Novel Predictions for Boundedly Rational Agents: A Bayesian Analysis

Rafael Fuchs (Rafael.Fuchs@campus.lmu.de)
Munich Center for Mathematical Philosophy and Graduate School of Systemic Neurosciences,
LMU Munich, 80539 Munich (Germany)

Stephan Hartmann (S.Hartmann@lmu.de)
Munich Center for Mathematical Philosophy and Department of Psychology,
LMU Munich, 80539 Munich (Germany)

Abstract

There is no guarantee that the set of possible theories that boundedly rational agents consider contains the true theory. And yet, these agents update their beliefs as new evidence comes in, leading to a conclusion about a particular domain. In this paper, we investigate under which conditions such agents arrive at sufficiently accurate beliefs compared to ideal agents. In doing so, we work within the framework of objective Bayesianism and draw on the literature on novel predictions in philosophy of science.

Keywords: Bounded rationality, Novel predictions, Maximum Entropy, Objective Bayesianism

Introduction

Standard epistemology assumes that agents are ideally rational. In particular, it is assumed that the agent's beliefs are distributed over a given set of mutually exclusive hypotheses that contains the true hypothesis. Given enough evidence, the agent will eventually converge to the truth, that is, the agent will eventually become maximally confident in the true hypothesis. Since the true hypothesis was part of the original set of alternatives, it is not necessary to expand it in the course of the deliberation. However, real agents are different. We have no guarantee that the set of alternative hypotheses we are considering at any given time contains the true hypothesis. This can be seen in many examples where the best considered theories ultimately turn out to be wrong. Real agents are thus limited not least in the sense that they cannot perform an exhaustive search in the space of all possible theories to guarantee that the set of hypotheses under consideration contains the true hypothesis.

In this paper, we develop an objective Bayesian analysis of this kind of bounded rationality and ask under what conditions such a bounded agent can form sufficiently accurate beliefs compared to an ideal agent. More precisely, the question is whether such a boundedly rational agent is able to correctly rank these hypotheses according to their probability given all currently available evidence, as an ideal agent could do for the same (restricted) set of hypotheses and the same set of evidence. Interestingly, this question is related to a long-standing debate in the philosophy of science about the confirmatory value of novel predictions and accommodations of known data. When a new theory successfully predicts novel phenomena not previously recognised by other theories, this is considered strong evidence in favour of that theory. Therefore, many have come to believe that such novel predictions confirm a scientific theory more strongly than 'mere accommodations' of known data. The observation of novel predictions therefore speaks more in favour of a theory than if this data has already been used in the development of the theory.

The corresponding philosophical position is known as predictivism, and the predictivism-debate concerns the question whether (ceteris paribus) novel predictions confirm a scientific theory better than accommodations of known data (Lakatos, 1976; Howson, 1988; Scerri & Worrall, 2001; Barnes, 2005; Worrall, 2005). This debate is not only theoretical, because much is at stake for those sciences that have to rely more on accommodations or ex post explanations because they have limited experimental control (like social sciences or climate science). If accommodations confirm systematically less than novel predictions (or not at all), then we – as boundedly rational agents – may have only little hope of coming to accurate conclusions within these disciplines.

The aim of this paper is to develop a Bayesian analysis that identifies optimality conditions for boundedly rational agents. Our central optimality result derives from an accuracy-first approach, where agents assign their rational beliefs by minimising a measure of expected inaccuracy. This approach of minimising expected inaccuracy applies to all Bayesian agents, whether bounded or ideal. Here we take the ideal agent with the most accurate belief system as a benchmark against which we compare the constrained agent that does not know all theoretical alternatives. We find that, given sufficient knowledge of the empirical implications of their current theories, the ranking of hypotheses matches that of an ideal agent when considering the same set of alternatives.

The paper proceeds as follows. In the next section, we introduce bounded rationality in the context of the predictivism debate and explain the central challenge of bounded agents in more detail. We then present our formal analysis, based on the framework of minimising epistemic inaccuracy, and derive our central result. We conclude with a brief discussion of the formal results with respect to different epistemic contexts and point out limitations and open questions for future work.

The Challenge for Bounded Agents and Previous Modeling Approaches

In this section, we provide the conceptual background on bounded rationality in relation to the predictivism debate.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Bounded rationality in economic models takes into account computational limitations of agents to make their decision making behaviour more realistic. Ideal agents are characterised by perfect information: they know all available options and can instantly evaluate them. Thereby, ideal agents always perform an exhaustive search of the option space and find the global maximum, without having to take into account the computational costs of this search. Boundedly rational agents (see Simon (1955); Kahneman (2003); Gigerenzer & Selten (2002)), on the other hand, do not perform an exhaustive calculation to find the global optimum. Rather, they may only compare a couple of options and then settle on a satisficing solution, i.e. an option that is "good enough". This may be due to bounded agents not being aware of all options, or, relatedly, due to the cost of performing an extensive information search. For any physical being, computation (e.g. reasoning through alternatives) takes time and consumes energy. As resources are consumed, the agent may be more prone to committing errors, and as time progresses, the costs of inaction may outweigh the potential benefits of finding an alternative that is even better. Given these constraints, choosing an option that is "good enough" (e.g. exceeds a quality threshold) is a sensible decision.

In the epistemic context, we interpret the "options" as mutually exclusive hypotheses, and one central dimension of epistemic "utility" is *accuracy*, i.e. proximity to the truth, concerning all variables or propositions of interest: in criminal cases, questions of interest include the identity of the true perpetrators, and some details of the act (whatever is relevant for the appropriate verdict); in scientific theory development, central questions concern the identity of the true laws governing a domain of interest, while more local questions may concern the accuracy of concrete predictions.

Bounded epistemic agents face an important problem, in particular when the set of possible alternative hypotheses is large, or even infinite. Since the agents cannot perform an exhaustive search of all logically possible alternatives (e.g. scientific theories), there is always some fundamental uncertainty as to whether their set of epistemic alternatives at a given time really contains the *true* alternative. This is especially true in scientific contexts: there is fundamental uncertainty as to whether a given theory describes the true laws that govern the system of interest and whether future predictions of the theory will be accurate. In criminal cases, the set of suspects is considerably smaller, but the evidence concerning the case may be scarce, and relevant information pointing to the true suspect may be hidden, thus creating the risk that the true suspect gets overlooked.

Given that bounded agents don't have the resources to investigate arbitrarily large sets of alternative theories, a possible way forward is a more systematic search that is informed by their current evidence. This can happen in broadly two ways: either, agents are forced to look for alternative theories, because none of their current theories fits the evidence, or they independently look for alternative explanations, which fit the given data at least as much as the current theories.

In the first case, where evidence is anomalous for all theories, there is a strong indication that the true theory is *not* in the set of currently entertained alternatives. Hence, to solve the anomaly, agents are looking for a new hypothesis that fits the data well, that is, under which the given data has a high likelihood. However, high likelihood by itself is also no guarantee that any of the newly introduced hypotheses is the true hypothesis: one possible risk consists in the introduction of overfitting theories that explain the existing data very well, but only due to the idiosyncracies of the dataset (i.e. the hypothesis also models the noise that is in the data). The danger of overfitting is discussed in the literature concerning accommodations of old data (Hitchcock & Sober, 2004) as well as in Bayesian contexts (Sterkenburg & De Heide, 2022).

The second case, where new alternatives are put forward independently, often occurs in competitive argumentative contexts, such as lawsuits or criminal trials: here, one possible strategy for the defense attorney, rather than challenging the evidence directly, is to offer an alternative explanation, which may discount some of the weight put on the defendant as the space of alternatives increases (Hahn & Hartmann, 2020). If this competitive argumentative setup works properly, it may enable a systematic search for alternatives from multiple angles, thereby reducing the risk of prematurely settling on a wrong alternative.

In any case, the introduction of new theories necessitates a re-evaluation of the whole belief system. Since the probabilities of alternative hypotheses must sum up to unity, introducing an alternative that has positive probability necessarily decreases the probability of the old alternatives. So, the first question is what probability should be assigned to the new hypothesis, and how each of the old alternatives should be changed in response. This is known as the problem of awareness extension or language change in Bayesian decision theory and epistemology (Williamson, 2003; Karni & Vierø, 2015; Bradley, 2017; Wenmackers & Romeijn, 2016; Steele & Stefánsson, 2021). Furthermore, the resulting new credence distribution should be determined in the light of the given evidence. By Bayes' rule, the probability of each hypothesis h_i (for $i = 1, \ldots, n$) is updated on the evidence as follows:

$$P'(h_i) = P(h_i|e) = \frac{P(e|h_i) \cdot P(h_i)}{\sum_{j=1}^{n} P(e|h_j) \cdot P(h_j)}, \qquad (1)$$

where $P(e|h_i)$ is the likelihood of h_i for data e, and $P(h_i)$ is the prior probability of h_i, which is updated to the posterior $P(h_i|e)$ after the observation of e. Thus, the posterior of the new hypothesis can be computed from the likelihoods and the new prior distribution that is re-assigned after the new hypothesis was introduced. If each h_i is a statistical hypothesis that fixes the likelihood $P(e|h_i)$, the agent only needs to 'reconstruct' the prior $P(h_n)$ of the new hypothesis h_n that she would have assigned before e was observed, and renormalise the other hypotheses proportionally (so everything sums up to

1). This amounts to the known counterfactual solution to the Problem of Old Evidence. However, some philosophers have argued that precisely this step of figuring out one's counterfactual beliefs is tricky at best, or even practically impossible. Earman (1992), e.g., argued that it might not be possible to disentangle the conception of h_n (and the hypothetical prior $P(h_n)$) from the observation of e. For example, he argued that one of Einstein's main tenets in developing general relativity was precisely the explanation of the anomalous Mercury perihelion. So, if this phenomenon didn't exist in an alternate timeline, it might be questionable whether anybody would even conceive of the possibility of general relativity, or consider its initial plausibility to be high enough so that it was worthy of investigation. On the other hand, if I still assign some plausible-looking prior to h_n 'under the impression of e', and then compute the posterior, I may be *double-counting* the evidence, according to Earman. So, how could I be sure I have reconstructed an adequate prior without being overconfident (e.g. in the light of old evidence that fits well with the given evidence)? Sterkenburg & De Heide (2022) showed that overfitting hypotheses can negatively affect truth convergence of open-minded agents. Intuitively, we can connect this to Earman's worry that assigning a too high prior to a new hypothesis in the light of fitting data may generate too high confidence in overfitting hypotheses, which lead the agent astray. But conversely, if the agent discounts these hypotheses too much, in order to avoid overconfidence, they may run into the opposite problem of being underconfident. More generally, if the temporal order of hypothesis-introduction affects the agent's credences, this may lead to bad distortions, and has to be avoided by finding a principled method for prior assignments.

One solution proposal that has been put forward in the light of these considerations is to exclude old data from the assessment of new theories, and to focus instead on generating new predictions that can confirm the new theories independently. If the new theories are any better than their predecessors, they should lead to the discovery of novel facts, and make more specific predictions that can be tested independently, which is taken as evidence that they are aiming at the truth (White, 2003). After all, the new theory has to fit the old data anyway – at least as much as the old theories – in order to be considered as a real alternative in the first place. This is in line with the philosophical intuition of strong predictivism, which is expressed in the "null-support-thesis": this principle asserts that accommodations *cannot* confirm a theory at all (Giere, 1984; Glymour, 1980).

A classical argument for the null-support-thesis was already put forward by Peirce (1932) and Hempel (1966). It states that, for any finite set of data points, there are infinitely many mutually exclusive hypotheses that are equally consistent with the data. If a hypothesis is designed to fit the known data, there are infinitely many alternatives, and therefore the data couldn't give us a reason to prefer this particular hypothesis over any of the alternatives. The data never 'had a chance to refute the hypothesis', and so they also cannot confirm it. Howson (1988, 1990) has presented a rebuttal of this argument. Whether a hypothesis is 'designed' to fit the data does not matter as to whether the data supports the hypothesis. In fact, maximum likelihood estimation precisely selects the model that best fits the data. Hence, the original argument is not sufficient to rule out the possibility that old evidence *does* support a hypothesis that is designed to fit the data. Furthermore, at least in some important cases, old evidence does really seem to confirm new theories – a famous example is general relativity, which could finally explain the shift of Mercury's perihelion that posed a serious anomaly for the Newtonian theory. Hence, in its full generality, the null-support thesis seems to be too strong, but we still need to develop a criterion for deciding when accommodation of old evidence confirms a theory, and how much. Weak predictivism, on the other hand, maintains that a restricted class of accommodations can confirm, but different authors give different recommendations (White, 2003; Scerri & Worrall, 2001; Barnes, 2005; Frisch, 2015).

Thus, at this point, several questions arise. First of all, we may ask, whether it is in fact true that novel predictions generate excess confirmation, just in virtue of their novelty. Furthermore, *if* the predictivist intuition is correct, the question remains by how much accommodations of old data should be discounted. If accommodations have to be significantly discounted (even if not completely), the prospects seem dire in all those situations where we have to rely (almost exclusively) on existing data to confirm our theories. If, on the other hand, old data must count to *some* extent, the question is how we should account for it. In order to tackle this question and to see how novel predictions and accommodations ought to be compared, we first need to solve the problem of how priors have to be assigned, in order to avoid distortions due to temporal order effects. We tackle this issue in the next section, by introducing an objective Bayesian account, based expected inaccuracy minimisation.

An Objective Bayesian Model

In this section, we start by tackling the question of rational prior assignment from an objective Bayesian perspective. The objective prior assignment will be grounded in considerations of epistemic inaccuracy minimisation. Hence, it is an optimal solution, which justifies the way in which we go about comparing the confirmatory value of old evidence and novel predictions. Finally, we will use this framework to answer the question whether *ceteris paribus*, novel predictions confirm a theory more than accommodations of old evidence.

First of all, in keeping with our preceding discussion of interpreting epistemic utility as accuracy, we need to ground prior assignments in a principle of epistemic optimality. This can be achieved within the framework of expected inaccuracy minimisation. If the goal of an epistemic agent is to have maximally accurate beliefs, we can use scoring rules to measure the inaccuracy of the agent's subjective degrees of

belief, relative to the truth. In this context, strictly proper scoring rules are of special interest. A scoring rule $S(\omega, P)$ is a real-valued function of a materialised outcome $\omega \in \Omega$, and the agent's degree of belief in that outcome $P(\omega)$. The score of S represents the inaccuracy of the agent's belief (P is fully accurate if $P(\omega) = 1$).

S is called *strictly proper* if its expected value $\mathbb{E}_Q S(\Omega, P)$ under a given distribution Q (e.g. relative frequency) is minimal if and only if $P = Q$. Thus, the score S measures how well-calibrated a belief P is relative to an optimally predictive distribution Q (e.g. empirical frequencies).

For every strictly proper scoring rule, the situation of complete ignorance is represented by a uniform distribution. However, often agents have some (yet incomplete) prior information. Hence, the agent needs to minimise their expected inaccuracy subject to those constraints that incorporate their prior information. In the case of scientific theories that specify statistical hypotheses, prior information is contained in the likelihood distributions $P(E|t)$ of each theory t over observational variables. Such objective likelihoods (Jeffreys, 1998; Wenmackers & Romeijn, 2016; Howson, 2017) encapsulate the empirical consequences of a theory, in terms of how likely each observable outcome is, given that theory. With this prior information in place, the agent obtains further information by observing variables, which leads them to updating their beliefs (and has to be considered as prior information when introducing new theories). Standard updating proceeds via Bayesian conditionalisation, but we will introduce a model that is more general. We will also show that by minimising expected inaccuracy, it is always possible to recompute every update so that we can also reconstruct every state of prior information.

As mentioned, in our model evidence can be more complex than what is captured by simple conditionalisation. For one thing, evidence can be uncertain in the sense that the probability is raised, i.e. $P'(e) > P(e)$, but still $P'(e) < 1$, which requires Jeffrey's (1965) rule. Going further, evidence can be even more complex, e.g. when the agent learns about *correlations* between observable variables – for example between temperature and pressure. This involves learning conditional probabilities $P'(a|b)$, which go beyond Jeffrey's rule. Thus, the most general update involves minimising an f-divergence (which is a class of functions that measure 'how different' two probability distributions are). There is only one f-divergence that is also related to a strictly proper scoring rule (Eva et al., 2020), that is, by minimising the distance from the prior distribution, the agent also minimises their expected inaccuracy: this is the Kullback-Leibler divergence (Kullback & Leibler, 1951), which is associated with the logarithmic scoring rule $S(\omega, P) = -\log P(\omega)$:

$$D_{KL}(Q||P) = \sum_{\omega \in \Omega} Q(\omega) \log \frac{Q(\omega)}{P(\omega)} \tag{2}$$

The KL-divergence is also called *relative entropy*, because it is closely related to Shannon's (1948) measure of expected information content, which is called *entropy*:

$$H(P) = -\sum_{\omega \in \Omega} P(\omega) \log P(\omega) \tag{3}$$

Alternatively, when P is defined over random variables[1] $X_1, \ldots, X_n$, we can also write $H(X_1, \ldots, X_n)$ instead of $H(P)$.

Objective Bayesianism is centered around the *Principle of Minimum Information* (Williams, 1980), which maintains that a rational agent ought to assign probabilities that are only as extreme as is necessitated by known evidence, and otherwise remain as equivocal as possible. Technically, this means that (i) $H(P)$ (eq. (3)) is *maximised* (Jaynes, 1968), to make the prior as uniform as possible, and (ii) $D_{KL}(Q||P)$ (eq. (2)) is *minimised*, to update the prior distribution just as much as is necessitated by the new evidence. The principle of minimum information works as an extension of Proposition 1, which enables us to identify the optimal, inaccuracy-minimising prior for any set of probabilistic constraints

In order to capture novel predictions, we also allow for partially specified likelihood distributions. As mentioned, the likelihood of a theory represents its empirical consequences. If a new theory introduces a novel phenomenon E', this new E' is outside the domain of previous theories – thus, the likelihood of those theories is not yet defined for E'. In this case, maximising entropy (MaxEnt) given all known constraints yields a precise prior distribution, even if not all likelihoods are specified.

With this framework, we can now propose a general strategy through which a bounded agent can identify the ideal agent's credences on a restricted domain. Here is the setup: The ideal agent entertains all logically possible and mutually exclusive theories (random variable T) and a collection of evidential variables that contain all possible observations (*set* of random variables $\mathbf{E}$). Importantly, the t_i, t_j ($i \neq j$) are mutually exclusive. This means that we are considering theories that are maximally specified. The bounded agent entertains a subset $T' \subset T$ of all possible theories and a subcollection $\mathbf{E}' \subset \mathbf{E}$. Introducing a new theory corresponds to adding a new value t_{n+1} to T' (denote the expanded variable as T'_{n+1}.) Analogously, introducing a novel prediction corresponds to adding a new observable magnitude E^{m+1} to $\mathbf{E}'$. We denote the ideal agent's probability distribution at time step n as Q^n, and the bounded agent's analogously as P^n. Both agents have a set of prior constraints $\mathcal{K}_T$ (ideal) and $\mathcal{K}_{T'} \subset \mathcal{K}_T$ (bounded), which contain information the likelihood distributions of each element of T and T', respectively. The prior distributions Q^0 and P^0 are obtained by MaxEnt given the respective set of constraints.

Upon obtaining new constraints $\mathcal{E}$ containing information about empirical probabilities $Q(E^j|E^k)$ for $E^j, \ldots, E^k \subseteq$

[1] We use capital letters A, B, C to denote random variables and lowercase letters $a_1, \ldots, a_n$ to denote variable-values or propositions (the binary values are a and $\neg a$). Thus $P(a_i)$ (short for $P(A = a_i)$) is the probability that the variable A takes the value a_i. Furthermore, $P(A, B)$ refers to the whole joint distribution over the variables A, B, and $P(A|A')$ refers to the distribution over the variable A if its set of possible values is restricted to a subset $A' \subseteq A$.

E, the ideal agent updates $Q^n \rightarrow Q^{n+1}$ by minimising $D_{KL}(Q||Q^n)$ subject to $\mathcal{E}$.

In contrast to that, the bounded agent can not only learn new empirical constraints, but also develop new theories, and derive their observational consequences. Thus, the bounded agent can expand the value range of T' and the set $\mathbf{E}'$. We will show that, in order to combine both learning modes so as to approximate the ideal agent's order of posteriors (expressing which theories are the best-supported ones), the bounded agent can employ the following strategy:

1. Upon learning new constraints $\mathcal{E}$ containing information about empirical probabilities $P(E^j|E^k)$ for $E^j, E^k \subseteq \mathbf{E}'$, the update proceeds by minimising $D_{KL}(P^{n+1}||P^n)$ subject to $\mathcal{E}$.

2. Upon becoming aware of a new theory t_{n+1} (and novel predictions $E^{m+1}, \ldots$), the bounded agent *first* recomputes their prior P^n over the extended set of variables, by MaxEnt given the extended set of prior constraints $\mathcal{K}_{T'_{n+1}}$. *Second*, the agent minimises $D_{KL}(P^{n+1}||P^n)$ given $\mathcal{E}$.

Thus, there are two kinds of variables of which the agent can be unaware: (1) alternative hypotheses over the same domain (set of observables), and (2) further empirical consequences (predictions) of a given hypothesis. Perhaps intuitively, one might suspect that(1) has a stronger effect on the agent's accuracy than (2). Thus, if we consider a fixed set of hypotheses at a given point in time (i.e. the agents have done their best to come up with a set of plausible hypotheses, and now they need to evaluate them), we can ask how knowing the predictions of all theories may affect their ranking, in comparison to an ideal agent, who knows all conceivable theories and empirical variables. We formalise the predictions of a theory via the theory's likelihoods, which are captured as prior constraints (as introduced above). It turns out that being aware of all implications of the current theories (i.e. knowing all prior constraints on likelihoods) is sufficient for emulating the ideal agent's ranking over the given set of hypotheses. However, when this condition fails, it is possible that the bounded agent deviates from the ideal ranking, as we will see in the next section. Formally, we define this awareness condition as follows.

Empirical Closure: If there are variables $E^k, \ldots, E^l$ that are not entertained by the bounded agent (i.e. not in **E'**), then for every t_i in T', there are no prior constraints on the conditional probability distributions $P(E^k, \ldots, E^l|\mathbf{e}, t_i)$ for every value configuration **e** of **E'**.

This means that the given theories have nothing to say about variables $E^k, \ldots, E^l$ outside the current domain, or in other words, all empirical implications (for which the given theories *do* posit prior constraints) are known by the agent. Even though knowing all empirical consequences is demanding for real agents, it is in principle realistic, as opposed to knowing all possible theories. Furthermore, it is certainly reasonable as a goal of rational inquiry: if a theory is formulated, all of its potential implications need to be derived (even if they cannot be tested right away), because this is needed to guarantee optimal credences.

With these assumptions in place, the following proposition holds:

Proposition 1. *Consider a single update of each agent's priors (Q^0 and P^0) upon learning some fixed (possibly empty) set of empirical constraints $\mathcal{E}$. Then, under empirical closure, it follows that $P^1(T') = Q^1(T|T')$, where $P^1(T')$ is the bounded agent's posterior over T', and $Q^1(T|T')$ is the ideal agent's posterior over T, restricted to the value-range of T'.*

As a corollary, we get that ceteris paribus (for the same set of likelihoods), accommodations of old evidence and novel predictions yield the same degree of confirmation.

Discussion

In this section, we use our results to answer the questions raised in our conceptual discussion, and point out limitations that can be tackled in future research.

First, we have answered the question of how to (re-)assign priors in a principled, non-arbitrary way – namely, by maximising the agent's expected accuracy via the logarithmic scoring rule, which leads to the principle of minimum information. Since both ideal and bounded agents will follow this principle of epistemic rationality, we can also directly compare the bounded agent's beliefs and resulting accuracy to the ideal benchmark. In doing so, we find that it is possible for the ideal agent to emulate the ideal belief system under the condition of empirical closure.

As a consequence of this, we can also propose a first answer to the question of predictivism: ceteris paribus, novel predictions and accommodations have the same confirmatory power. This is, because the prior distribution can be recomputed at any point, independently of the temporal order in which hypotheses were formulated, and it only depends on the distribution of likelihoods. To see this, consider equation 6 (in the Appendix). Since $H(E|t_k)$ is maximal if $P(E|t_k)$ is uniform, and minimal (zero) if $P(e|t_k) = 1$ for exactly one $e \in E$ (i.e. maximally extreme), it follows that $P(t_i)$ depends on how extreme its likelihood distribution is, relative to the other theories $t_{j\neq i}$. That is, the theory that is relatively the strongest (intuitively: 'rules out the most observations') has the lowest (absolute) prior probability. On the flip side, if the likelihood distribution is very concentrated, the corresponding theory will get a much higher confirmatory boost, if the predicted outcome (that is most likely under that theory) is observed and differs from the predictions of the other theories. With this prior probability at hand, we can also recompute the posterior for every set of observations – including old evidence, without having to worry about distortions due to an overconfident prior.

This may be good news regarding those cases, in which the agent has to rely mostly or exclusively on old data, because old data can count just as much as new observations. As shown in proposition 1, it is possible for the bounded agent to emulate the ideal agent for a given set of hypotheses and

empirical observations. However, this positive result depends on the bounded agent being sufficiently aware of all relevant empirical implications (empirical closure), which can be a demanding requirement. If empirical closure is violated, on the other and, there are cases where the bounded agent will even get the ordinal ranking of hypotheses wrong.

Consider the following example: $T = T' = \{t_1, t_2\}$, and $\mathbf{E'} = \{E_1\} \subset \mathbf{E} = \{E_1, E_2\}$, with the following conditional entropies: $H(E_1|t_1) = 0.5$, $H(E_1|t_2) = 0.9$, $H(E_2|t_1) = 0.9$, and $H(E_2|t_2) = 0$ (where, for simplicity, E_1 is conditionally independent of E_2, given T). Now, the prior distribution of the ideal agent (before observing the value of either E_1 or E_2) is

$$P_i(t_1) = \frac{e^{0.5+0.9}}{e^{0.5+0.9} + e^{0.9+0}} \approx 0.62246,$$

and thus, $P_i(t_1) > P_i(t_2)$. However, for the bounded agent, who is only aware of $\mathbf{E'}$, we have

$$P_b(t_1) = \frac{e^{0.5}}{e^{0.5} + e^{0.9}} \approx 0.40131,$$

and thus, the ordering is reversed: $P_b(t_1) < P_b(t_2)$. This is due to the fact that $H(E_e|t_2)$ is minimal, which means that t_2 predicts a particular value of E_2 with certainty. That is, the prediction of t_2 with respect to E_2 is maximally specific, as opposed to that of t_1. Therefore, under the full algebra, t_2 is discounted much more than t_1, which is the reason for the reversal, if E_2 is left out. Conceptually, we can understand this as follows: If the bounded agent is ignorant of E_2, they are in fact missing a major prediction of t_2, and therefore severely misjudge the potential empirical content of that theory that could set it apart from its competitor t_1 (if the prediction is not confirmed, t_2 is refuted, and if it is confirmed, t_2 gets a much higher confirmatory boost than t_1). On the other hand, if the agent misses a prediction that is not very specific in comparison to the competitor, the correct order may still hold. In our example, suppose that $H(E_2|t_2) = 0.7$ instead of $H(E_2|t_2) = 0$. Then, with everything else unchanged, $P_b(t_1) > P_b(t_2)$, even if $\mathbf{E'} = \{E_1\}$ (i.e. the bounded agent is not aware of E_2). Thus, for the bounded agent to get the theory-ranking approximately right, it is primarily important to be aware of *strong* predictions that set apart some theories in a certain subdomain, even if those predictions cannot be tested right away. Moreover, if the agent is aware of all empirical implications of all theories under consideration, the agent is guaranteed to get the ordering *and* the relative preference for each hypothesis right, as stated in Proposition 1.

As mentioned, empirical closure is a strong requirement, and bounded agents in our model are still ideal Bayesian reasoners in the sense that they immediately know (with maximal probability) all logical truths that hold in their language. Furthermore, we have restricted our attention to epistemic accuracy, i.e. proximity to the truth. However, in practical contexts (e.g. the aforementioned lawsuit), other values besides accuracy may also be important. Last but not least, time and computational resources may still make our solution unattainable in cases where the space of alternatives is large. Taking these computational constraints into account in more detail will be a fruitful task for future research.

Conclusion

In this paper, we have shown that bounded Bayesian agents who are not aware of all conceivable theories can approximate the ideal agent's credences if they are at least aware of all empirical implications of their theories. As a consequence, novel predictions and accommodations have the same confirmatory power, ceteris paribus. This follows from the central result that bounded agents can use a method, based on expected inaccuracy minimisation, to figure out their optimal degrees of belief, relative to the given subset of theories that they currently entertain. This provides hope for disciplines which have limited experimental control. Therefore, in concluding our paper we suggest that, in future research our result can be fruitfully applied to study more concrete issues that arise in the special sciences, such as model confirmation in climate science, and issues concerning the replication crisis in psychology.

References

Barnes, E. (2005). On Mendeleev's predictions: Comment on Scerri and Worrall. *Studies in History and Philosophy of Science*, *36*(4), 801–812.

Bradley, R. (2017). *Decision Theory with a Human Face*. Cambridge: Cambridge University Press.

Earman, J. (1992). *Bayes or Bust?: A critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

Eva, B., Hartmann, S., & Rafiee Rad, S. (2020). Learning from conditionals. *Mind*, *129*(514), 461–508.

Frisch, M. (2015). Predictivism and old evidence: A critical look at climate model tuning. *European Journal for Philosophy of Science*, *5*, 171–190.

Giere, R. (1984). *Understanding Scientific Reasoning*. New York: Holt, Rinehart, and Winston.

Gigerenzer, G., & Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Glymour, C. N. (1980). *Theory and Evidence*. Princeton: Princeton University Press.

Hahn, U., & Hartmann, S. (2020). Reasonable doubt and alternative hypotheses: A Bayesian analysis. (PhilSci-Archive: http://philsci-archive.pitt.edu/18472/)

Hempel, C. G. (1966). *Philosophy of Natural Science*. Englewood Cliffs: Prentice-Hall.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, *55*(1), 1–34.

Howson, C. (1988). Accommodation, prediction and Bayesian confirmation theory. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 1988(2), pp. 381–392). Cambridge.

Howson, C. (1990). Fitting your theory to the facts: Probably not such a bad thing after all. In C. W. Savage (Ed.), *Minnesota Studies in the Philosophy of Science, Vol. XIV* (pp. 224–244). Minneapolis: University of Minnesota Press.

Howson, C. (2017). Putting on the Garber style? Better not. *Philosophy of Science*, *84*(4), 659–676.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*(3), 227–241.

Jeffrey, R. C. (1965). *The Logic of Decision*. Chicago: University of Chicago Press.

Jeffreys, H. (1998). *The Theory of Probability*. Oxford: Oxford University Press.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475.

Karni, E., & Vierø, M.-L. (2015). Probabilistic sophistication and reverse bayesianism. *Journal of Risk and Uncertainty*, *50*(3), 189–208.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lakatos, I. (1976). *Falsification and the Methodology of Scientific Research Programmes*. Berlin: Springer.

Peirce, C. S. (1932). *Collected Papers* (Vol. II). Cambridge, MA: Harvard University Press.

Scerri, E. R., & Worrall, J. (2001). Prediction and the periodic table. *Studies in History and Philosophy of Science Part A*, *32*(3), 407–452.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

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

Steele, K., & Stefánsson, H. O. (2021). Belief revision for growing awareness. *Mind*, *130*(520), 1207–1232.

Sterkenburg, T. F., & De Heide, R. (2022). On the truth-convergence of open-minded bayesianism. *The Review of Symbolic Logic*, *15*(1), 64–100.

Wenmackers, S., & Romeijn, J.-W. (2016). New theory about old evidence. *Synthese*, *193*(4), 1225–1250.

White, R. (2003). The epistemic advantage of prediction over accommodation. *Mind*, *112*(448), 653–683.

Williams, P. M. (1980). Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, *31*(2), 131–144.

Williamson, J. (2003). Bayesianism and language change. *Journal of Logic, Language and Information*, *12*(1), 53–97.

Worrall, J. (2005). Prediction and the 'periodic law': A rejoinder to Barnes. *Studies in History and Philosophy of Science Part A*, *36*(4), 817–826.

Appendix

Proof of Theorem 1

In order to find the maximum entropy prior P' for the bounded agent, subject to known likelihoods, we maximise the Lagrangian function

$$L = H(P') + \sum_{\phi \in \mathcal{K}} \lambda_\phi \cdot \phi, \tag{4}$$

where each $\phi \in \mathcal{K}$ is a constraint on known likelihoods, such that for a given value configuration $e_j^1, \ldots, e_k^m, t_i$ of $E^1, \ldots, E^m$ and T_b there is $a \in [0,1]$ such that $P(e_j^1, \ldots, e_k^m | t_i) = a$. Taking the partial derivatives of L and setting them equal to zero yields predefined values for all likelihoods $P(e|t_i)$ that are constrained by $\mathcal{K}_b$, while

$$P(e_j^1, \ldots, e_k^m | t_i) = \frac{1 - \sum_{\phi_{t_i} \in \mathcal{K}} \phi_{t_i}}{|\{\psi_{t_i} : \psi_{t_i} \notin \mathcal{K}\}|}, \tag{5}$$

where, for fixed t_i, ϕ_{t_i} are those constellations of $E^1, ..., E^m$ for which $P(e|t_i) \in \mathcal{K}$ while ψ_{t_i} are those constellations with $\psi_{t_i} \notin \mathcal{K}$. That is, the probabilities of *unconstrained* likelihoods are evenly spread out over the remaining mass that is left over by the constrained ones.

Finally, the prior distribution over T_2 is given by

$$P_b(t_i) = \frac{e^{H(E^1, \ldots, E^m | t_i)}}{\sum_{j=1}^{n+1} e^{H(E^1, \ldots, E^m | t_j)}}, \tag{6}$$

where

$$H(E|t_k) = -\sum_{e \in E} P(e|t_k) \log P(e|t_k) \tag{7}$$

is the entropy of E conditional on theory t_k.

By empirical closure, there are no constraints on $E^{m+1}, \ldots, E^q$ *given* any value of T_b, which leads to a uniform conditional distribution and full conditional independence of $E^{m+1}, \ldots, E^q$ given T_b. Thus, for $t_i \in T_b$, the joint entropies are additive:

$$H(E^1, \ldots, E^m, \ldots, E^q | t_i) = H(E^1, \ldots, E^m | t_i) + H(E^{m+1}, \ldots, E^q | t_i)$$

, which means that a constant factor α is added to each exponent in equation 6 to obtain the prior of the ideal agent, conditional on T_2:

$$P_i(t_j | T_b) = \frac{e^{H(E^1, \ldots, E^m | t_j) + \alpha}}{\sum_{k=1}^{n+1} e^{H(E^1, \ldots, E^m | t_k) + \alpha}}, \tag{8}$$

which cancels out, and therefore yields the same values.

Finally, upon observing any evidence $\mathcal{E}$ that is expressible in terms of $E^1, \ldots, E^m$, both agents minimise $D_{KL}(Q||P^o)$ subject to the same set of constraints $\mathcal{E}$, where $P^o = P_b$ for the bounded agent, and we consider $P^o(\cdot) = P_i(\cdot|T_b)$ for the ideal agent. Since these are identical, it follows that $Q_b(T_b) = Q_i(T_i|T_b)$, as required. □