prestige as an identifier of expertise is not a viable solution to the novice/2-expert problem.

4 Results

A baseline model is used as the control group. In the baseline model, scientists do not use testimony when forming their beliefs. Each agent forms beliefs solely based on their evidence, but makes judgements about the reliability of others. Meta-expertise, or lack thereof, does not interfere with expertise in the baseline model. Prizing apart conferred recognition and testimony-soliciting relations allows us to understand the effects of social updating. The baseline model is implemented by setting $c = 1$ for all agents.[13]

We should expect that in the baseline model, centrality is a strong predictor of accuracy, since the more accurate an agent's credences are, the better their track record will be in the long run. Since track records are fully transparent to all agents, and each agent confers recognition to the top $m\%$ of agents ranked by track records, the correlation between centrality and accuracy will strengthen with time. This is in fact what we see in the simulations. Fig.1 shows the relation between centrality and accuracy in the baseline model.

[12] Authority is inter-defined with another metric, "hub"—a high hub score is assigned to individuals good at recognizing authorities, while a high authority score is assigned to those recognized by individuals with high hub scores. For our purposes, an agent with a high hub score is someone with meta-expertise, and an agent with a high authority score is someone whose work is deemed reliable by meta-experts.

[13] The parameter $c \in [0, 1]$ determines how much the scientist relies on her own evidence versus peer testimony, and $m \in [0, 1]$ determines how openminded a scientist is. The larger m is, the more the scientist is willing to consult peers with bad track records.

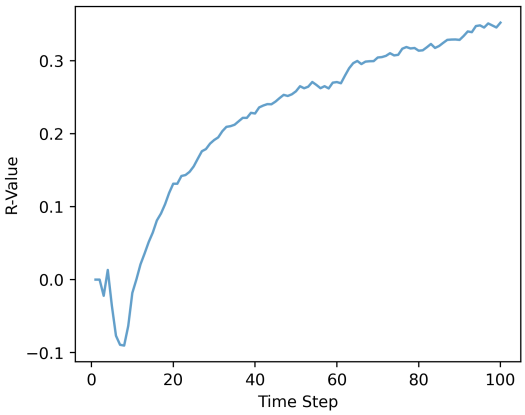

Fig. 1 R-values, averaged over 600 runs, between centrality and Brier accuracy of credences at each time step. Higher R-values indicate more meta-expertise in the scientific community.

Fig. 1 shows how the R-value evolves over 100 time steps, averaged across 600 runs.[14] We can observe that there is initially little correlation between centrality and accuracy, but the correlation quickly emerges as track records begin to track the accuracy of each agent's credences about the coin bias.

Given the idealizations, these results are not surprising. But they are helpful starting points in our investigation, providing a baseline for comparison when we add the effects of testimony to the model.

4.1 The Effects of Testimony

One might expect that, not only knowing the complete track records of others, but also updating one's beliefs in light of that information should increase overall accuracy. Those who have been relatively unsuccessful in predicting the coin tosses would then be able to align their beliefs with those who have been more successful.

It comes as a surprise, then, that overall accuracy is reduced with testimonial updates (fig. 2). Moreover, testimonial updates drastically reduce the community's ability to collectively identify which agents have accurate beliefs, even given the complete availability of objective track records. Fig. 3 is a comparison of the baseline and testimony models in terms of meta-expertise. The correlation between recognition and accuracy is much weaker with testimonial updates.

How can we explain these surprising results? There are two interacting factors to consider. The more obvious one is that track records are only good indicators of the accuracy of credences once we have accumulated enough past predictions. The less obvious factor is that the availability of track records quickly reduces the diversity of opinions within the community; everyone's credences move toward a small subset of highly ranked agents.

[14] I also ran simulations for longer than 100 time steps but found no qualitative difference.

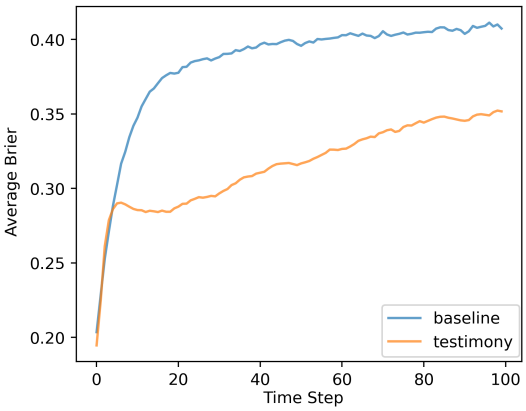

Fig. 2 Comparison of average Brier accuracy at each time step between baseline and testimony models.

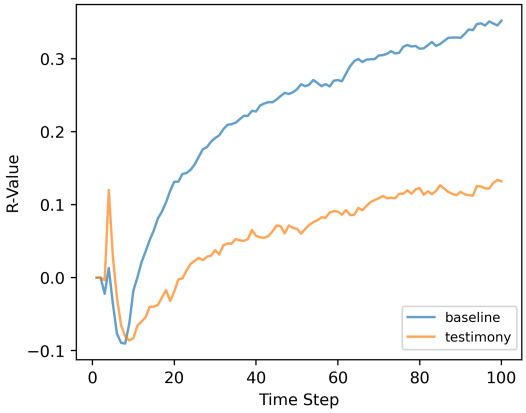

Fig. 3 R-values, averaged over 600 runs, between centrality and Brier accuracy of credences at each time step. Higher R-values indicate more meta-expertise in the scientific community.

Since track records only become reliable indicators of accuracy after a large enough number of predictions, in the initial stages of inquiry, some agents will have good track records by mere luck. And these lucky agents have a lot of influence over the beliefs of others, since everyone knows about their successful predictions, and begins to solicit their testimony. Their inaccuracy quickly spreads across the community.

But what happens once more past predictions are assessed, and track records become more reliable? Shouldn't agents be able to then identify those with accurate credences, solicit testimony from them, and become more and more accurate?

Unfortunately, the negative effects of this kind of unwarranted opinion monopoly are not simply reversed with time.[15] Once monopoly of opinion happens, meta-expertise can hardly improve. Social updating based on the testimony of the most successful agents leads to low variance in credences between agents. And low variance makes it difficult to distinguish the accuracy of agents' beliefs about the coin's bias based on track records of coin toss predictions. For example, suppose the bias is 0.7. The mean Brier score across a series of coin flip predictions will hardly differentiate two agents with credences 0.7 and 0.72.

In turn, it is more difficult for an epistemic group to manifest meta-expertise using information on track records. When there is less variance, meta-expertise also becomes less crucial. Given the marginal difference in accuracy within the community, identifying the agent with the most accurate credences would yield only limited improvement.

Importantly, if the opinion leaders are consistently those with the most accurate credences, then it is not a problem when all agents partially align their credences with those opinion leaders. Opinion monopoly is only problematic when combined with the fact that track records, especially in early stages, do not consistently track the accuracy of beliefs.

To test out this explanation, we can remove these two factors, one at a time, in order to better understand their effects. The next section looks at models where agents either take into account peer testimony only once enough past predictions have been accumulated, or choose their informants in a way that creates less monopoly.

4.2 Less Monopoly, More Patience

The first modified model ensures that agents do not rely on track records to select testimony until enough instances of past predictions have been accumulated. This guarantees a high chance that good track records indicate accuracy. Looking at data from previous trials, after 50 coin tosses, the track records accumulated are generally enough for significant correlation to emerge between track records and the accuracy of beliefs. So in this first modified model, agents wait until the 51^{st} time step to begin soliciting testimony based on track records. Call this the "more-patience" model.

The second modification decreases the influence that agents with top track records have on the rest of the group. Call this the "less-monopoly" model. In the less-monopoly model, half of the agents do not select their informants based on track records, but choose randomly instead.[16] An agent with $m = 0.2$, for instance, will randomly choose 20% of the community as their informants each time. This dilutes the influence that opinion leaders have compared to the original model, where a small subset of agents are extremely influential, and half of the group has no impact on others' beliefs.

[15] The lower accuracy compared to the baseline model might be related to Golub and Jackson (2010)'s finding that severe imbalance in epistemic influence in a community can decrease accuracy.

[16] I also tested models with different track-record-to-randomness ratios. No qualitative difference was found when increasing or decreasing the proportion of agents randomly seeking testimony.

The results support the explanation I put forth in the previous section. Looking first at meta-expertise (fig. 4), both the more-patience and the less-monopoly models show improvement from the original testimony model. Still, the correlation between accuracy and centrality is not as strong as in the baseline model. This is not surprising, given that, in the absence of testimony exchange, there is more opinion variance in the baseline model than any other model. Since agents' predictions in the baseline model are more spread out, track records better distinguish those with accurate credences from those with inaccurate credences.

There is no initial correlation between accuracy and centrality in the first half of the more-patience model, since all agents withhold judgment until enough track records are accumulated. At the 51^{st} time step, when agents begin to assess their peers based on the accumulated information on past predictive success, the correlation between centrality and accuracy spikes to the same level as the baseline model. But this correlation weakens as soon as testimonial updates reduce the variance of opinions.

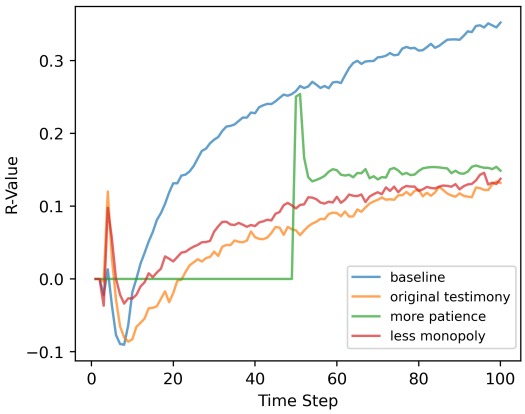

Fig. 4 R-values, averaged over 600 runs, between centrality and Brier accuracy of credences at each time step. Higher R-values indicate more meta-expertise in the scientific community.

As for overall accuracy (fig. 5), we can see that in the more-patience model, accuracy is greatly improved once agents begin to solicit testimony. Patience enables an epistemic community to reap the benefits of public track records.

In the less-monopoly model, overall accuracy improves compared to the original testimony model. However, the improvement is not nearly as significant as the more-patience model. This is explained by a difficult tradeoff the epistemic community faces. To reduce monopoly and avoid premature convergence of opinions, some agents cannot be selective with their informants. This means they will sometimes update on the testimony of peers with low predictive success, which is usually an indicator of low accuracy. Maintaining a diverse range of opinions comes at the cost of reduced mean accuracy.

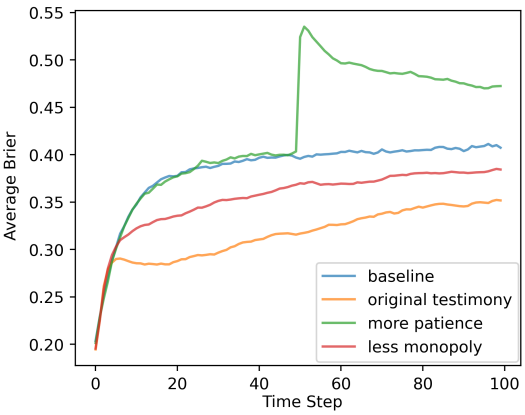

Fig. 5 Comparison of average Brier accuracy at each time step for baseline, the original testimony, less-monopoly and more-patience models (Higher is more accurate).

I take this result to suggest that, although there is amelioration in both modified models, more-patience is a better solution than less-monopoly in this context, since patience leads to significant improvement on both meta-expertise and expertise.[17]

4.3 Practical implications

So far we have seen that within an epistemic community,

1. Choosing whose testimony to solicit based on track records can reduce accuracy and group meta-expertise.
2. Delaying assessments of others' expertise until sufficient historical data on predictive success is accumulated significantly improves both meta-expertise and expertise.
3. Reducing opinion monopoly leads to a modest improvement in both meta-expertise and expertise.

An important question to consider, then, is whether the solutions in 2 and 3 are actually feasible.

Can we afford to have more patience? There are many situations where we cannot. This might be because there are not many instances of predictions to begin with. Not every scientific problem is like weather forecasts—wait a few more days, and you get a longer track record. Often times, experiments or scientific measurements are expensive, and resources are limited. When track record data are difficult to come by, it is unlikely that we can wait until we are certain that track records reflect the accuracy of underlying beliefs to start soliciting testimony accordingly. Moreover, there is sometimes pressure to take into consideration the opinions of prominent figures in a domain. Failure to

[17] This result is robust across a wide range of other parameter values tested.

take into account well-known results can be considered a flaw in one's research. Therefore, patience creates a higher barrier to publication for scientists.

If the practical cost of more patience is too high, can we settle for reducing monopoly? Is it realistic to have a significant number of unselective agents willing to solicit testimony even from scientists with relatively poor track records? Here we face a collective action problem. What we want is a mix of scientists with varying levels of exigence; this differentiation helps the community distinguish experts from those with weaker cognitive performance. Only a subset of scientists have to take the risk of randomly selecting informants, and trusting those with relatively poor track records. However, given that past predictive success is an indication of accurate beliefs, it is individually rational to be selective about testimony. Therefore, it is not clear how to motivate some but not others in the community to be unselective. (A similar social dilemma is discussed in Bruner and Holman (2022).)

In the absence of a practically implementable solution, we face a deep problem when a lot of attention is given to a small subset of scientists with widely cited early success, dictating subsequent research. While the model does not tell us whether this phenomenon has in fact occurred, it does call for caution when using track records as a guide to identify experts. The expert identification problem cannot simply be solved by making track records available and assessible.

Furthermore, the decrease of meta-expertise in the original testimony model casts doubt on the seemingly promising suggestion that laypeople could simply trust those recognized by the expert community. Even with perfect judgment of one another's track records, the mere fact that agents also update their credences on those of trusted testifiers can drastically reduce the community's meta-expertise.

While it is difficult to extrapolate findings from simplified simulations to real-world scenarios, some of the phenomena uncovered here could partially explain certain epistemic failures. For example, psychology is facing a "replication crisis," where experimental findings in many publications cannot be replicated, and are likely spurious (Nissen, Magidson, Gross, & Bergstrom, 2016). One plausible mechanism underlying this crisis is the publication bias towards positive, news-worthy results. Once these results are published, even if they are spurious effects, they immediately gain a lot of traction, and influence later research. The prestige conferred based on early results can irreversibly prevent the psychology community from later on forming accurate beliefs on the topic, as the model in this paper predicts. For instance, a series of published results about priming in psychology turned out not to be replicable, and yet priming quickly became a widespread practice in psychology research early on (See Dijksterhuis and Knippenberg (1998) and Williams and Bargh (2008)). The parallel between the model and this real-world scenario is imperfect, but it suggests that one possible explanation of reduced quality in research might be the monopoly of opinions created by high recognition attributed based on early successful results.

Sometimes it pays off to trust your own initial results even if it speaks against the view of the opinion leaders in early stages of an inquiry. Sometimes it pays off to still take seriously the opinions of those whose predictions were not initially successful.

5 Connections with existing work

Some might wonder how the above results connect with the Zollman effect (Zollman, 2007). In the Zollman (2007) model, agents face a problem where they have to determine which of two coins has a higher chance of heads, by flipping one of them each round, and updating also on the data observed by other agents they are connected with. Zollman (2007) finds that sparsely connected communities are more likely to reach the correct consensus than communities where information is widely shared between agents.

This is surprising, and is also explained by premature convergence on the false result when information traverses quickly. However, in the Zollman model, once false information circulates, agents stop observing one of the two coins. This set up is only realistic in a limited range of cases where collecting evidence from the hypothesis believed to be "worse" comes at a higher cost than not finding the truth. For instance, we might think of each coin as a medical treatment. Once we have some confidence that one treatment has a higher chance of success, we might be hesitant to explore the option that seems worse, so as to maximize the expected chance of survival for patients. In other kinds of questions, however, this will not hold. If the cost of flipping each coin is similar, then there is no reason why the scientific community would stop exploring one hypothesis without sufficient evidence. The effect uncovered here, by contrast, does not depend on this kind of stopping condition. There is only one coin to investigate, and there is no need for the inquiry to stop in order for the effect to occur.

We can therefore see that the same negative effect of premature convergence can happen even without any stopping condition in the model. Moreover, in Zollman's model, the structure of the testimony network is fixed, whereas in this model, it is constantly changing based on new information about objective track records. The dynamics of the network allow us to observe that premature convergence irreversibly reduces not only overall accuracy, but also the extent to which centrality tracks accuracy.

Bruner and Holman (2022) investigate how networks can self-assemble by reinforcement learning. To my knowledge, their model, which builds on Barrett, Skyrms, and Mohseni (2019), is the most similar to mine in the literature.[18] The task in their model is estimating a proportion (e.g., coin bias, or the probability of a future event). Agents simultaneously update evidentially, with varying reliability, and socially, by pooling information that might be "tainted" by results of previous pooling. After each time step, the truth

[18] Barrett et al. (2019)'s original model does not allow simultaneously updating on evidence and testimony.

is revealed, and the probability that the agent pools with the same agents as they did in the previous round increases in the next round if their estimation was accurate.

The networks that evolve over time by reinforcement learning are truth-conducive—the most individually reliable agents tend to observe nature, whereas unreliable agents consult the reliable agents.[19]

Both Bruner and Holman (2022)'s model and mine involve self-assembling networks, and agents learn both socially and evidentially. But their results are much more optimistic. What gives? There are several factors that can account for the different results. First, reinforcement learning is slow. In their model, before the final truth-conducive network emerges, there is a long period of time where agents choose to consult a peer or observe nature almost randomly. Since each successful consultation only changes the probability distribution of an agent's actions by a little, it takes a long exploratory stage before agents begin to consistently choose the same actions, thereby reducing the risk of premature convergence on falsehood. Second, in their model, each agent only gets information about the success or failure of the peers that they choose to consult in that round. By contrast, in my model, agents have access to the track record of all the agents in the community at once, and are therefore able to solicit testimony from those with a good track record. Third, in their model, agents not only learn who to trust, but also how many of them to trust. It is only with this flexibility that a network where the most reliable agent observes nature, and the rest simply listen, can emerge. On the other hand, in my model, the openmindedness parameter m is fixed. The most realistic way of modelling scientific inquiry is likely somewhere in between—scientists can learn and adjust the number of peers from whom to solicit testimony, but not to the extend that a small number of scientists are completely self-reliant, and the rest are completely reliant on others.

If we only look at Bruner and Holman (2022)'s model, we might think that track records are indeed the gold standard, and fail to exercise caution when it is called for. The combination of their optimistic result and my pessimistic result shows that fully available track records are not a straightforward solution, nor is prestige. It also matters how quickly and widely track records are made available, and scientists' flexibility in using or ignoring peer testimony.

Finally, Schurz (2012) endorses "meta-induction," the strategy to imitate the best performer, or a weighted combination of a subset of best performers. And Herrmann (2022) draws on the machine learning technique called "prediction with expert advice," and argues that political decision making should be guided by different expert advice weighed by past success. Although the strategies they study use track records, there are notable differences from my model. First, both of them study how a complete novice should make decisions guided by expert information, so agents are divided into two kinds: putative experts and novices. Since I am primarily interested in group meta-expertise

[19] This is similar to the findings in Barrett et al. (2019). There are multiple models in Bruner and Holman (2022). The one that is particularly similar to mine has noise, as opposed to systematic bias.

within the scientific community, there is no complete novice in my model. Each agent also makes observations in addition to receiving testimony. Second, the epistemic goal in their models is relative—the best case scenario is for everyone in the community to match the performance of the best expert. In my model, success is defined, instead, by two non-relative metrics: accuracy in predictions and the community's ability to jointly identify experts.

6 Conclusion

I have argued that, contrary to received wisdom, even perfectly accessible and evaluable track records might fail to guide us to the experts. Using a computational model, I show that when agents not only assess one another based on track records, but also update their own credences partially based on those of the best-performing testifiers, a kind of premature convergence of opinions may occur. As a consequence, accuracy and meta-expertise within the community decrease.

There are two possible solutions: either wait until long-enough track records have been accumulated before assessing expertise, or ensure that some agents are unselective with their sources of information. The former solution is more effective than the latter in all the parameter values tested. Realistically, however, it is unlikely that we can afford to wait. And it is unclear how the second solution can be implemented, given that it is individually rational to be selective about sources of testimony.

There is a lot more work to be done here, and plenty of other potential solutions to be studied. I hope to have offered a cautionary tale, and a more complete picture of the problem of expert identification, by pointing out a new depth to the difficulty of assessing expertise.

Supplementary information. The code of the simulations, written in Python, is available at [omitted].

Acknowledgments. [Omitted]

References

Babic, B., Gaba, A., Tsetlin, I., Winkler, R.L. (2021). *Resolute and correlated bayesians.* (manuscript)

Barrett, J.A., Skyrms, B., Mohseni, A. (2019). Self-assembling networks. *British Journal for the Philosophy of Science*, *70*(1), 1–25.

10.1093/bjps/axx039

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, *92*(5), 1170-1182.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1-3.

Bruner, J.P., & Holman, B. (2022). Pooling with the best. G. Ramsey & A. de Block (Eds.), *The dynamics of science: Computational frontiers in history and philosophy of science* (chap. 2). Pittsburgh, PA: University of Pittsburgh Press.

Coady, D. (2012). *What to believe now: Applying epistemology to contemporary issues.* Wiley-Blackwell.

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. A. Hajek & C. Hitchcock (Eds.), *Oxford handbook of philosophy and probability.* Oxford: Oxford University Press.

Dijksterhuis, A., & Knippenberg, A. (1998, 04). The relation between perception and behavior, or how to win a game of trivial pursuit. *Journal of personality and social psychology*, *74*, 865-77.

Douven, I., & Riegler, A. (2009, 11). Extending the Hegselmann–Krause Model I. *Logic Journal of the IGPL*, *18*(2), 323-335.

Goldman, A.I. (2001). Experts: Which ones should you trust? *Philosophy and Phenomenological Research*, *63*(1), 85–110.

Golub, B., & Jackson, M.O. (2010, February). Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, *2*(1), 112-49.

Hegselmann, R., & Krause, U. (2006). Truth and cognitive division of labour: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, *9*(3).

Herrmann, D.A. (2022). Prediction with expert advice applied to the problem of prediction with expert advice. *Synthese*, *200*(4), 1–24.

10.1007/s11229-022-03809-5

Katz, L. (1953, March). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39-43.

Mohseni, A., & Williams, C.R. (2021). Truth and conformity on networks. *Erkenntnis*, *86*(6), 1509–1530.

10.1007/s10670-019-00167-6

Newman, M.E.J. (2010). *Networks: an introduction.* Oxford; New York: Oxford University Press.

Nguyen, C.T. (2020). Cognitive islands and runaway echo chambers: Problems for epistemic dependence on experts. *Synthese*, *197*(7), 2803–2821.

Nissen, S.B., Magidson, T., Gross, K., Bergstrom, C.T. (2016, dec). Research: Publication bias and the canonization of false facts. *eLife*, *5*, e21451.

Schurz, G. (2012). Meta-induction in epistemic networks and the social spread of knowledge. *Episteme*, *9*(2), 151–170.

Williams, L., & Bargh, J. (2008, 04). Keeping one's distance the influence of spatial distance cues on affect and evaluation. *Psychological science*, *19*, 302-8.

Zollman, K.J.S. (2007). The communication structure of epistemic communities. *Philosophy of Science*, *74*(5), 574–587.