In Search for Optimal Methods: New Insights About Meta-Induction

Gerhard Schurz

Journal for General Philosophy of Science (2023) 54:491–522
https://doi.org/10.1007/s10838-023-09649-2

ARTICLE

In Search for Optimal Methods: New Insights About Meta-Induction

Gerhard Schurz[1]

Accepted: 30 March 2023 / Published online: 4 September 2023

Abstract
In this paper, the contributions to the account of meta-induction (Schurz 2019) collected in this volume are critically discussed and thereby, new insights are developed. How broad and expandable the program of meta-induction is can be learned from Ortner's contribution. New insights about the transition from the a priori justification of meta-induction to the a posteriori justification of object-induction emerge from the reflection of Shogenji's paper. How meta-induction may be applied also to religious prophecies and that their meta-inductive justification does not fail for a priori reasons but because of missing evidence for predictive success is learned from the discussion of Pitts' contribution. That meta-induction does not rely on a particular prior distribution, while the no free lunch theorem depends implicitly on a uniform prior, is the major conclusion drawn from the discussion of Wolpert's article. How the problem of induction is treated in different versions of the Bayesian account is learned from the discussion of Willliamson's paper. That meta-induction can also be employed for abduction, and that abductive theory-revision can offer meta-inductive aggregation methods is a new insight emerging from the reflection of Aliseda's contribution.

Keywords Meta-induction · Online learning under expert advice · Problem of induction · No free lunch theorem · Religious prophecy · Abductive belief revision

1 Meta-Induction: Epistemological Account and Scientific Research Program—or Lessons from Ortner

In what follows, "induction" is broadly conceived as the projection of observed patterns from the past to the future. The problem of induction was raised by David Hume 250 years ago. Hume argued that all standard methods of justification fail when applied to the task of

✉ Gerhard Schurz
schurz@hhu.de

[1] Department of Philosophy, Heinrich Heine University Duesseldorf, 40225 Duesseldorf, Germany

justifying induction. More specifically, induction cannot be justified by induction from the observation of its past success as follows:

Inductive Justification of Induction: Induction has been successful in the past, thus by induction it will be successful in the future.

This argument is circular and without justificatory value. As Salmon (1957, 46) has pointed out, counter-induction (that predicts the opposite of induction) can be pseudo-justified in the same circular manner:

Counter-inductive Justification of Counter-induction: Counter-induction was unsuccessful in the past, thus by counter-induction it will be successful in the future.

The circularity problem besets also the justification of a prior probability distribution—for short: a prior—for the predictions of event probabilities. While the predictive success of a chosen prior depends on the future course of events, this future course can only be probabilistically assessed by assuming (explicitly or implicitly) a particular prior. Bayesians reply that the influence of the prior can be washed out by conditionalizing the posterior probabilities on increasing amounts of evidence, but this reply has two limitations: (1) Not all prior distributions can be washed out in this way, not even in the long run. For example, Carnap's (1950, 564–566) $m^{\dagger}$ measure that assigns a uniform probability (density) to all possible event-sequences, or states of the world, cannot be washed out, because it makes inductive learning impossible; in what follows we call such a distribution a *state-uniform* distribution. (2) In the short run the situation is worse: for every finite amount of evidence there exists a suitably biased prior that prevents learning from this evidence (Schurz 2019, 66, prop.).

In conclusion, the crucial challenge of Hume's problem is to find a *non-circular* justification of induction. Such a justification has to be *a priori* in the sense that it does not assume anything about the future or the unobserved part of the world. A justification attempt of this sort was proposed in Reichenbach's "best-alternative" account to induction. Reichenbach (1949) argued that induction is the best one can do to achieve successful predictions. What Reichenbach attempted here is an *optimality justification*. Optimality justifications are epistemologically weaker than reliability justifications. They do not establish that the predictive success of a prediction method is reliable, in the sense of being greater than a certain threshold that is greater than random success. An a priori demonstration of the reliability of induction is impossible because of the possibility of skeptical scenarios in which *no method* can be successful. But skeptical scenarios are compatible with optimality justifications, because even in skeptical scenarios a method may be optimal in the sense of "being the best of a bad lot".

By *object-induction*, abbreviated as OI, we mean induction applied at the level of events. Reichenbach's best-alternative account failed because it was developed for OI. What blocks Reichenbach's account is the possibility of methods that are superior to OI, e.g., methods based on clairvoyance or on other 'paranormal' effects. This possibility cannot be excluded a priori (Skyrms 1975, ch. III.4).

Schurz (e.g., 2008, 2019) develops a new optimality approach by applying induction at the level of methods, called *meta-induction*, abbreviated as MI. In general, an MI method tries to find an optimal prediction method by basing its prediction on the predictions and the observed track records of the set of (simultaneously) accessible methods; this set is called the *pool* of *candidate* methods. There are different versions of meta-induction. The simplest meta-inductive method is imitate-the-best, abbreviated as ITB. It predicts what the hitherto best method predicts (ties are resolved by taking the 'first' best method in an assumed ordering). More complex MI methods predict a weighted average of predictions of the accessible methods, with weights being correlated to their success.

The candidate pool may contain object-level methods of various sorts, inductive as well as non-inductive ones.[1] The simplest version of an OI method is the straight rule that projects the observed frequency of an event as its probability to the future; the binarized version of this rule predicts 1 (event happens) if the event's probability is greater-equal 0.5, and 0 (event does not happen) otherwise. More complicated OI methods are the Carnapian λ-rules, Bayesian updating of priors, etc. Generally considered, the number of different OI methods is countless, but the meta-inductive account does not depend on a sharp general definition of an 'inductive method'. On the other hand, clear examples of non-inductive methods are counter-inductive methods, blind guessing methods and agent-based methods relying on (purported) clairvoyance.

Meta-induction can handle the clairvoyant objection because if there would really be an accessible prediction method that is superior to scientific induction, meta-inductivists would base their predictions on this method. More generally, meta-induction avoids the objection of formal learning theory (Putnam 1965; Kelly 1996) against 'absolute' optimality because it restricts its optimality theorems to epistemically *accessible* methods. These are prediction methods whose predictions and track records are accessible to the epistemic agent (cf. Schurz 2019, def. 5-3 for a precise definition). This restriction is justified because methods that are epistemically inaccessible are epistemically irrelevant.

By transforming the best-alternative account to the meta-level, the optimality of meta-induction becomes mathematically demonstrable. The demonstration is carried out in the framework of *prediction games*. A prediction game consists of an infinite sequence of (binary or real-valued) events e_1, e_2, …, whose normalized values $v \in Val$ lie in the real-valued interval [0,1], together with a given meta-inductive method MI and a candidate pool $\mathbf{C} = \{M_1, \ldots, M_m\}$ of candidate methods (or 'players') that are either given as algorithms simulated by MI or as external agents whose predictions are available to MI. More formally, the set of possible event values Val is a subset of [0,1]; "v" ranges over possible event values ($v \in Val$), $e: N \rightarrow Val$ is the event variable, with $N = \{0,1,2 \ldots\}$ being the set of time points or rounds of the game, and e_n refers to the actual (true) event at time n. For binary events, $Val = \{0,1\}$, where 1 stands for the event's occurrence and 0 for its non-occurrence.

In each round $n \in N$ each method $M \in \{MI\} \cup \mathbf{C}$ delivers a prediction $pred_{n+1}$ (M) of the next event e_{n+1}. The predictions, too, take their values in [0,1]. Importantly, in real-valued prediction games it is permitted to predict *weighted averages* or *probabilities* of event values; thus the set of possible prediction values, Val_{pred}, may be a superset of Val ($Val \subseteq Val_{pred} \subseteq [0,1]$). The predictions of real-valued games are scored by a *convex* loss function, loss ($pred_n$, e_n), normalized within the interval [0,1]. Convexity means that for any two predictions and weight $w \in [0,1]$ the loss of their w-weighted average is not greater than the w-weighted average of their losses. Typical convex loss functions are the absolute distance $|pred_n - e_n|$ or the squared distance $(pred_n - e_n)^2$ between e_n and $pred_n$. Note that instead of the next event, also the k 'next' future events, or a parameter of a sample of future events may

[1] For epistemological reasons the candidate pool is even allowed to contain other meta-level methods.

be predicted; this doesn't change the results, given a suitably adapted loss function (Schurz 2019, Sect. 7.4).

The score obtained by a method M in round n is defined as 1-loss ($pred_n$ (M), e_n), M's cumulative score or *absolute success* achieved in round n, Suc_n (M), is the sum of M's scores until round n, and M's average score or *success rate* at round n, suc_n (M), is defined as Suc_n (M)/n. For binary predictions with an absolute loss function, their success rate is identical to their truth frequency.

Based on theorems in machine learning (Cesa-Bianchi and Lugosi 2006; Schurz 2019, memo (6.8.), theorem 6.9) shows that a certain form of meta-induction is universally optimal among all accessible methods. This method is called *attractivity-based* meta-induction, abbreviated as aMI, and it predicts a weighted average of the predictions of the candidate methods:

(1) $$\begin{aligned}&\mathrm{pred}_{n+1}(\mathrm{aMI}) = \textstyle\sum_{1\le i\le m} \mathrm{w}_n(\mathrm{M_i})\cdot \mathrm{pred}_{n+1}(\mathrm{M_i}) \,/ \textstyle\sum_{1\le i\le m}\mathrm{w}_n(\mathrm{M_i}),\\ &\text{where } \mathrm{w}_n(\mathrm{M_i}) = \exp(\eta\cdot \mathrm{S}_n(\mathrm{M_i})), \text{ with } \eta = \sqrt{8\cdot \ln(m)/(n+1)}\end{aligned}$$

(2) *Optimality result for aMI*: For every possible event sequence and candidate pool:

(2.1) MI's 'long run' success rate is never worse and sometimes better than the maximal success rate (max_n) of the candidate methods ($\mathrm{limsup}_{n\to\infty}(max_n - suc_n(aMI)) \le 0$).

(2.2) In the 'short run', small losses of aMI compared to the actually best method, so-called 'regrets', are unavoidable; however, these regrets have the following tight worst-case bound:

$$\mathrm{max_n} - \mathrm{suc_n(aMI)} \le 1.77\cdot\sqrt{\ln(m)/n},$$

so they converge quickly to zero when n grows larger than m.

This optimality result holds even in 'paranormal' environments that host clairvoyants or adversarial methods that try to deceive aMI, as well as in chaotic environments in which the method's success rates do not converge to a stable performance ordering but oscillate forever. Not all meta-inductive methods are universally optimal in this sense. For example, ITB is not access-optimal: ITB's success rates may be driven down to zero by deceiving methods that lower their success rate as soon as ITB imitates them (cf. Schurz 2019, 128). aMI is optimal because the weight it assigns to a candidate method reflects its 'attractivity' (or 'regret'), which means that w_n (M_i) increases with M_i's success but becomes zero or negligibly small if M_i continues to perform worse than aMI. So if the candidate pool contains a sustainably superior method M*, aMI will soon assign all weight to M* and behave as ITB.

Finally, note that the universal optimality of aMI does not imply that aMI is universally dominant, in the sense that aMI beats every other method in at least one world. There are different variants of attractivity-weighted MI that are equally long-run optimal with different short-run properties; so aMI is not universally dominant. However, the following restricted dominance theorems have been proved (Schurz 2019, prop. 8.4, 8.5):

(3) *Dominance results for aMI*: aMI dominates (a) all independent methods and (b) all meta-methods that are not universally access-optimal. The latter ones subsume (among

others): (b1) all one-favorite methods (who at each time point imitate the prediction of one accessible method), (b2) success-based weighting (using success rates as weights), (b3) linear regression for a linear loss function, and (b4) 'simply' non-inductive meta-methods (as defined in Schurz 2019, def. 8.2).

A final remark: the optimality (and partial dominance) of meta-induction must not be misunderstood as if meta-induction *in isolation* were the best possible strategy. Without competent candidate methods, meta-induction cannot be successful. What the optimality theorem justifies is that meta-induction is universally recommendable *ceteris paribus*, as a strategy applied *on top* of one's toolbox of candidate methods. At the same time it is always reasonable to improve one's pool of candidate methods, which is, however, not an objection against the a priori justification of meta-induction.

By itself, the a priori optimality of meta-induction does not entail anything about the rationality of object-induction. Which prediction method, or combination of methods, is meta-inductively evaluated as optimal is an *a posteriori* matter that depends on the empirically given track record of the accessible methods. The possibility of superior non-inductive methods cannot be excluded a priori. However, the a priori justification of MI provides us with the following *a posteriori justification* of OI: to the extent that object-inductive prediction methods were observed as more successful than all accessible non-inductive prediction methods, we are meta-inductively justified in continuing to favor object-inductive prediction methods in the future. This argument is no longer circular, because a non-circular justification of meta-induction has been independently established. In conclusion, the proposed solution to the problem of induction consists of two parts: (i) the a priori (mathematical) justification of meta-induction and (ii) the a posteriori (empirical) justification of object-induction based on (i).

Apart from its epistemological application to foundation-theoretic epistemology (cf. Schurz, 2022a), meta-induction is a broad research-program serving the purpose of improving scientific forecasting. There is not only one but there are many different object-inductive methods with different performance strengths in different environments; so especially in situations involving unforeseeable changes of the environment, meta-induction can serve as a means to optimize scientific forecasting by aggregating the predictions of different methods. Viewed from this perspective, strategies of meta-induction have been generalized and refined in various ways, especially in the machine learning literature where meta-induction is developed under the designation of "online-learning under expert advice" (OLEA). In particular, prediction games and corresponding optimality results have been generalized in three respects:

1.) *Probabilistic* (or Bayesian) *prediction games*: In a probabilistic prediction game (Schurz 2019, Sect. 7.1; Sterkenburg 2020), the candidate methods predict probability distributions over a finite space of event values. Methods are identified with probability distributions P_i and the prediction $pred_{n+1}(P_i)$ of each method $P_i \in \mathbf{C}$ delivered in round n is a probability distribution over the possible values $v \in Val$ of the next event,

$$P_{i,n}\left(e_{n+1} = v | e_1, \ldots, e_n\right)$$

conditional on the actual events in the past and possibly on further method-specific evidence that is left implicit. Formally, probabilistic games are a species of real-valued games. The deviation of the predicted probabilities from the event value that has been actually realized is scored by a *proper* scoring function such as the quadratic loss. It is well-known that under proper scoring, forecasters will maximize their average score if they predict the objectively correct probabilities. Attractivity-based meta-induction predicts the weighted average of these probability distributions

$$\mathrm{P_{aMI,n}}\,(\mathrm{e_{n+1}}=\mathrm{v}|\mathrm{e_1},\ldots,\mathrm{e_n}) = \sum_{1\le \mathrm{i}\le \mathrm{m}} \mathrm{w_n}\,(\mathrm{P_i})\cdot \mathrm{P_{i,n}}(\mathrm{e_{n+1}}=\mathrm{v}|\mathrm{e_1},\ldots,\mathrm{e_n}),$$

with weights defined as for the ordinary aMI.

One advantage of meta-inductive probability aggregation over standard Bayesian learning by conditionalization lies in the fact that is not restricted to a particular class of prior distributions, but permits the inclusion of any prior distribution in one's candidate pool, even of induction-hostile priors such as the mentioned state-uniform distribution. Yet, by applying the optimality result (2) meta-induction grants an optimal probability distribution over the next event(s) conditional on evidence about past events. A second advantage is that the dynamics of meta-inductively aggregated probabilities are more flexible than Bayesian conditionalization, since the weights of the candidate distributions are updated during the game.

2.) *Discrete prediction games* (Schurz 2019, Sect. 6.7): In discrete games, predicting mixtures of event values is impossible or forbidden, thus $\mathrm{Val_{pred}} = \mathrm{Val}$. For example, if events are binary the forecaster must decide to predict either 1 or 0. Two ways of generalizing the optimality result (2) to discrete games have been proposed in the literature:

2.a) *Randomization* (raMI): This method chooses between predictions of possible event values according to a probability distribution, which predicts each event value v (at the given time) with a probability that equals the sum of the normalized weights of all candidate methods predicting value v. The optimality result for raMI in discrete prediction games holds not only for convex but even for all loss functions, but at the cost that it applies not to the actual but to the probabilistically *expected* success rate of raMI, with the same short-run regret bound as for aMI (Cesa-Bianchi and Lugosi 2006, Sect. 4.1-4.2). The epistemological disadvantage of the optimality result for raMI is that it presupposes non-adversarial environments, in the sense that the predicted events have to be probabilistically independent from raMI's random choice of predictions. For a solution of Hume's problem this restriction is unacceptable. Schurz (2008, Sect. 8) develops an alternative solution that works without this restriction, namely:

2.b) *Collective meta-induction* (caMI): Here a collective of k meta-inductivists, $\mathrm{caMI_1}$, …, $\mathrm{caMI_k}$, approximates the predictive probabilities of caMI by the frequencies of their discrete predictions as close as possible. The optimality result for caMI holds approximately for the actual *average* success rate of the $\mathrm{caMI_i}$'s[2]; this result is universal and holds for all loss functions, and even in adversarial environments.

3.) Another important generalization is the extension of prediction games to sets of candidate methods that may *grow unboundedly* in time. A universal optimality theorem has been proved for unboundedly growing sets of methods under the mild restriction that

[2] There is an addition regret of $\frac{1}{2\cdot k}$ due to the *rounding* of raMI's ideal probabilities by finite frequencies that can be driven towards zero by increasing the number k of caMI's.

the number of accessible methods, m(n), grows less than exponential in time, or formally, $\lim_{n\to\infty} m(n)/e^n = 0$ (Schurz 2019, theorem 7.3).

4.) A further straightforward generalization is that from prediction games to *action games*—which in the machine learning literature are known under the category of "multi-armed bandits". Since formally a choice of an action can be identified with the prediction that this action will be most successful among the available options, the optimality theorem for predictions applies equally to meta-induction over action games (cf. Ortner in *this volume*; Schurz 2019, Sect. 7.5).

Considering meta-induction as a research program for forecasting sciences brings us to the contribution of Ortner in this volume, who presents several important variants of aMI that are not mentioned in Schurz (2019). Note that Ortner uses "T" instead of our "n" for the number of times or rounds and "K" instead of our "m" for the number of candidate methods. In Sect. 3.2 Ortner introduces a third possibility of transferring aMI to discrete prediction games with binary events, one that avoids randomization or collective meta-induction—namely *weighted majority* meta-induction, which I abbreviate as mMI. The worst-case regret bounds of mMI are as good as that for aMI (cf. Shalev-Shwartz and Ben-David 2014, Sect. 21.2.1). The strategy mMI has the disadvantage that its optimality result is restricted to the natural (absolute) loss function, while the optimality result for the randomized or collective aMI holds for all loss functions; the *hedge* algorithm that generalizes mMI to arbitrary loss functions uses again randomization (cf. Ortner *this volume*, Sect. 3.3.2).

In Sect. 4 Ortner turns to action games (multi-armed bandits) with limited information about success records. To handle them the meta-inductive strategy has to be modified by a success-independent weight component, which yields an optimality theorem with slightly worsened short-run regrets. An interesting idea are the *best-of-both-worlds* strategies that Ortner presents in Sect. 4.4. This idea picks up a problem that is also discussed in Schurz (2019, 162). As mentioned, there are several long-run optimal versions of attractivity-weighted MI; the version with the best *worst-case* short-run regret bounds is aMI, whence aMI is the uniquely recommendable MI method that one should apply on top of one's candidate pool. But there are other MI methods that have improved short-run performance in particular environments. For example, for regular event sequences the short-run regret of MI with linear weights is slightly smaller than that of aMI, at the cost of higher short-run regrets for irregular event sequences (Schurz and Thorn 2022, Sects. 5-6). Moreover, if we know that we are in a particular type of environment E, e.g. in an IID environment (an identical independent distribution), then we can know that a particular object-inductive method M_E (for example Laplace's rule of induction) will achieve highest success among all object-level methods. In this case we should use method M_E from the start and don't need meta-induction—although monitoring alternative methods is nevertheless wise since we can never be sure about the future. In Schurz and Thorn (2016, 52) the following *division of epistemic labor* is proposed to handle these complications: If a forecaster X has the justified belief that the (future) environment in which her predictive targets lie is of type E, and that in E a particular method M_E is optimal, then X should nevertheless use aMI on top of her candidate pool but include M_E into her candidate pool. This guarantees that if the future environment is indeed of type E, aMI's success will quickly converge against M_E's success, while if X's conjecture about the future environment is false her predictions will still be optimal w.r.t. her candidate pool. Ortner (Sect. 4.4) presents an alternative way to handle this situation: Given a generally optimal 'general-purpose' method M_g over a given candidate pool **C** (i.e., M_g plays the role of aMI) and a method M_E that yields better predictions than M_g over **C** in environments of type E, one can construct a *best-of-both-world algorithm* B (M_g, M_E), that behaves *almost* as good as M_g in arbitrary environments and almost as good as M_E in environments of type E (Ortner *this volume*, theorem 8). Observe that the above-mentioned strategy of putting M_E into aMI's candidate pool leads to a very similar result: The method aMI applied to the extended candidate pool $\mathbf{C}\cup\{M_E\}$ has an almost as good universal optimality result as aMI applied to **C** (since the number m is increased by 1), while in environments of type E, aMI will perform almost as good as M_E (for not too early times n). A closer comparison of the (dis)advantages of the two strategies of combining general-purpose optimality with local optimality constitutes an important future research question.

In Sect. 5 Ortner deals with "changing environments". The heading is somewhat misleading since, as emphasized by Ortner himself, already the standard version of aMI handles all sorts of changing environments. What the section is about is meta-induction over strategies that may switch between several base methods, which in Ortner's case are possible actions, represented by the arms of a multi-armed bandit. Ortner presents an MI method that achieves an absolute regret bound of the order of magnitude of $(s{\cdot}m{\cdot}n{\cdot}\log(m{\cdot}n))^{-0.5}$ in this setting, thus a vanishing regret rate of $(s{\cdot}m{\cdot}\log(m{\cdot}n)/n)^{-0.5}$, where s is the maximal number of allowed switches. The corresponding setting for sequences of prediction methods is briefly described in Schurz (2019, Sect. 7.3.2). The optimality of a variant of aMI in regard to all possible sequences of base methods whose number of switches grows sublinearly with the number of rounds offers a meta-inductive solution to Goodman's problem (cf. the next section).

2 From Justifying Meta-Induction to Justifying Object-Induction—or Lessons from Shogenji

In his interesting article Shogenji examines the second part of my proposed solution to Hume's problem—the transition from the a priori justification of meta-induction (MI) to the a posteriori justification of object-induction (OI), based on the superior track record of OI. With "induction" Shogenji means always common-sense or scientific object-induction and we follow his convention whenever we use "induction" without specification. Shogenji confronts the second part of my proposal with a twofold challenge, that can be summarized as follows:

Challenge 1: There is a tension between the justification of meta-induction and that of ordinary induction, because 1a) ordinary induction is unlike meta-induction and 1b) my justification of ordinary induction via meta-induction becomes difficult in the light of my skeptical arguments against the justifiability of the reliability of induction, in particular in view of certain skeptical arguments relying (i) on Goodman's problem and (ii) on the problem of induction-hostile prior distributions.

Challenge 2: Because of challenge 1b) the justification of OI by MI and OI's track record requires additional arguments that defeat the skeptical arguments of type (i) and (ii).

Shogenji proposes two such arguments, but then he argues that with help of these arguments a direct justification of OI would be possible without the recourse to MI.

Briefly summarized, my defense against these challenges will consist in showing that (i) the justification of OI by MI does not need the additional arguments proposed by Shogenji and (ii) Shogenji's proposed arguments that attempt to give a direct justification of OI do not work.

Let me first turn to Shogenji's challenge 1a). As Shogenji correctly observes, MI is not simply induction applied at the meta-level of the success rates of methods; it uses these success rates as weights but then aggregates the method's predictions into a prediction of its own. Shogenji (*this volume*, Sect. 2) writes that MI does not fit with our ordinary inductive practice, because we don't form weighted averages of predictions. I think this claim of Shogenji is only correct for the most simple inductive scenarios (e.g., strict induction over the color of ravens). In many other fields of ordinary and scientific induction we *do* form averages over the predictions of different methods or 'experts'. In common-sense reasoning, meta-induction is realized as an important form of *social learning* from experts or authorities; if the opinions of these authorities (e.g., two medical doctors) diverge, ordinary people will often intuitively weigh their opinions. In science, meta-inductive methods are realized in different ways; one important realization are methods of *meta-analysis*, in which the results of many studies are transformed into a common scale and aggregated into a general result.

In conclusion, meta-induction is not as far away from our inductive practice as Shogenji insinuates; but he is certainly right that meta-induction is different from pure object-induction applied at the level of events. This brings us to Shogenji's challenge 1b), according to which the justification of OI in terms of MI plus OI's superior track record "does not seem very strong if the arguments cited by Schurz (or their variants) against the reliability of induction hold up" (Shogenji in this volume, Sect. 2, 3rd §). This is a misunderstanding: in Schurz (2019, Chs. 2 and 4) I argue merely against the possibility of an *a priori justification* of the reliability of induction. But since the justification of OI is a posteriori, these arguments do *not* pertain to my justification of OI. Nevertheless, Shogenji then puts forward two objections that indeed pertain to the justification of OI, insofar they seem to undermine the thesis of OI's superior track record.

The first counterargument of Shogenji invokes *Goodman-type* methods. In analogy to Goodman's predicate "grue" = "green before a future time point t and blue afterwards", a Goodman-type method G_t predicts inductively until an arbitrary chosen future time point t, after which it starts to predict anti-inductively. As Shogenji points out, given the future time point t equals t_{n+1}, the track record of the Goodman-method G_{n+1} is just as good as that of OI but G_{n+1} makes the opposite prediction as OI. Since one *could* have added the non-inductive method G_{n+1} to the candidate pool, one cannot say that OI has a better track record than *any* conceivable non-inductive method.

My reply is twofold. First, my claim about the superior track record of OI that Shogenji quotes in his Sect. 2 is about prediction methods that people have actually used in the past (such as religious divination, soothsaying or astrology), so that some sort of track record is available. For Goodman-methods of the past (whose switch point may lie before the actual present) no track record is actually available; these methods are intuitively so strange that nobody would use them; nobody would even conceive them as 'one' method but rather as an unmotivated switch between two methods. Nevertheless, the problem of Goodman-methods is a theoretically serious challenge, since it is at least *possible* to include them into one's candidate set and to calculate their success records post-facto. Schurz (2019) discusses Goodman's problem in several sections (Sect. 1.2, 4.1, 4.2) and offers a solution to this problem in Sect. 9.1.5, that is not discussed by Shogenji. Instead, Shogenji develops his own proposed solution which goes as follows. Assuming a strict uniformity (e.g. all emeralds are green), then among all Goodman-methods G_t (with $t>t_n$ = the present time), there is only one method whose prediction differs from the prediction of OI, namely G_{n+1}, while all others G_t's (with $t>n+1$) predict as OI. It would be arbitrary, says Shogenji, to include merely G_{n+1} in the pool; if one includes one Goodman-method then one should include all or at least many Goodman-methods in **C**. But then, almost all methods in the pool will predict like OI, which means that the weight that aMI assigns to OI's prediction is high and induction is saved.

This is an interesting idea, but unfortunately it does not work, because it is restricted to Goodman-methods with merely *one* switch-point. For sake of generality one has to consider Goodman-methods with arbitrary many (say k) switch points, switching k (future) times between arbitrarily chosen (inductive, anti-inductive or other) methods. The generalized Goodman problem with multiple switch-points has been examined at the level of hypotheses by Steel (2009, 476 f.) and at the level of prediction methods by Schurz (2019, Sect. 9.2.5). If events are binary, then among all concatenated methods with arbitrary many future switch points there are as many methods that predict the next event correctly as there are methods that predict it incorrectly (by the symmetry of binary branchings); so Shogenji's proposal breaks down.

A different solution is proposed in Schurz (2019, memo (9.5)). By the finiteness of the cognitive resources of human beings, the candidate pool of methods is finite, but it is allowed to grow in time and may also contain Goodman-type methods, since the optimality theorem for aMI holds for all kind of methods. But there is an important additional problem and result. Given a set of qualitative *base* methods B, the number of Goodman-methods that piece together k base methods (with arbitrarily chosen future switch points $n+t_1, \ldots, n+t_k$, for $t_1<\ldots<t_k$) grows exponentially with k, and since k should be allowed to grow with time n (k=k (n)), tracking the success records of all Goodman-methods would soon become unfeasible. It turns out, however, that it is enough to track the success records of the base methods in order to achieve optimality in respect to the extension of B by all Goodman-methods that concatenate these base methods, provided the number k (n) of switch points grows sublinearly with n (Schurz 2019, 271, (9.5)).

Shogenji does not discuss this solution proposal. Rather, he argues that it is possible to justify a preference of OI over Goodman-type prediction methods, without the recourse to MI. For this purpose, Shogenji (*this volume*, fn. 4) employs a variant of a curve fitting criterion such as Akaike's information criterion. In the domain of curve-fitting with variable (e.g. polynomial) functions containing parameters that are freely adjustable to the data, it is an established fact that the expected accuracy of a fitted function (in regard to arbitrary new data) decreases with the number of parameters that have been freely adjusted to fit the actual data. Now, Shogenji applies this idea to Goodman-type methods and argues that the generalized Goodman-method G_x with a *variable* future switch point x contains the freely adaptable parameter x, which is one free parameter more than the object-inductive straight rule (OI).

First of all it should be noted that curve fitting criteria make inductive assumptions (e.g. a fixed dispersion with Gaussian error distribution; cf. Burnham and Anderson 2002, 63). But my major objection is that Shogenji's proposal seems to rest on an inappropriate application of Akaike's (or other) curve fitting criteria, for the following reason: The assumption of these criteria is that the free parameters are *adjustable* to the data, i.e. by varying their values the actual data are approximated better or worse. But this is not the case for the future switch points of Goodman-methods. OI and Goodman-methods fit the actual data precisely equally well (independently of which future time point is inserted for the parameter "x"); therefore the curve-fitting criteria do not apply. One may see this also by the following argument: If one fixes the variable time x in G_x to an arbitrary particular value t (> t_n), then the 'fixed' Goodman-method G_t has no more freely adjustable parameter. So by applying Akaike's information criterion the expected success of G_t comes out as exactly equal to that of OI; thus the application of Akaike's curve-fitting criteria breaks down for all Goodman-methods with fixed switch points.

We now turn to the second counterargument of Shogenji against the justification of OI by MI and OI's superior track record. This second argument (beginning with Sect. 4) pertains to prediction methods that are based on *state-uniform* probability distributions, i.e. on uniform distributions over 'state-descriptions' of the considered worlds, implemented as infinite sequences of (supposedly binary) events. Shogenji realizes the problem posed by state-uniform distributions when he writes that "in the absence of any empirical knowledge [...] there is nothing obviously wrong to assign the same probability to all state descriptions"—and yet this probability assignment makes induction impossible, as explained in the beginning of Sect. 1. Shogenji translates this challenge into the context of meta-induction by considering every state-description as a 'constant' prediction method (i.e., a method that predicts a particular event sequence irrespective of past observations). Shogenji proposes to solve this problem by considering the uncountable class of all constant methods as *one* method whose average prediction is always 1/2 (since at any time n there are as many constant methods predicting 1 as there are constant methods predicting 0). But since in induction-friendly environments, OI is predictively much more successful than constantly predicting 1/2, Shogenji thinks that the challenge is refuted.

This is a nice idea, but I cannot see how it works. First of all, there are uncountably (2^{∞}) many of these 'methods'. Putting all of them into the candidate pool would exceed the finite bounds of humans' cognitive resources.[3] Only a finite fraction of them can be simultaneously accessed and tracked. But if there is only a finite number of them in the pool, Shogenji's proposed solution no longer works. However, meta-induction provides another solution for this case: if each state description S_i has the same probability to be included in the pool, then since the S_i's cannot learn from the past, it will be overwhelmingly probable that all S_i's included in the pool have a by far worse success record than OI. Of course it is not excluded—be it by 'magic' or accident—that a chosen state description S* always predicts correctly, in which case S* would be indistinguishable from a perfect clairvoyant and MI would assign to S* the highest weight.

In the final paragraph of Sect. 4, Shogenji gives the impression as if OI would be more successful than the constant 1/2-forecaster in *every* possible environment. This is not the case, as there are various binary sequences for which OI's success rate is 0, while the suc-

[3] Note, however, that there is a meta-inductive optimality result for infinite candidate pools and players with infinite cognitive resources (Schurz 2019, 265, theorem 9.2).

cess rate of the constant 1/2-forecaster is always 1/2 (Schurz 2019, 19 f.; Schurz and Thorn 2022, Sects. 5–6). For the same reason Shogenji's claim that this argument provides an independent non-circular justification of OI's reliability without the help of MI breaks down. In his Sect. 5, Shogenji argues, based on a paper related to machine learning, that prior distributions should not be considered as epistemic but as objective distributions that can be revised in the light of evidence. In Sect. 4 we shall argue that this view rests on a confusion. For related reasons I do not share Shogenji's intuition about the asymmetry between the two events E_1: "I win a lottery with 1024 tickets" and E_2: "I toss with a coin 10 heads in a row", although a detailed analysis would go beyond the scope of this paper.

3 Meta-Induction in Application to Prophecy and Religious Worldviews—or Lessons from Pitts

In his exciting contribution to this volume, Pitts also challenges the second part of the meta-inductive account, the justification of OI by MI and OI's superior track record. But Pitts' challenge is very different from that of Shogenji. It is not concerned with 'virtual' methods alternative to scientific induction that have been invented by analytic philosophers, but with the most widespread examples of real-life alternatives to science: strategies of prophecy and clairvoyance, that make up a major part of the religious narratives and are defended by Pitts from a religion-friendly viewpoint. Pitts starts his paper with a brief reconstruction of recent attempts of justifying induction and their failures. He agrees with my explanation of why a justification of induction is needed. Against those who consider induction as self-evident, Schurz (2019, 16) points out that "Millions of people do in fact believe in superior non-inductive methods, be it God-guided inner intuition, clairvoyance, or other supernatural abilities". Pitts sympathizes with this attitude, but comments on my passage in a slightly polemically way by saying: "Schurz appears actually to understate the case by a couple of orders of magnitude. By some counts more than half the world's population is Christian, Muslim, or Hindu." Let me reply that this is a slight distortion of (i) words and (ii) facts, since (i) speaking of "millions" is just a metaphor for a very high number (not a numerical count) and (ii) for many people their membership in a religious confession does not at all imply that they literally believe in the doctrines of this confession (cf. Inglehart and Norris 2003, 55; Table 3). Only a very small percentage of Christian people do really believe in the miraculous events reported in the Bible that are so important for Pitts.

Pitts then focuses on one key aspect of the optimality justification of MI, namely its radical *openness* towards epistemic possibilities (Schurz 2019, 203 f.). From an a priori viewpoint, an esoteric prediction method such as clairvoyance or God-guided divination can also be meta-inductively justified, if it would be more successful than ordinary or scientific object-induction, henceforth simply called "induction". While I agree with Pitts' observations about the non-dogmatism of meta-induction, I doubt that any of the religious prediction methods to which Pitts refers was ever significantly successful; my criticism of Pitts' considerations will be confined to this point. For illustration, in his Sect. 1 (11th §) Pitts writes that:

> "If induction is not known in advance to be reliable … then miracle-like events are not known in advance to be impossible or even highly improbable: dead people might cease being dead, for example."

This is correct, but his conclusion is by far too quick—namely (Pitts in this volume, Sect. 1, 11th §):

> "It is therefore unclear on what grounds one can exclude miracle reports from history, at least as data potentially to be taken seriously, as rationalists in philosophy and theology have been doing since Spinoza, Hume (in his critique of miracles) and Schleiermacher."

I want to reply that this is *not* 'therefore' unclear, because, as I will argue, miracle reports from history do *not* count as facts in the sense understood by Pitts, namely as facts that confirm miracles. Moreover, Hume was not a rationalist but an empiricist, and he did not exclude miracles on a priori grounds, but based on an a posteriori argument that weighs the (im)probability of a purported miracle against the (im)probability of an erroneous testimony (see below). It is true that there were rationalist or idealist philosophers such as Spinoza or Bradley who held the view that miracles are impossible a priori; so Pitts has his point (cf. McGrew 2019, Sect. 3.1). But an open-minded empiricist and in particular a meta-inductivist would reject this view as just another sort of dogmatism.

Pitts' central claim (Sect. 4, 3rd last §) is that religious people believe occasionally in non-inductive methods—e.g. when they trust biblical reports about miracles (such as resurrection, multiplication of bread or healing from death)—but when they do this they do it on *meta-inductive* grounds, relying on the purported success of these non-inductive methods. Pitts gives various historical reports for his claim. For many of his examples he seems to be right and the demonstration of this fact is one of the merits of his paper. But there are also other passed-on examples where religious beliefs are not based on meta-inductive reasoning but on pure obedience to God's authority, such as the case of Abraham's obedience to God's command to sacrifice his son.

However that may be—what almost always distinguishes religious arguments from that of empirical science is a circumstance that is *orthogonal* to the induction problem, namely that they are typically based on *pseudo-evidence* as opposed to proper evidence. Religious reports about miracles typically rely on merely purported evidence whose content consists in 'testimony' based on imagination, hearsay or wishful thinking, but not on evidence meeting the objectivity standards of scientific observations. In Sect. 3 (2nd §) Pitts gets to this point. He correctly criticizes dogmatic historians that 'reconstruct' or select historical facts in the light of theories that are then claimed to be confirmed by these historical facts. Discarding testimonial reports as non-factive because they don't fit with one's world-view is of course circular. Precisely for this reason any empirical and in particular any meta-inductive method needs a level of *theory-neutral* or method-neutral observations. The theory-neutrality of observations has been defended in Schurz (2014, 74, def. 2.9-1; 2019, 56, (4-4)), based on the criterion of intersubjective ostensive learnability. Only sense experiences can satisfy the criterion of ostensive learnability, not inner experiences based on intuition or imagination.

Meta-induction assumes that past observations can be reliably recorded (Schurz 2019, 198). Reliable observations are required to satisfy several (scientific) standards of objectivity; the two most important ones for our topic are these: (i) They must be described in an intersubjectively shared observational language, and (ii) they must have been recorded by several *mutually independent* observers or empirical sources, as many as possible. The reason for requirement (ii) is that each observer or empirical source is error-prone, but provided the reliabilities are at least greater than random success, the mutual agreement of many (n) conditionally independent observational reports increases the conditional probability of the hypothesis and raises it to certainty for $n \rightarrow \infty$ (which is a variant of the famous Condorcet jury theorem; cf. Bovens and Hartmann 2003, 62; Schurz 2022b, 9, theorem 1). Two cases can be distinguished:

1.) Usually the reported type of fact can be reproduced provided it is true. In this case the reported facts should have been reproduced in independent experiments (if possible double-blinded; Schurz 2014, 202).
2.) Sometimes reported facts cannot be reproduced. In fact, reports of paranormal or supernatural events are typically alleged to be non-reproducible—for example, because God sent his son only *once* to humanity. This should already make one doubtful; but let us assume we accept the non-reproducibility. The more important it is in these cases that there are several mutually *independent* testifiers or testimonial sources that report this fact. But I do not know of any historical report about miracles or paranormal events that is convincingly testified by independent sources. Rather these 'reports' have resulted from a chain or tree of story-writers that have more-or-less rewritten the stories of their predecessors, similarly as in the development of rumors.

The view that testimony is reliable per se is unjustified; one needs positive reasons that justify the reliability of a purported informant (cf. Schurz 2019, 285). What thereby counts as observable evidence is the fact that a proposition p has been *reported* (by some source), abbreviated as R(p) and the reliability of this report is given as the probability that p is true given it has been reported, i.e. $P(p|R(p))$. A report is minimally reliable if this probability is greater than random success, in the binary case greater than 1/2. While reports of observational events by ordinary people (e.g., that it starts raining) are usually highly reliable, common-sense reports about future happenings (e.g. the climate change) are rather low, and especially reports about miracles seem to be unreliable in most cases. Among other arguments, this follows already from the fact that most of the religious stories handed down by mankind have been found to be false by scientific evidence (I dispense with presenting here a long list). I am saying "most" but not "all" because several of these stories are empirically irrefutable; but "most" is sufficient to infer that the statistical frequency that a report R(p) tells the truth given that p describes a religiously motivated miracle is rather low.

Pitts is much more than me inclined to believe in the biblical stories. He senses my observation that until today scientific induction has been by far more successful than religious prophecy as a big "leap" (Sect. 3, 6th §), but I think by the arguments given above my points should be obvious for everyone who is not a science-denier. I do not deny that several religious people believe the stories told by their religion and trust them even more than scientific forecasts, but I think this is usually the case because these people lack scientific education and their beliefs are based on wishful thinking; cognitive psychology is full of studies confirming the manifold cognitive biases resulting from wishful thinking (cf. Piattelli-Palmarini 1994). If some of the Christian miracle reports would really be scientifically confirmed, this would *not* be “something of a disaster for the scientific rationality”, as Pitts writes (2nd half of Sect. 3), but rather, it would be a new and highly important scientific hypothesis. For example, if one could really multiply one piece of bread into bread for 5000 people (as Jesus purportedly did), then all major hunger crises could be solved—but unfortunately no miracle report has ever been reproducible and thus never led to any verifiable beneficial consequences, except the comforting Placebo effects of wishful thinking (cf. Schurz 2022c, 101). In conclusion, Pitts is right that super-natural powers cannot be excluded a priori, but wrong in saying that MI would “beg the question” against super-natural powers (Pitts, *this volume*, end of Sect. 3)—MI does not beg the question, because the rejection of super-natural powers is based on their poor empirical track record.

Likewise, the major argument of the empiricist philosopher David Hume is not based on a priori reasoning. The quote of Hume given by Pitts (end of Sect. 3) is distorting; central for Hume’s view is his “balance of probabilities” argument (Hume 1748/2000, 87–88):

> “When anyone tells me, that he saw a dead man restored to life, I immediately consider with myself, whether it be more probable, that this person should either deceive or be deceived, or that the fact, which he relates, should really have happened. I weigh the one miracle against the other; and according to the superiority … I pronounce my decision, and always reject the greater miracle. If the falsehood of his testimony would be more miraculous, than the event which he relates; then, and not till then, can he pretend to command my belief or opinion”.

Admittedly there are other quotes in which Hume overestimates the conclusiveness of OI—contrary to his theoretical skepticism, which is a well-known tension in Hume. Sometimes he seems to say that no evidence whatsoever can confirm a miracle, but the most common interpretation of Hume is the one fitting with the above quote (cf. McGrew 2019, Sect. 3.1.2). According to this interpretation of Hume, if a miraculous event (e.g., someone walking on water) would really be reported by sufficiently many independent witnesses, this would overthrow the conclusion of event-based induction that speaks against the miracle. But no such case is known; all that one has are narratives and legends perpetuated in oral history and folk writing.

At the end of Sect. 8 Pitts writes that since Jesus’ miraculous deeds are more than 2000 of years away from us, it would be too much to demand the historical documentation of many independent witnesses. But independent from (non-)reproducibility, there is another strong argument against the reliability of miracle reports—namely, that in the history of mankind there have been thousands of different religions (Wallace 1966 estimated them as 100,000), all reporting different sorts of super-natural powers or Gods bringing about miracles. They cannot all be true because they are in mutual competition and contradiction. This argument is likewise found in Hume and is quoted by Pitts in his Sect. 9:

> “it is impossible that the religions of ancient ROME, of TURKEY, of SIAM, and of CHINA should, all of them, be established on any solid foundation. Every miracle, therefore, ... has it the same force, ... to overthrow every other system ... it likewise destroys the credit of those miracles on which that system was established” (Hume 1748/2000, 90).

We can infer from this argument that the truth chance of a report about a religious or supernatural miracle is low, independent of its content; the mere condition that it has a mystic or religious source makes it improbable.

This concludes my major argument against Pitts' attempt of justifying biblical legends by meta-induction applied to their track record. In the final part of this section we have a closer look on Pitts' concrete examples in Sect. 4 ff. In the beginning of his Sect. 4, Pitts presents evidence for the "use of meta-induction in the Hebrew Bible to motivate apparently unreasonable (by ordinary standards) behaviors on religious grounds". Like Pitts we assume that "whether these stories or anything like them actually happened, can be set aside for now". As already mentioned, Abraham's obedience to God's command to sacrifice his son Isaac doesn't seem to be explainable by meta-inductive reasoning but rather by unconditional submission to God's will, but the other examples do indeed seem to be instances of meta-inductive reasoning. While by ordinary induction from military experiences, the Israelites had only a small chance to conquer the promised land of Canaan, their success was predicted meta-inductively by Moses on the grounds of God's having successfully rescued the Israelites from slavery in Egypt, indicating the superior power of God's help, as Pitts tells us. When the majority of Israelites still reasoned object-inductively that presumably they could not conquer Canaan, God complained to Moses "How long will they [the Israelites] not believe in me, in spite of all the signs which I have wrought among them?“ (Numbers 14:11 RSV; cf. Pitts, 5th § of Sect. 4). Pitts rightly concludes that according to these stories, God recommended to his people that they should reason meta-inductively and, based on the observable track record of God, should weigh his predictions higher than that of ordinary induction. This does not mean that God himself is a meta-inductivist, as Pitts writes; God is rather an omniscient clairvoyant who doesn't need meta-induction, but he expects his cognitively more restricted people to reason meta-inductively. A similar diagnosis applies to the third example of Pitts about the prophet Hanani. Of course (as re-emphasized by Pitts himself), whether these legends are true is an entirely different question. What these examples indicate is that meta-induction is strongly *entrenched* in common sense reasoning, but they do not indicate that religious faith has a meta-inductive a posteriori justification.

Beginning with Sect. 5 Pitts speaks frequently about meta-induction as a "logic of prophecy". Understandably, Pitts intends to exploit meta-induction for religious purposes, but meta-induction is *definitely not* a "logic of prophecy". It is not a "logic" at all; this phrase is misleading, because "logic" suggests an a priori justification, while a meta-inductive justification of prophecy could only be established by its track record. Meta-induction is rather a most general epistemic method that is not only applicable to scientific hypotheses but also to beliefs or world-views of common-sense. In his Sect. 5 Pitts portrays King Croesus's test of the Greek oracles of Delphi as a genuine instance of a predictive test. If the reported results were indeed true (which is doubtful), this test would indeed constitute a meta-inductive confirmation of the success of these oracles. Pitts then turns to Cicero's distinction between *prophecy-related* (so-called 'natural') *divination* (messages from a God via prophets) and *technical divination* (observation of animal entrails, directions of bird flight, etc.). In the meta-induction account, these divination methods are classified as follows: While prophecy-related divination is meta-induction based on track records of 'experts', the classification of technical divination depends: (i) if technical divination is confirmed by induction form observed correlations, it is object-induction, (ii) if it is believed because a prophet says so, it is meta-induction, and (iii) if it is believed by mere superstition it is a non-inductive method. If Sambursky's report is correct that the Stoics denied an essential difference between scientific inference and technical divination (cf. Pitts *this volume*, end of Sect. 6), this speaks for option (i).

At the end of his Sect. 15 Pitts writes that given that meta-induction was used as an attempted justification of prophecy by the Stoics and implicitly by the Israelites, "it is too quick merely to assume or assert without investigation that experience [...] never vindicates prophecy". But as explained above, this question has already been investigated, since there are thousands of pieces of historical and systematic evidence confirming that religious or other mythical stories did not really happen but were based on fiction and wishful thinking. I think that the burden of proof here is on the side of the defenders of prophecy and miracles, who owe us *at least one* documentation of a successful prophecy or miracle that meets the two above-mentioned empiricist standards. I do not know of any single documentation that even remotely meets these standards; if there was one, empirical scientists would immediately jump on this documentation and turn it into a sensational scientific report. Although in some places Pitts concedes critical doubts about the truthfulness of biblical legends, in other places he seemingly assumes uncritically religious testimony as evidence (e.g., at the end of Sect. 8). Apart from these drawbacks Pitts' paper is an inspiring reflection of the use of meta-induction in domains outside of science.

4 Meta-Induction as a Solution to the No Free Lunch Theorem—or Lessons from Wolpert

Wolpert claims in the abstract and in Sect. 4 of his paper that my account would favor the induction-friendly frequency-uniform prior distribution. Let me start this section by emphasizing that this claim is wrong. On the contrary, in several passages in Schurz (2019) it is emphasized that meta-induction is not bound to any particular prior distribution (e.g. on pages 71 f., 167, 240-244). Rather, what I object to Wolpert's no free lunch (NFL) theorem is that this theorem rests on a particular prior, namely the induction-hostile state-uniform prior. Although the justification of meta-induction works even for the state-uniform prior, this justification becomes much stronger if one allows for different possible priors that are evaluated and aggregated by probabilistic meta-induction, including induction-friendly as well as induction-hostile priors. But nowhere in my book do I express a preference for frequency-uniform priors and I wonder how Wolpert came to this misunderstanding.

Wolpert defends his account against my objection that the NFL theorem for predictions depends on a state-uniform prior, by presenting versions of this theorem that apparently do not assume a state-uniform prior. The goal of this section is to demonstrate that in fact these versions *do* assume a state-uniform prior, at least implicitly, by the consideration of (unweighted) sums or averages over all possibilities.

Wolpert's paper starts with a nice introduction presenting a game-strategy devised by Parrondo as an early example of a strategy of meta-induction, or online learning under expert advice (OLEA), as it is called in machine learning. In Parrondo's setting, methods are represented by sequences of bits of their payoffs, and a simplified version of Parrondo's strategy, call it P, imitates the prediction (or action) of the method that has highest cumulated payoff. Obviously, P is a version of ITB. Wolpert explains why P is a good strategy, but it should be added that ITB is not universally optimal.

In Sect. 2 Wolpert turns to the NFL theorems. They apply only to prediction methods that are *non-clairvoyant*, in the sense that the total information about the past events and success rates screens off the next event from its prediction—which is Eq. (3) in Sect. 2 of Wolpert's paper. In Sect. 2 (below Eq. (5)) Wolpert presents two versions of NFL theorems that are only inessentially different. Both versions compare the sum or average of the loss or cost of prediction methods *over all possible event sequences* (or states of the world) f, with the result that this cost sum or average cost is the same for all methods. There is a second and more important distinction, that between a strong and a weak variant of the NFL theorem. The *strong variant* of the NFL theorems is presented by Wolpert. This variant presupposes a *homogeneous* loss function in the sense of Wolpert (1996, 1349)—which is arguably a too strong condition on loss functions—while the weak NFL theorem assumes a merely weakly homogeneous loss function (see below).

Let C be the set of all possible losses resp. "one-shot" costs c, i.e. the possible differences between a prediction and an event (formally $C = \{c: \exists pred \in Val_{pred} \exists e \in Val: c = loss(pred,e)\}$). The strong variant of the NFL theorem (in both of Wolpert's versions) applies to each possible cost value $c \in C$ and asserts, in simplified worlds, that the probability of having loss c averaged over all environments is the same for all non-clairvoyant methods. More precisely, version 1 of Wolpert's NFL theorem asserts that for all $c \in C$, the sum of the probabilities of a method's attaining cost c in world state f, summed over all possible f's (conditional on data of size m) is the same for all methods (note that Wolpert's variable C_{OTS} ranges over these possible c's).[4] Wolpert's version 2 asserts that for all $c \in C$, the probability of a method attaining cost c in world state f (conditional on a data sequence d) is the same for all methods, given a state-uniform probability distribution P(f) over the f's. Now, Wolpert says that a

> "secondary implication of the NFL theorems is that if it so happens that you assume/believe that P(f) is uniform, then the average over f's used in the NFL for search theorem [=version 1, G.S.] is the same as P(f) in version 2".

I don't think this implication is "secondary" because *summing up* the probabilities of attaining cost c in f over all f's is essentially the same as averaging over these probabilities (since dividing their sum by their number gives the average) which is in turn essentially the same as calculating the overall probability of attaining cost c by a uniform prior distribution over the f's (since the average of these probabilities over all f's equals their expected probability according to a state-uniform prior over the f's).

The condition of *homogeneity* requires that for *every possible loss value* $c \in C$, the number of possible event values $e \in Val$ for which a given prediction pred leads to a loss of c is the same for all possible predictions $pred \in Val_{pred}$. Homogeneity is satisfied only for prediction games with a zero-one loss function, which gives a maximal loss of one if the prediction differs from the event and a zero-loss if the prediction equals the event (cf. Schurz 2019, 326, def. 9.1). Obviously homogeneous loss functions are unreasonable whenever predictions and/or events are graded. For example, the prediction "0.9" of the event "1" is better than the prediction "0.1" (since the distance between 0.9 and 1 is much smaller than that between 0.1 and 1), although for homogeneous loss functions both predictions are equally bad and attain a score of zero. Therefore Schurz (2019, 237, def. 9.2) and Schurz

[4] We ignore here Wolpert's probability $\pi(q)$ of choosing the predicted event q, because q is fixed.

and Thorn (2022) concentrate their investigation on weakly homogeneous loss functions, that are mentioned by Wolpert (1996) in a small paragraph on p. 1354 ("More generally, for an even broader set of loss functions …"). A loss function is *weakly homogeneous* if for each possible prediction pred, the *sum* (or average) of the losses over all possible events is the same. For binary games with real-valued predictions and absolute loss function, weak homogeneity is satisfied, since for every possible prediction pred ∈ [0,1], loss(pred,1) + loss (pred,0) = 1 – pred + pred = 1 (Schurz 2019, 327, def. 9.2).

The *weak variant* of the NFL theorem makes the corresponding assertion not for each cost value c∈C separately, but merely for the sum or average of all cost values. In version 1 the weak NFL theorem says that the average cost over all possible event sequences f (conditional on data size m), defined as $\Sigma_{f,c}P(c|f,m)\cdot c$, is the same for all methods, and in version 2 it says that the probabilistically expected cost of a method (conditional on a data sequence d), defined as $\Sigma_f P(f)\cdot\Sigma_c P(c|d,f)\cdot c$, is the same for all methods according to a state-uniform distribution P(f) over the f's. Finally, note that loss-functions for real-valued events do not even satisfy the condition of weak homogeneity and Wolpert's version of the NFL theorem does not hold for real-valued events; however, a weaker version of the NFL theorem applies to them (as proved in Schurz 2019, prop. 9.3).

We now turn to Wolpert's arguments against my diagnosis that the NFL theorem for predictions depends on a state-uniform prior. These arguments and my objections to them apply equally to the strong and the weak variant of Wolpert's NFL theorems. In his first argument Wolpert (4th § after equ. (5)) says that

> "it must be emphasized that simply allowing [the prior—G.S] P(f) to be non-uniform, by itself, does not invalidate the NFL theorems",

and some lines later he says that the

> "NFL theorems do not assume that the universe is governed by a uniform prior in some objective sense."

Here we meet an important confusion that is also found in other machine learning texts (for example also in the paper quoted in Shogenji's Sect. 5, as mentioned in my Sect. 4), namely the following: When epistemologists speak of a *prior* probability they mean always a *subjective-epistemic* probability, i.e. a rational degree of belief, but *not* an objective probability (be it a statistical propensity or an objective single case chance). A 'prior' probability is defined as a distribution that one adopts or should reasonably adopt, *prior to experience*; this notion *only makes sense* for an epistemic notion of probability, but not for an objective one, because objective probabilities are *independent* from whether the subject has experience or not. When machine learners speak of an "objective prior", they just mean the *true unconditional* probability function over the possible states of a type of system; but this is entirely different from a prior in the epistemic sense. For this reason, Wolpert's accusation in Sect. 4 (5th §) that "Schurz argues that one should adopt a single, specific prior … a uniform prior over frequencies" is *not only* incorrect because I never make any such assertion; in addition Wolpert's critique of this position—which is the position of Laplacean inductivists—is inappropriate because Wolpert assumes wrongly that the frequency-uniform prior is meant in the objective sense. Wolpert attempts to refute this misunderstood position by pointing out that "all of statistical physics is based on a uniform distribution over patterns, not over frequencies". Wolpert's misleading critique culminates in his devious diagnose in the last paragraph of his paper that

> "Schurz's proposal for a uniform prior over frequencies runs afoul of thousands (tens of thousands?) of previous experiments concerning the real, physical world".

This wrongs me twice: first because it is *not me* who assumes frequency-uniform distributions but Laplacean inductivists, and second I know quite well that distributions of micro-canonical ensembles in thermodynamics are not frequency-uniform, as Wolpert rightly observes, but his observation is *besides the point*, because the frequency-uniform distributions to which induction-friendly probability theorists refer are meant as epistemic and not as objective probabilities.

Having clarified this confusion, let us get to Wolpert's second major argument against my diagnosis that the NFL theorems are based on a state-uniform epistemic prior. Namely, Wolpert writes (in the 2nd half of his Sect. 2) that

> "allowing P(f)'s [i.e., the priors over event sequences—G.S.] to vary provides us with a new NFL theorem. In this new theorem, rather than compare the performance of two learning algorithms by uniformly averaging over all f's, we compare them by uniformly averaging over all P(f)'s".

As Wolpert continues, this uniform averaging results again in an NFL theorem (in both of his versions). This is no wonder—because a uniform average over all objective priors over the space of possible event sequences is just *a second order version of a uniform epistemic prior* that results in a uniform expected first order prior. For example, suppose that events are binary (0 or 1) and $p =_{def} p(1)$. Assuming a uniform (2nd order) prior density $D(p)$ over all possible (1st order) priors $p \in [0,1]$, the resulting expected 1st order probability of the event 1 is given as ${}_0\int^1 p \cdot D(p)\, dp = {}_0|^1 p^2/2 = 1/2$, which is uniform at the 1st order level.

Thus, Wolpert's proposed method of averaging over possible prior distributions is just another version of a state-uniform prior distribution. In conclusion, Wolpert's attempts to escape the diagnosis that the NFL theorems for prediction depend on a state-uniform prior do not work, and his claim in the 3rd-last § of Sect. 2 that this diagnosis is "simply wrong" seems to apply to itself.

Let us now briefly explain the solution to the challenge provided by the NFL theorems proposed by meta-induction. It follows from the dominance results for aMI (recall result (3) in Sect. 1) that aMI enjoys free lunches over all methods that it dominates. How can that be in view of the NFL theorems—is this not a contradiction? My answer distinguishes between the long run and the short run perspective. In both perspectives, the answer is no. In regard to the long run perspective, the contradiction is only apparent, because the state-uniform probability distribution that Wolpert assumes assigns a probability of zero to all worlds (infinite event sequences) in which aMI dominates the inferior methods (cf. Schurz 2019, 70 f., 241); so these worlds do not affect the probabilistic expectation value of the method's success. But although the state-uniform prior of worlds in which aMI meta-induction dominates inferior methods is zero, there are many—indeed uncountably many—such worlds and it is precisely in these worlds that intelligent prediction methods can have any chance at all. We should not exclude these induction-friendly worlds from the start by assigning a probability of zero to them, which means that we should not restrict the epistemic priors to uniform priors.

Within the short-run perspective, the defense of meta-induction against the NFL challenge is more difficult, because here aMI suffers a small regret. Here we argue as follows. What counts are two things: (a) To reach *high* success in those environments which *allow* for high success by their intrinsic properties (uniformities). This is what independent inductive methods do. (b) To *protect* oneself against high losses (compared to average success) in induction-hostile environments. This is what cautious methods do, such as the method "averaging" that always predicts the average of all possible event values. The advantage of aMI is that it combines *both* accomplishments—reaching high success rates whenever possible and avoiding high losses; a demonstration of this fact by computer simulations is found in Schurz and Thorn (2022, Sect. 5). In conclusion, aMI achieves 'the best of both worlds', although this comes at the cost of a small short-run regret of aMI that is acceptable given the mentioned advantages of aMI. In the case of discrete events with linear loss function, the NFL theorems imply that the state-uniform average of this short-run regret is the same for all methods; but the advantages (a) and (b) even hold under this induction-hostile assumption. For quadratic loss functions or more induction-friendly priors the short-run advantages of meta-induction get amplified (cf. Schurz and Thorn 2022, tables 3-8). Wolpert's notion of "head-to-head minimax distinctions" in his Sect. 4 comes close to my proposed solution for the short run: the maximal regret of the methods is minimal for aMI and yet aMI climbs to high successes in regular environments.

Finally a remark on Wolpert's nice construction of a competition between two meta-level algorithms in his Sect. 3—a meta-inductive method based on cross-validation, and a corresponding meta-anti-inductive method. Both meta-methods have access to the same candidate pool of methods; we abbreviate the two meta-level methods as MI and MAI. Schurz (2019, 93, 157) calls such competitions prediction *tournaments*, as opposed to prediction *games*, since in tournaments it is assumed that the preferred meta-inductive method cannot access the competing meta-methods. Wolpert observes that for every prior P(f) over event sequences for which MI performs well, there exists corresponding prior P*(f) for which AMI performs equally well. This is certainly correct, but it does not affect the optimality result, because it assumes that MAI is not accessible to the method MI, while the optimality theorem applies only to accessible methods. As soon as MI is allowed to access AMI's predictions MI's success is granted to converge to AMI's success in environments in which AMI is optimal.

5 The Problem of Induction for Probabilistic Frameworks—or Lessons from Williamson

In his Sect. 1 Williamson writes that "Schurz (2019, ch. 4) argues against probabilistic accounts of induction", but this misrepresents my position. What I attempt to show is that probabilistic accounts of induction do not help in *justifying* induction, because they themselves make inductive assumptions—such as countable additivity, non-dogmaticity, exchangeability, the principal principle and uniformity (Schurz 2019, 75, Sect. 4.7). I do not deny that probabilistic accounts are highly important for explicating inductive reasoning; on the contrary I am a friend of probabilistic accounts and emphasize their usefulness in several places—for example in Sect. 3.3 and 4.3-4, in which I present various probabilistic results, including results about the *principal principle*—PP for short—and the related *principle of the narrowest reference class*—PNRC for short—that are in the focus of Williamson's paper.

In conclusion, all that I doubt is that probabilistic accounts can justify principles of induction. Williamson himself doubts that a solution to Hume's problem is possible at all and considers the problem of justifying induction as "largely academic" (see his Sect. 1)—but in Sect. 3 about Pitts' contribution we have seen that this is not so, because the justifiability of induction is in the center of science-versus-religion debates. In any case, Williamson focuses his paper on his probabilistic explication of induction, while the problem of justifying induction seems to be 'charmed away' in his account. Let us see where this problem has gone.

Williamson starts the systematic part of his paper in Sect. 2 with objections against subjective Bayesianism and logical probability theory with which I largely agree. He then comes to his own preferred account that he calls an "empirically-based Bayesianism". This account rests on a version of the direct inference principle that connects subjective probabilities (degrees of belief) with objective probabilities—either with generic-frequentistic probabilities (Howson and Urbach 1996, 345) by means of the PNRC, or with single-case chances (Lewis 1980) by means of Lewis' PP. Frequentistic probabilities are usually conceived as *statistical* probabilities, which are by definition *frequency-limits* in potentially infinite sequences of realizations of a random experiment. In what follows I write a capital "P" for statistical probabilities (Williamson writes P* instead of P) and a capital B for degrees of belief. There is a big difference between finite frequencies and frequency limits: the statistical tendency of a fair coin to land on heads with $P=1/2$ is not definable by any of its finite frequencies. Unfortunately Williamson never distinguishes between frequencies and frequency-limits, which is one step of the "magic" that obscures the induction problem. But Williamson's magic goes farther: it is contained in his phrase that a piece of evidence E (i.e. a finite sample) "determines" a statistical probability P, in his explication of the PNRC at the beginning of Sect. 3 that goes as follows—where "Ac" stands for "individual c has property A", and E stands for the given total evidence:

Williamson's PNRC: $B_E(Ac)=x$ if E determines that (i) the frequency (limit) $P(A|R)$ equals x (some number between 0 and 1) and determines that (ii) R is the unique narrowest reference class containing c for which (the information about) $P(A|R)$ is available and (iii) E contains no information more pertinent to (the probability of) Ac than the information (i) and (ii).

Clarifications: 1) We write R for Williamson's narrowest reference class $\hat{\mathrm{p}}$, A for α, and the statistical probability in the conditionalized form; so we write P(A|R) instead of Williamson's $\mathrm{P^*}_{\hat{\mathrm{p}}}(\alpha)$.

2) The predicates A and R are monadic and should be read as furnished with an implicit individual variable x—thus $P(A|R)=P(Ax|Rx)$ expresses the statistical probability that an arbitrarily member x of (the extension of) R has property A. The PNRC is generalizable to

relational predicates as indicated in Schurz (2019, 68, fn. 4), but we focus here on monadic predicates.

3) We would prefer to write Williamson's epistemic probability $B_E(Ac)$ in the conditionalized form B(Ac|E), but we don't do this because Williamson advocates a "non-standard Bayesianism" in which the principle of strict conditionalization that equates $B_E(Ac)$ with B(Ac|E) (for a suitable epistemic prior function B) is given up. Williamson calls this (abandoned) principle *CBCP*, for conditional-belief-conditional-probability. By this step Williamson attempts to solve some apparent problems of the PNRC, but as we shall argue below the problems diagnosed by him result from an inappropriate understanding of clause (iii) of the PNRC. I call this clause the *no-defeat clause* because it requires that the evidence should not contain additional information that defeats the information in (i)+(ii). In a more adequate formulation of the PNRC the diagnosed problems disappear and the principle CBCP is saved.

After this technical clarification we come to our main observation: In Williamson's account the *problem of induction* is hidden behind the phrase "E determines that the frequency (limit) P(A|R) equals x", because this phrase expresses *nothing but* an inductive inference from a sample frequency—in Williamson's example in Sect. 3 "17 out of 100 21 years old males get a cough"—to the inductive-statistical generalization "the probability of getting a cough among 21 years old males is approximately 17%". What justifies this inductive inference? What justifies us in assuming that this small sample is representative for the population of all 21 years old males? Maybe in the near future the frequency of coughs will increase drastically because of the spread of new viruses, in which case all of our present samples become unrepresentative and we should better be skeptical about induction. *These* questions make up the problem of induction in Williamson's account, but Williamson never discusses these questions; he simply assumes without comment that a finite sample inductively "determines" the frequency limit in the population. In this sense, Williamson's account sweeps the problem of induction under the carpet, if I may say so.

Anyway, let us assume that the inference from sample frequencies to statistical probabilities over the infinite domain of arbitrary future applications can be justified—we think, it can by meta-induction—and turn to the problem that is most important for Williamson, namely the coherent interpretation of the PNRC. What Williamson's PNRC does are two things: first, following Reichenbach it equates the rational degree of belief in Ac with the statistical probability of A in the narrowest (or maximally informative) reference class R containing c, according to our available evidence, and second, it assumes an inductive inference from an evidence E to the statistical probability hypothesis P(A|R).

Now we get to Williamson's major concern. He thinks that the PNRC lead to incoherences. He gives the following example, in which c is a 21 years old man called Cheesewright, R is the property of being a 21 years old male, A the property of getting a cough, S the property of having tuberculosis, X the proposition $P(A|R)=0.17$ and Y the proposition $P(A|R\wedge S)=0.97$. In Sect. 3 Williamson writes that the PNRC would warrant the following five conditional epistemic probabilities about c—which determine one's *actual* degrees of belief by the PNRC, provided conditions (ii) and (iii) are satisfied:[5]

[5] I write "B" instead of Williamson's "P∅", because "P∅" does not fit with Williamson's formulation of the PNRC. Also, I do not know why Williamson prefers here the conditional formulation "B(α|β)" instead of "$B_\beta(\alpha)$".

1. $\mathrm{B}(\mathrm{Ac}\,|\mathrm{Rc}\wedge \mathrm{X}) = 0.17.$
2. $\mathrm{B}(\mathrm{Ac}\,|\mathrm{Rc}\wedge \mathrm{Sc}\wedge \mathrm{Y}) = 0.97.$
3. $\mathrm{B}(\mathrm{Ac}\,|\mathrm{Rc}\wedge \mathrm{X}\wedge \mathrm{Y}) = 0.17.$
4. $\mathrm{B}(\mathrm{Ac}\,|\mathrm{Rc}\wedge \mathrm{Sc}\wedge \mathrm{X}\wedge \mathrm{Y}) = 0.97.$
5. $\mathrm{B}(\mathrm{Ac}\,|\mathrm{Rc}\wedge \neg\mathrm{Sc}\wedge \mathrm{X}\wedge \mathrm{Y}) = 0.17.$

As Williamson convincingly shows (cf. equation (i) in Sect. 3), these five probabilistic claims, taken together, are incoherent. What I want to show is that Williamson's diagnosis *does not* follow, if the notion of defeat in clause (iii) of the PNRC is understood in the right way. I think that all four conditional probabilities 1–4 are warranted by the PNRC. For example, if our evidence contains the conditioning proposition of claim 4, then by clauses (ii) and (iii) of the PNRC claims 1 and 3 are defeated and only claim 4 can determine our actual degree of belief. I agree with Williamson's diagnosis that giving up any of claims 1–4 would undermine the PNRC. What I deny is the warrant of claim 5. As Williamson proves in the end of his Sect. 3,, the knowledge ¬Sc, i.e. that c (Cheesewright) doesn't have tuberculosis, decreases the probability of Ac further below 0.17, for the reason that $P(A|R\wedge S)=0.97 > P(A|R)=0.17$ implies $P(A|R\wedge\neg S)<0.17$, which gives us

$$\mathrm{B}(\mathrm{Ac}\,|\mathrm{Rc}\wedge \neg\mathrm{Sc}\wedge \mathrm{X}\wedge \mathrm{Y}) < 0.17. \tag{5}$$

Thus, Williamson's claim 5 is rejected and given that $Ra\wedge\neg Sa$ is the narrowest known reference class, his claim 1 is defeated by (5).

Williamson sees the plausibility of rejecting his claim 5, but he argues that the acceptance of this argument would undermine any simple application of the PNRC in the form of claim 1. Namely, even without knowing whether or not c has tuberculosis, we know that *there are* various (additional) properties, S, that increase the probability of A and with high plausibility, c does not possess all of them. So we know with high plausibility that there are some properties S such that $P(A|R\wedge S)>P(A|R)$ and ¬Sc. Williamson thinks that this fact (let us take it for granted) is a defeater of claim 1, so by condition (iii) of the PNRC, claim 1 cannot represent our actual degree of belief. I will show now that this impression is *wrong*, for the reason that the additional information is not about an additional relevant property, but about the *existence* of some property, and the no-defeat clause in the PNRC should not count existential information as a defeater; so the PNRC *does not force us* to give up claim 1 as our actual degree of belief. Let us reflect this in more detail: Assume our antecent information is $Rc \wedge P(A|R)=0.17$, so the corresponding conditional degree of belief is $B(Ac|Rc\wedge P(A|R)=0.17) = 0.17$. This conditional belief determines our actual degree of belief in Ac if $Rc \wedge P(A|R)=0.17$ is our total relevant evidence (i.e., the conditions of the PNRC are satisfied). The new additional information is now not the information "$\neg Sc \wedge P(A|R\wedge\neg S)< 0.17$" (which would indeed defeat claim 1), but the existential proposition:

$$\exists S\,(P(A|R\wedge\neg S)<0.17 \wedge \neg Sc). \tag{6}$$

This existential proposition should not count as defeating evidence, for the following reason: If existential propositions of this sort were admitted, then the conditionalization by (6) would not be maximally relevant, because by a similar argument we know that there are also various properties Q that decrease the probability of A and c does not possess all of them, which entails:

(7) $\exists Q\,(P(A|R\wedge\neg Q)>0.17 \wedge \neg Qc)$.

So by Williamson's arguments we would have also a defeater going into the opposite direction, which proves that the information contained in these existential conjectures does not satisfy the condition of maximal relevance. What one would need is the plausibility of a property that conveys to A probability $\neq 0.17$ that is maximally relevant, i.e.,

(8) $\exists S\,(P(A|R\wedge\neg S)<0.17 \wedge \neg Sc \wedge \neg\exists Q\,(P(A|R\wedge\neg S\wedge Q) \geq 0.17 \wedge Qc)$,

but we have no evidence for (8).

To avoid incoherent reasoning with the PNRC, we propose to formulate it in a different manner. Our revision of PNRC follows three guidelines:

1.) We eliminate the phrase "the evidence determines" in condition (i) of the PNRC that hides the problem of induction, and we separate the problem of induction from the PNRC by relativizing the PNRC to the "set of accepted beliefs" that contains singular observational evidence as well as accepted general (statistical) hypotheses that have to be inductively confirmed. More precisely, we are interested in determining the belief B (Ac|B), where B is the set of all of accepted conjunctively elementary beliefs.[6] We represent B by the conjunction:

$B=R^{+}c\wedge E(a_i)\wedge H$, where:

- R^{+} c is the conjunction of all singular beliefs in B about the individual c,
- E (a_i) is the conjunction of all singular beliefs in B about other individuals a_i different from c, and,
- H is the conjunction of all quantified hypotheses in B including statistical hypotheses.

2.) When we transfer the statistical probability of Ax to the target individual c, only singular evidence about the individual c is relevant. Thus, in the statistical probability P(Ax|…) only the individual constant c gets replaced by a corresponding individual variable x, but *not* other individual constants a_i. In the monadic case this is obvious, because if $Fc\wedge Ga$ is our total singular evidence for Ac and we apply the PNRC to both properties F and G, we obtain $P(Ax|Fx\wedge Gy)$ which is equal to P(Ax|Fx) by statistical independence. But even in the relational case our guideline applies, because (for reasons I cannot explain here) we must existentially quantify over variables different from x. In conclusion, in our revised PNRC we identify B(Ac|B) with $P(Ax|R^{+}x)$.

3.) If we follow Williamson's PNRC we should split up $R^{+}c$ further into the conjunction $Rc\wedge Qc$, where Rc is a unique strongest subconjunction of $R^{+}c$ for which we possess information about P (Ax|Rx), and Qc is the remainder subconjunction for which statistical probabilities $P(Ax|Rx\wedge Qx)$ are not known. Williamson proposes to identify B (Ac|B) with P (Ax|Rx) *provided* clause (iiii) is satisfied, i.e. the remainder evidence $Qx\wedge E(a_i)$ is not relevant. We know that E (a_i) cannot be relevant; so the only problem is the assessment of the relevance of Qa, but this is difficult. A simple strategy is to assume *by default* that properties of c whose statistical probabilities are unknown are irrelevant, i.e. $P(Ax|R^{+}x)=P(Ax|Rx)$. In many cases this would be too simple, because we may have *unsharp* information about the value of $P(Ax|Rx\wedge Qx)$, e.g. that $P(Ax|Rx\wedge Qx)$ lies in some interval. In these and other

[6] A statement B is conjunctively elementary if it is not logically equivalent with a conjunction of*shorter* statements (for details see Schurz 2022b, 3, def. 1).

cases (e.g., when there is no unique strongest subconjunction Rc) one has to *estimate* the statistical probability $P(Ax|R^{+}x)$ (cf. Kyburg 1961, 222-226; Thorn 2012). So we propose to use the *estimated* statistical probability $P_{est}(Ax|R^{+}x)$. Because of this step, there is no need in our in our revised version of PNRC for a separate "no-defeat" condition (iii).

Revised PNRC: $B(Ac|B) = B(Ax|R^{+}x \wedge E(a_i) \wedge H) = P_{est}(Ax|R^{+}x)$, where $R^{+}x$, $E(a_i)$, H and $P_{est}(Ax|R^{+}x)$ are characterized as above.

This concludes our defense of the PNRC against Williamson's objections. An extensive discussion that does justice to all details of the following sections of Williamson's paper would require a paper of its own; so I confine myself to a few remarks. In his Sect. 4, Williamson raises an analogous 'undermination' argument against Lewis' principal principle (PP) for single case chances, which according to our diagnosis relies on a too liberal understanding of an "admissible" evidence E. Only facts up to time t (the present) are admissible; therefore the proposition $A \leftrightarrow F$ in Eq. 8 of Williamson's Sect. 4 is not admissible, because A is about the future. Since Williamson thinks that both the PNRC and the PP are undermined by his objections, he develops a non-standard version of objective Bayesianism that rejects the principle CBCP. In our view, giving up CBCP has drastic disadvantages, but fortunately there is no need for such a step because Williamson's undermination arguments can be defeated.

There is also an inductive inference smuggled into Williamson's principles of chance and frequency calibration in his Sect. 5, which are Williamson's elaborated versions of the PP and the PNRC, respectively. Both calibration principles assume in their if-condition that "according to current evidence E, the current chance function [or frequency (limit) function] P* lies in a set of probability functions". Obviously, the question which chance or frequency function "accords with the evidence" is nothing but the problem of induction.

At the end of Sect. 5, Williamson presents his version of the Williams-Stove symmetry argument that is critically analyzed in Schurz (2019, Sect. 4.6). In premise (ii) of this argument Williamson assumes the inductive inference from a sample frequency to a confidence interval I_τ (cf. Schurz 2014, 191 f., 226-232, for the inductive assumptions behind that method). In Schurz (2019, 73 f.) I point out that Williamson's argument fills a gap in the Williams-Stove version, namely the employment of the PP (or the PNRC) in passing from the general claim

(9) With high statistical probability (τ) over samples, the population frequency lies in the confidence interval I_τ around the sample frequency.

to the instantiation of claim (9) for the given *particular* sample s:

(10) With high epistemic probability, the population frequency lies in the confidence interval I_τ around the sample frequency of s.

I agree with Williamson that given the PP (or PNRC) these two steps (steps iv and v in Williamson's paper) are correct. What is criticized in Schurz (2019, 74) is the additional step in Williamson (2013) in which he conditionalizes the epistemic probability in (10) on the additional information about the particular *frequency value* in sample s:

(11) With high epistemic probability, the population frequency lies in the confidence interval I_τ around the sample frequency of s which is (say) 0.8.

The latter step violates the admissibility condition. Interestingly, in the version of the argument in Williamson's paper this last step is omitted; but in his subsequent discussion step (11) is again assumed. In Sect. 6 Williamson discusses my criticism in Schurz (2019, Sect. 4.6), in particular my coin tossing counterexample in which our prior evidence that the coin is fair counteracts against an unbiased inductive projection of the observation of an improbable sample frequency of 0.3 to the population modulo confidence interval. When Williamson writes "this evidence may be enough to resist the inference that the frequency or chance is in the 95% confidence interval [around the value of 0.3—G.S.]", I take this as an indication that Williamson agrees with my point. What he nevertheless denies is the conclusion drawn by Maher (1996) and supported by my analysis, namely that probabilistic induction in Williamson's style depends implicitly on a uniform prior distribution. Williamson tries to avoid this conclusion by his rejection of the principle CBCP; he even argues that the evidence E need not be included in the domain of the probability function B. Similarly Williamson agrees with my challenge that state-uniform distributions prohibit induction, but argues that by giving up the CBCP, this challenge can be avoided. For me, however, giving up CBCP seems to lead into a dead end, because this step implicitly amounts to giving up the attempt of explicating the probabilistic relations between our beliefs and the evidence.

6 From Meta-Induction to Abduction—or Lessons from Aliseda

The method of meta-induction is not only applicable to inductive inferences, but also to abductive inferences to the best explaining theoretical hypothesis or theory. The distinguishing characteristics of theoretical hypotheses is that they contain theoretical concepts, or latent variables—these are concepts or variables *not* contained in the empirical evidence that the theories attempt to explain. An example is the explanation of the trajectories of the planets by the sun's gravitational force in Newtonian mechanics; gravitational forces are not part of the observed astronomical data (the trajectories of planets), but they are theoretically postulated to explain these data.

When we apply meta-induction to theories we evaluate them as tools for predictions. There are two important differences compared to standard meta-induction in prediction games. First, theoretical hypotheses typically predict or explain *empirical regularitie*s. We therefore assume that the entities to be predicted or explained are *samples* of data, abbreviated as s_1, s_2, …. Each sample s_i consists of observed values of correlated variables. The theoretical hypotheses deliver predictions of the 'dependent' variables conditional on some (chosen) 'independent' variables. The second difference is that the 'predictions' of the sampled data need not be proper predictions but may also be explanations, i.e., data whose values have already been observed. Thereby we must restrict the data confirming the competing hypotheses to *use-novel* data—data that have not been used in fitting free parameters to the data. Thus, the success of competing theoretical hypotheses is evaluated by a sequence of use-novel data samples.

Of course, the competing theoretical hypotheses cannot be directly compared with the observed data, because they make assertions about the unobservable. Rather, they are evaluated by an aggregated measure of success in potential predictions, combined with simplicity, as described in Schurz (2022a, Sect. 5.1); we speak here of *instrumentalistic success evaluation* (for a similar account see Feldbacher-Escamilla 2022 about "meta-abduction").

The instrumentalistic success of a theory warrants belief in its empirical predictions, but not necessarily belief in its theoretical content. Meta-induction over predictive success is compatible with the instrumentalistic position in philosophy of science, exemplified by the empiricism of van Fraassen (1980). According to this position, we are warranted to believe in the empirical adequacy of well-confirmed scientific theories, but not in their realistic truth. For example, a scientific instrumentalist will believe in the reality of planets and their trajectories, but not in the reality of the gravitational force, while the scientific realist tends to infer from their instrumentalistic success the reality of gravitational forces. The latter inference is *abductive* in nature and not reducible to meta-inductive optimality; its justification requires stronger and at the same time more controversial epistemological principles. In Schurz (2022a, Sect. 5.2, memo (10)) it is argued that if a theory T dominates every equally successful competitor theory T' in the sense that T' contains an isomorphic copy of T as a submodel but involves additional complications, then the abductive inference from the predictive success of T to T's realistic truthlikeness is justified.

In this section, however, we don't speak about the abductive inference from instrumentalistic success to realistic truthlikeness; we will rather say more about meta-inductive theory-aggregation as a means of optimizing the instrumentalistic success of theory-generating abduction. This perspective leads us to Aliseda's excellent reconstruction and elaboration of my account of scientific theory revision (Schurz 2011). In Schurz (2018) this account is called the "construction paradigm of theory development". As Aliseda makes it clear, the account intends to complement (rather than to replace) the standard rational choice paradigm of theory development (Rott 2001).

The standard AGM approach of belief revision (Alchourrón et al. 1985) describes the development of a theory or belief set T induced by (empirical) input information e in terms of expansions, contractions and revisions. In the following we speak of theories and belief sets interchangeably (note that theories have several important parts or subsets; cf. Aliseda's explication in her Sect. II). The *expansion* of a theory T by a T-compatible input e is denoted as T+e and defined as $Cn(T\cup\{e\})$ (where "Cn" is the consequence operator). If e contradicts T, then one first contracts T by ¬e before one can expand by e. The *contraction* of T by e is denoted by T÷e and intended to be some preferred T-subset which does not entail e; different methods of defining contraction operations have been suggested (cf. Gärdenfors 1988, Rott 2001). Finally, the *revision* of T by a T-incompatible input e is denoted by T*e and defined as a sequence of a contraction and an expansion, $T{*}e = (T \div \neg e) + e$; this definition is also called the Levi-identity (after Levi 1980).

As worked out in Schurz (2011), AGM revision is purely corrective in the following sense: if a new observation e obtains, then e is just added to T or to T÷¬e, but T doesn't learn from e in the sense that T is enriched by new (inductively or abductively inferred) hypotheses so that the new system can *explain* e. Precisely this is the task of *abductive* theory expansion and revision that is in the focus of the work of Aliseda (2006) and Schurz (2011, 2018). In an *abductive expansion*, the theory T receives a new empirical input e that is consistent with T but cannot be explained by T; in this case T is expanded not only by e but also by a new hypothesis h that can explain e together with T; Aliseda (Sect. II)) calls such an empirical input e an abductive novelty. In an *abductive revision* T receives an empirical input e that is not only unexplainable but even inconsistent with T; in this case one first forms a suitable contraction $T \div \neg e$ which is then abductively expanded by e and a suitable hypothesis h that can explain e together with $T \div \neg e$.

So far I have explained the standard theory of abductive expansion and revision as developed by Pagnucco (1996), Aliseda (2006) and others. The standard account is indeterminate insofar it does not specify criteria in what a good explanation consists nor does it give operations by which good explanations can be found, if they can be found at all. What Schurz (2011) adds to this standard account are more detailed operations by which abductive theory expansions and revisions are carried out in science.

Aliseda starts her discussion of abductive theory revision in Sect. III with a discovery in Schurz (2011, 95 f.) and later in Niiniluoto (2018), namely the failure of the Levi identity for the following reason: Assume an element h of a belief set T that explains each member of a set of pieces of evidence E (= $\{e_1, \ldots, e_n\}$), and e is a new piece of evidence contradicting h; so that $T \div h$ is the minimal contraction of T consistent with e. Then it is inefficient to remove first h from T and generate an alternative hypothesis h* from scratch—what one would have to do according to Levi identity—because the contraction $T \div h$ would fail to explain the old evidence in E. It would be hard to find 'from scratch', after having forgotten h, a new hypothesis h* (in the context of $T \div h$) that simultaneously explains each member in E and e. In any case, such a h* would not be the result of a standard abductive expansion of the contracted theory $T \div h$, as this standard abductive expansion would only explain e but not E. An appropriate revised hypothesis h* must result from an abduction expansion of $T \div h$ by $E \cup \{e\}$; so Levi's identity fails.

What scientists typically do in such a situation is try to construct the hypothesis h* by a *direct revision* of the old hypothesis h in view of $E \cup \{e\}$, such that h* explains e and at the same time preserves the explanations of the old evidence E (Schurz 2018, 477–478). Aliseda (Sect. III) calls this revision operation *theory-refinement;* however in some cases such a refinement may involve a quite radical change of parts of the theory. In Schurz (2011, 7) I give a concrete example of a direct revision for a hypothesis h: $Y = c_o + c_1 \cdot X$ that is a linear quantitative law relating two magnitudes X and Y; my example was the *ideal gas law*. When new data come in implying that for low values of X, Y is lower than predicted by the linear relationship h, scientists do not simply remove h from T, but replace h by a modified hypotheses h* in which a new non-linear term is added (in the ideal gas law case a negative quadratic term). So in this case, the operation of hypothesis refinement consists in the addition of new terms to a functional equation.

Schurz (2011, Sect. 4.3.5) describes a general abductive expansion and revision operation for scientific theories about dynamical systems; these are systems about particles whose movement in space and time is explained via 'generalized' forces. One of my examples is the derivation of the elliptic orbits of planets from Newton's theory T together with the auxiliary assumption that the only non-negligible force acting on the planet is the gravitational force of the sun. When Adams and Leverrier in 1846 recorded a significant deviation of Uranus' orbit from the predicted orbit, they did not just remove this auxiliary hypothesis, but replaced it by the new auxiliary hypothesis "there exists an hitherto unobserved small planet, called Neptune, whose gravitational force deflects Uranus' orbit", which together with the remaining part of the theory could explain the observed orbit of Uranus.

The existence of the planet Neptune was first merely postulated, but later observed with stronger telescopes. This was regarded as a great success of Newtonian physics. In the final part of her paper Aliseda introduces an important extension of the account of abductive theory revision that accounts for this kind of theory development: abductive hypothesis refinement by *existential instantiation*. Here the hypothesis in question contains an existential statement, e.g. about the existence of some hitherto unknown disturbing factor, and the abductive theory refinement replaces this existentially quantified variable by a concrete entity that instantiates the disturbing factor whose existence is postulated, which increases the theory's content. Aliseda illustrates her idea by an example from medicine.

Let me finally consider abductive theory revision within the meta-inductive framework. How should the weighted average of several theoretical hypotheses be interpreted? In situations where there is a unique best hypothesis or theory, the interpretation is easy: then aMI (attractivity-weighted meta-induction) converges to ITB (Imitate the best); thus the best theory is meta-inductively selected and its competitors are ignored because of their negligible weights. More difficult is the general case where several theoretical hypotheses have a non-negligible weight. This brings us to the application of meta-induction to abductive theory revision. Let us make the situation precise: T is a theory consisting of elementary statements (not closed under logical consequences), $T = \{s_1, \ldots s_n\}$, including core axioms and auxiliary hypotheses, and $E = \{e_1, \ldots e_n\}$ is the set of pieces of evidence successfully explained or predicted by T, and e_{n+1} a new piece of evidence contradicting T. Let H be a 'minimal' and least important subset of T (typically auxiliary hypotheses located in T's periphery) so that the set-theoretic difference T–H is consistent with e_{n+1} (thus T–H is a suitable contraction of T by $\neg e_{n+1}$). Then the abductive revision problem consist in finding a new H* that simultaneously explains e_{n+1} and does not lose the explanatory force in regard to E. In the above example finding this H* was straightforward, but often in science it is extremely hard to find such a H*. What one often finds is a H* that together with T–H is able to explain e_{n+1} and several of the old pieces of evidence in E, but H* can no longer explain certain other data in E, we call them E* ($\subset$ E) (E* has also been called "Kuhn-loss"; cf. Hoyningen-Huene 1993). In this situation we have *two competing theories*, the old theory T that explains E but not e_{n+1}, and a competing theory $T^* = (T–H) \cup H^*$ that explains e_{n+1} and E–E* but not E*. A case in point is the famous classical "planetary" model of atoms of Rutherford and Bohr (H) within classical physics (T), that was later replaced by Bohr's quantized stability postulate for electron orbits (H*), as described in Schurz and Lambert (1994, 104).

In such a case aMI has to combine the two theories, weighted by their attractivity-weights. However, this is just a metaphorical way of speaking because what is really combined by aMI are not the two theories but their real-valued predictions. Working with two (or several) theories whose predictions have always to be weighted is rather complicated. It would be preferable if the meta-inductivist could also combine the two theories themselves to obtain a superior unified theoretical model. Whether this is possible cannot be said in general but depends on the particular context. For example, if conditionalized meta-induction discovers that H_1 is dominantly successful in a given domain of applications A_1 (H_2's weight is small in A_1), and vice versa for H_2 and A_2, then the meta-inductively recommended combination is $H_{aMI} = (A_1 \rightarrow H_1) \wedge (A_2 \rightarrow H_2)$. Moreover, if the hypotheses are structural equation models in the same variables, $Y = f_i(X)$, then they can be literally combined by a weighted averaging of the functional expressions: $Y_{aMI} = w \cdot f_1(X) + (1\text{-}w) \cdot f_2(X)$. Finding intrinsic theory-combinations is an important research program triggered by meta-induction over scientific theories.

Funding Open Access funding was enabled and organized by Projekt DEAL. This work was supported by the DFG grant SCHU 1566/9-1 as part of the priority program "New Frameworks of Rationality" (SPP 1516) and by the DFG grant SCHU 1566/11-2 as part of the DFG research unit "Inductive Metaphysics" (FOR 2495).

References

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change. *Journal of Symbolic Logic* 50: 510–530.
Aliseda, Atoacha. 2006. *Abductive reasoning*. Dordrecht: Springer.
Bovens, Luc, and Stephan Hartmann. 2003. *Bayesian epistemology*. Oxford: OUP.
Burnham, Kenneth P., and David R. Anderson. 2002. *Model selection and multimodel inference*. 2nd ed. New York: Springer.
Carnap, Rudolf. 1950. *Logical foundations of probability*. Chicago: University of Chicago Press.
Cesa-Bianchi, Nicolò, and Gabor Lugosi. 2006. *Prediction, learning, and games*. Cambridge: CUP.
Feldbacher-Escamilla, Christian J. 2022. Meta-abduction. Inference to the probabilistically best prediction. In *Philosophy of computing*, ed. Björn Lundgren, Nuñez Hernandez & Nancy Abigail, 51–72. Cham: Springer.
Gärdenfors, Peter. 1988. *Knowledge in flux*. Cambridge/Mass.: MIT Press.
Howson, Colin, and Peter Urbach. 1996. *Scientific reasoning: The Bayesian approach*. 2nd ed. Chicago: Open Court.
Hoyningen-Huene, Paul. 1993. *Reconstructing scientific revolutions*. Chicago: University of Chicago Press.
Hume, David. 1748/2000. *An Enquiry concerning Human Understanding*, ed. by Tom L. Beauchamp. Oxford: Clarendon Press.
Inglehart, Ronald, and Pippa Norris. 2003. *Rising tide: gender equality and cultural change around the world*. Cambridge/Mass.: CUP.
Kelly, Kevin T. 1996. *The logic of reliable inquiry*. New York: OUP.
Kyburg, Henry E. 1961. *Probability and the logic of rational belief.* Middletown/CT: Wesleyan University Press.
Levi, Isaac. 1980. *The enterprise of knowledge*. Cambridge/Mass.: MIT Press.
Lewis, David. 1980. *A subjectivist's guide to objective chance.* Reprinted in: Lewis, David. *Philosophical papers vol II, chapter 19*. New York: OUP 1986.
Maher, Patrick. 1996. The hole in the ground of induction. *Australasian Journal of Philosophy* 74/3: 423–432.
McGrew, Timothy. 2019. Miracles. *Stanford encyclopedia of philosophy* (spring 2019 edition). plato.stanford.edu/archives/spr2019/entries/miracles/.
Niiniluoto, Ilka. 2018. *Truth-seeking by abduction*. Cham: Springer.
Pagnucco, Maurice. 1996. *The role of abductive reasoning within the process of belief revision*. Dissertation. University of Syndey. www.cse.unsw.edu.au/~morri/Papers/morri.PhD.pdf.
Piattelli-Palmarini, Massimo. 1994. *Inevitable illusions*. Hoboken: Wiley.
Putnam, Hilary. 1965. Trial and error predicates and a solution to a problem of Mostowski. *Journal of Symbolic Logic* 30: 49–57.
Reichenbach, Hans. 1949. *The theory of probability*. Berkeley: University of California Press.
Rott, Hans. 2001. *Change, choice and inference*. Oxford: Clarendon Press.
Salmon, Wesley C. 1957. Should we attempt to justify induction? *Philosophical Studies* 8/3: 45–47.
Schurz, Gerhard. 2008. The meta-inductivist's winning strategy in the prediction game: a new approach to Hume's problem. *Philosophy of Science* 75: 278–305.
Schurz, Gerhard. 2011. Abductive belief revision. In *Belief revision meets philosophy of science*, eds. Erik Olsson & Sebastian Enqvist, 77–104. New York: Springer.
Schurz, Gerhard. 2014. *Philosophy of science: a unified approach*. New York: Routledge.

Schurz, Gerhard. 2018. Truthlikeness and approximate truth. In *Routledge Handbook of Scientific Realism*, ed. Juha Saatsi, 133–148. New York: Routledge.
Schurz, Gerhard. 2019. *Hume's problem solved: the optimality of meta-induction*. Cambridge/MA: MIT Press.
Schurz, Gerhard. 2022a. Optimality justifications and the optimality principle: New tools for foundation-theoretic epistemology. *Noûs* 56: 972–999.
Schurz, Gerhard. 2022b. Tacking by conjunction, genuine confirmation and convergence to certainty. *European Journal for the Philosophy of Science* 2: 46 (1–18).
Schurz, Gerhard. 2022c. Evolution of rationality. In *The handbook of rationality*, eds. Markus Knauff & Wolfgang Spohn, 101–113. Cambridge/MA: MIT Press.
Schurz, Gerhard, and Karel Lambert. 1994. Outline of a theory of scientific understanding. *Synthese* 101/1: 65–120.
Schurz, Gerhard, and Paul Thorn. 2016. The revenge of ecological rationality: strategy-selection by meta-induction. *Minds and Machines* 26 (1): 31–59.
Schurz, Gerhard, and Paul Thorn. 2022. Escaping the no free lunch theorem: a priori advantages of regret-based meta-induction. *Journal for Experimental and Theoretical Artificial Intelligence*. https://doi.org/10.1080/0952813X.2022.2080278.
Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding machine learning. From theory to algorithms*. New York: CUP.
Skyrms, Brian. 1975. *Choice and chance*. Encinco: Dickenson (4th ed. Wadsworth 2000).
Steel, Daniel. 2009. Testability and Ockham's razor: how formal and statistical learning theory converge in the new riddle of induction. *Philosophy of Science* 38: 471–489.
Sterkenburg, Tom. 2020. The meta-inductive justification of induction. *Episteme* 7/4: 519–541.
Thorn, Paul. 2012. Two problems of direct inference. *Erkenntnis* 76: 299–318.
Van Fraassen, Bas. 1980. *The scientific image*. Oxford: Clarendon Press.
Wallace, Anthony, F. C. 1966. *Religion: an anthropological view*. New York: Random House.
Williamson, Jon. 2013. Why frequentists and Bayesians need each other. *Erkenntnis* 78: 293–318.