The Meta-inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem*

Gerhard Schurz†‡

This article suggests a 'best alternative' justification of induction (in the sense of Reichenbach) which is based on *meta-induction*. The meta-inductivist applies the principle of induction to all competing prediction methods which are *accessible* to her. It is demonstrated, and illustrated by computer simulations, that there exist meta-inductivistic prediction strategies whose success is approximately optimal among all accessible prediction methods in arbitrary possible worlds, and which dominate the success of every noninductive prediction strategy. The proposed justification of meta-induction is mathematically analytical. It implies, however, an a posteriori justification of object-induction based on the experiences in our world.

1. Introduction and Conceptual Clarification. In this article I understand the notion of an *inductive inference* in the narrow 'Humean' sense, in which a property, regularity, or frequency is transferred from the observed to the unobserved, or from the past to the future. Other forms of non-deductive inferences such as inferences 'to the best explanation' are not considered. My article is an attempt at a positive solution to the problem of induction (or Hume's problem): how can we *rationally justify* inductive inferences? I will focus on *predictive* inferences, as the achievement of predictive success is the major concern of all methods of induction.

My setting shall cover *binary* as well as *real-valued* prediction games. I assume the *stream of events* consists of an infinite sequence $(e) := (e_1, e_2, \ldots)$ of events which are coded or measured by elements of the unit interval $[0, 1]$. Hence at each discrete time $n = 1, 2, \ldots$ an event $e_n \in$

*Received November 2005; revised March 2008.

†To contact the author, please write to: Department of Philosophy, University of Duesseldorf, Universitaetsstrasse 1, Geb. 23.21, Duesseldorf, Germany D-40225; e-mail: gerhard.schurz@phil-fak.uni-duesseldorf.de.

‡For valuable help I am indebted to Ronald Ortner, Eckhart Arnold, Markus Werning, Brian Skyrms, Nicholas Rescher, and an anonymous referee.

Philosophy of Science, 75 (July 2008) pp. 278–305. 0031-8248/2008/7503-0002$10.00

[0, 1] obtains. For example, (e) may be a sequence of daily weather conditions, stock values, or coin tossings, etc. In a *prediction game* several players P_1, P_2, . . . have to predict future events of the event sequence. In what follows, $p_n(P)$ denotes the prediction of player P *for* time n, which is delivered *at* time $n - 1$. Also the admissible predictions p_n are assumed to be elements of [0, 1]. In *binary* prediction games, predictions as well as events must take one of the two values 0 and 1 which code instantiations of a binary event-type E ('1' for 'E obtains' and '0' for 'E does not obtain'). The deviation of the prediction p_n from the event e_n is measured by a normalized loss function $l(p_n, e_n) \in [0, 1]$. We define the *natural* loss-function as the absolute difference between prediction and event, $l(p_n, e_n) := |p_n - e_n|$. However, my theorems will not depend on natural loss functions but hold for arbitrary (and in case of Theorems 4 and 5 for convex) normalized loss-functions.

A bit more notation: The *score* $s(p_n, e_n)$ obtained by prediction p_n given event e_n is defined as 1 minus the loss, $s(p_n, e_n) := 1 - l(p_n, e_n)$, and the *absolute* success $a_n(P)$ achieved by player P at time n is defined as P's sum of scores until time n, $a_n(P) := \sum_{1 \leq i \leq n} s(p_i(P), e_i)$. The *success rate* $\mathrm{suc}_n(P)$ of player P at time n is defined as $\mathrm{suc}_n(P) := a_n(P)/n$. Finally, $\bar{e}_n := (\sum_{1 \leq i \leq n} e_n)/n$ denotes the event's *mean* value at time n, and $\bar{e} := \lim_{n \to \infty} \bar{e}_n$ denotes the event's *limit* mean value, provided the mean values converge to a limit. (Recall that $\lim_{n \to \infty}(x) = a$ iff $\forall \varepsilon > 0 \exists n \forall m \geq n: |x_m - a| \leq \varepsilon$.) Note that for binary prediction games, (i) $\mathrm{suc}_n(P)$ equals the relative frequency of P's correct predictions until time n, (ii) $\bar{e}_n$ equals the relative frequency $f_n(E)$ of event E at time n, and (iii) $\bar{e}$ equals E's limiting frequency $\lim_{n \to \infty} f_n(E)$. Theorems 1–3 of my article will be *insensitive* to the difference between binary and real-valued prediction games; but Theorems 4 and 5 will depend on it.

Finally, I will avoid Goodman's 'new riddle' of induction by assuming that the events are explicated in terms of *qualitative* predicates or function symbols in the sense of Carnap (1947, 146), whose definitions in terms of primitives do not mention individual constants. Under this assumption, Goodman's paradox does not arise. What is left from it is the lesson that there exists an infinite number of anti-inductive generalization rules which conjecture that after some time in the future the regularities observed so far will be subverted (cf. Howson 2000, 43ff.). This lesson deepens Hume's problem, but it does not go beyond it.

2. Reichenbach's Best Alternative Approach and Its Shortcomings. David Hume has shown that all standard methods of justification seem to fail when applied to the task of justifying induction. Obviously, inductive inferences cannot be justified by deductive logic, since it is logically possible that the future is completely different from the past. More impor-

tantly, induction cannot be justified by induction from observation, by arguing that induction has been successful in the past, whence—by induction— it will be successful in the future. For this argument is *circular*, and circular arguments are without any justificatory value. Salmon (1957, 46) has shown that also anti-induction may be pseudojustified in such a circular manner. It is equally impossible to demonstrate that inductive inferences are reliable in a probabilistic sense—for in order to do so, one must presuppose that the relative event frequencies observed so far can be transferred to the unobserved future, which is nothing but an inductive inference. These were the reasons that led Hume to the skeptical conclusion that induction cannot be rationally justified, but is merely the result of psychological habit.

There have been several attempts to solve or dissolve Hume's problem, which cannot be discussed here. It seems that so far, none of these attempts has been successful in giving a *positive* justification of induction, which establishes in a *noncircular* manner that the inductive method is a superior prediction method in terms of its success frequencies. Let me emphasize that this is neither obvious nor guaranteed. Millions of people *do* in fact believe in superior noninductive methods, be it God-guided inner intuition, clairvoyance, or other supernatural abilities. Therefore a satisfying justification of induction would not only be of fundamental *epistemological* importance; it would also be of fundamental *cultural* importance as part of the enterprise of enhancing scientific rationality.

Assuming that it is impossible to demonstrate that induction must be successful (Hume's lesson), and that there are various alternative prediction methods, it seems to follow that the only approach to Hume's problem for which one can at least uphold the *hope* that it *could succeed* if it were adequately developed is Reichenbach's *best alternative approach* (Reichenbach 1949, sec. 91; Salmon 1974). This approach does not try to show that induction *must* be successful, but it attempts to establish that induction is an *optimal* prediction method—its success will be maximal among *all* competing methods in arbitrary possible worlds. Or in simplified words: if any method of prediction will work, then the inductive method will work (Rescher 1980, 207ff.). If the demonstration of the optimality of induction succeeded, then one could extend this result to a demonstration of induction's (weak) *dominance* over noninductive methods, because every noninductive method predicts suboptimally in a 'sufficiently normal' world. It must be emphasized that in demonstrating optimality one must allow *all* possible worlds, including all kinds of *paranormal* worlds in which perfectly successful future-tellers or anti-inductivistic demons do indeed exist. Restricting the set of worlds to 'normal' or uniform worlds would completely *destroy* the enterprise of justifying

induction. For then we would have to justify inductively that our real world is one of these 'normal' worlds, and we would end up in that kind of circle or infinite regress in which according to the *Humean skeptic* all attempts of justifying induction must end up.

Unfortunately, Reichenbach *failed* to establish an optimality argument with respect to the goal of predictions. He only demonstrated an optimality argument with respect to the goal of approximating the frequency limit (or mean). With respect to that goal, the argument is *almost trivial*: if the event sequence has a frequency limit, then the inductive straight rule, which transfers the observed frequency to the conjectured frequency-limit, *must* approximate this limit in the long run, while other noninductive methods may or may not approximate the limit; but if the sequence of events does not have a frequency limit, then *no* method can find the limit (Reichenbach 1949, 474–475). However, our ability to infer approximately correct frequency limits is practically not significant. What *is* of practical significance is our success in true *predictions*. In this respect, Reichenbach's approach fails. Nothing in Reichenbach's argument excludes that God-guided clairvoyants may be predictively much more successful than the object-inductivist. A perfect future-teller may have perfect success in predicting random tossings of a coin, while the Reichenbachian object-inductivist can only have a predictive success of .5 in this case. Reichenbach was well aware of this problem, and he remarked that if successful future-teller existed, then the inductivist would recognize this by applying induction to the success of prediction methods (1938, 358–359; 1949, 476–477). But Reichenbach did neither show nor even attempt to show that by this meta-inductivistic observation the inductivist could have equally high predictive success as the future-teller (this point has been highlighted by Skyrms 1975, Chapter III.4).

By *object-induction* (OI) I understand methods of induction applied at the level of events—the 'object level'. The general problem of Reichenbach's account lies in the fact that it is *impossible* to demonstrate that object-induction is an optimal prediction method (which is also a lesson of formal learning theory, see Section 3). In contrast to Reichenbach's approach, my approach is based in the idea of *meta-induction*. The meta-inductivist (MI) applies the inductive method at the level of competing prediction methods. More precisely, the meta-inductivist bases her predictions on the predictions and the observed success rates of the other (non-MI) players and tries to derive therefrom an 'optimal' prediction. The simplest type of MI predicts what the presently best prediction method predicts, but one can construct much more refined kinds of meta-inductivistic prediction strategies.

One should expect that for meta-induction the chances of demonstrating

optimality are much better than for object-induction. The crucial question of this article will be: is it possible to design a version of meta-induction which can be proved to be an (approximately) optimal prediction method? The significance of this question for the problem of induction is this: if the answer is positive, then at least meta-induction would have a rational and noncircular justification based on a mathematical-analytic argument. But this analytic justification of *meta-induction* would at the same time yield an *a posteriori* justification of *object-induction* in the real world: for we know by experience that in the real world, noninductive prediction strategies have not been successful so far, hence it would be meta-inductively justified to favor object-inductivistic strategies.

Let me finally discuss two possible objections to my approach. A first objection might complain that I presuppose that the predictions of the other players for the next time are *accessible* to the meta-inductivist. There might be possible worlds in which alternative players do not *give away* their predictions but keep them secret. Indeed, this is possible, and so I have to restrict my claim to *accessible methods*. What I intend to show is that among all prediction methods (or strategies) whose *output* is accessible to a given person, the meta-inductivistic strategy is always the best choice. I argue that this restriction is not a drawback. For methods whose output is not accessible to a person are not among her *possible actions* and, hence, are without relevance for the optimality argument.

A second objection could point out that my proposed justification of (meta-)induction (if it were successful) is not an epistemic but a practical justification, and hence is not a true solution to Hume's problem. I think this diagnosis is incorrect. For the goal underlying the optimality argument is maximization of true predictions, and this is clearly an *epistemic* and not a practical goal. Although an optimality justification is weaker than a reliability justification, it is nevertheless an epistemic justification. Moreover, Hume did not only argue that induction cannot be demonstrated to be reliable, but he argued for the much stronger view that no epistemically rational justification of induction is possible. Therefore I regard optimality arguments as a genuine though weak solution to Hume's problem.

3. Prediction Games. A prediction game is a *pair* $((e), \Pi)$ consisting of:

1. An infinite sequence $(e) := (e_n : n \in |N^+)$ of events $e_n \in [1, 0]$ as explained. Each discrete time unit $n = 1, 2, \ldots$ corresponds to one *round* of the game.
2. A set of players $\Pi = \{P_1, P_2, \ldots, x\text{MI}\,(x\text{MI}_1, x\text{MI}_2, \ldots)\}$, whose task is to predict, at any time n, which event will occur at the next time $n + 1$. The players include:

2.1. The object-inductivist OI := P_1 (OI has index 1), whose prediction method is explained below. OI has informational access to past events; his first prediction (at $n = 1$) is a guess.
2.2. A subset of alternative players $P_2, P_3, \ldots$; for example, persons who rely on their instinct, God-guided future-tellers, etc. In paranormal worlds, these alternative players may have any success and any information you want, including information about future events and about the meta-inductivist's favorites. Players of Type 2.1 or 2.2 are called *non-MI-players*.
2.3. One or several meta-inductivists of a certain type, whose denotation has the form 'xMI', where 'x' is a variable (possibly empty) expression specifying the *type* of the meta-inductivist. The meta-inductivist has access to the past events and the past and present predictions of the non-MI-players.

The simplest type of meta-inductivist from which I start my inquiry is abbreviated as MI. At each time, MI predicts what the non-MI-player with the presently highest predictive success rate predicts. If P is this presently best player, then I say that P is MI's present *favorite*, or simply that MI *favors* P. If there are several best players, MI chooses the first best player in an assumed ordering, say, from left to right. MI changes his favorite player only if another player becomes *strictly* better; otherwise he sticks to his present favorite. In what follows, $\text{fav}_n(\text{MI})$ denotes MI's favorite for time n, that is, the player with the first-best success-rate at time $n - 1$ among the non-MI-players in the assumed ordering. Observe that MI's favorite for time n is determined at time $n - 1$. MI's first favorite is OI. I assume that MI has always access to OI: even if no person different from MI plays OI, MI may constantly simulate the predictions of OI and use them if their virtual success supersedes that of the alternative players.

The simplest object-inductive prediction method is based on the already mentioned *straight rule*. In the case of *real-valued* events, this inductive rule transfers the observed mean value to the next event, that is, $p_{n+1}(\text{OI}) = \bar{e}_n$. In the case of *binary* events, the straight rule is merely used for conjecturing the frequency limit (cf. Salmon 1974, 89–95; Rescher 1980, Chapter 6). For the purpose of single event predictions one needs in addition the so-called *maximum rule*, which requires one to predict an event with maximal conjectured frequency (cf. Reichenbach 1938, 310–311). For binary events, this generates the prediction rule $p_{n+1}(\text{OI}) = 1$ if $f_n(E) \geq 1$, and else $= 0$, which can be summarized by saying that OI predicts the *integer-rounding* $[\bar{e}_n]$ of $\bar{e}_n$. OI's prediction rule is appropriate as long as the event sequence is *random* in the sense that it converges to a limit but there exist no correlations between past and future events which could be utilized in predictions. For random sequences with natural

loss functions, OI's success rate converges in the binary case against the maximum of $P(E)$ and $P(\neg E)$, and in the real-valued case against the limiting mean value of the absolute deviations, $\lim_{n\to\infty}(\sum_{1\leq i\leq n}|e_i - \bar{e}|/n)$. For nonrandom sequences refined object-inductivistic prediction strategies exist, whose success dominates OI's success (see Section 9.1).

In general, a prediction strategy is an *object-strategy* if its input information does not contain information about the predictions of other players; otherwise it is a *meta-strategy*. Besides meta-inductivistic strategies there exist other meta-strategies, for example, *blind-favorite* strategies who favor a certain player whatever happens. It is important to distinguish between a player *playing* a strategy S, and a player merely *favoring* a strategy S—in the latter case, the player does not play S but plays a metastrategy that favors S. A player is able to play a certain strategy S iff S is *internally* accessible to him, which means that he can reproduce the strategy with help of some mechanism. In contrast, a player P is able to favor a strategy S if S is *output-accessible* to P, that is, some other player Q plays S and P has access to Q's predictions. A strategy is called *accessible* to P if it is at least output-accessible (or even internally accessible) to P.

I identify prediction games with *possible worlds*. Apart from the definition of a prediction game, I make no assumptions about these possible worlds. In particular, my approach does not depend on some (problematic) distinction between 'uniform' or 'nonuniform' worlds (such a distinction is notoriously difficult; cf. Skyrms 1975, 34–35). The stream of events (e) can be arbitrary. Should (e) be nonrandom, then more refined object-inductivistic strategies may exist (as explained), but their players are assumed to be among the alternative players—nothing that concerns the behaviour of xMI hangs on that question. I also do not assume a fixed list of players—the list of players may vary from world to world, except that it always contains xMI and the (virtual) OI. I make the realistic assumption that xMI has finite computational means. On this reason I restrict my investigation to prediction games with *finitely* many players.

Prediction games are prima facie not interactive games in the *narrow* sense because the success rate of each player P is solely determined by 'nature' (the stream of events) and by P's own actions, that is, predictions. But prediction games are interactive in a *wider* sense, because the actions of the meta-inductivist depend on the actions of the non-MI-players, and the actions of certain alternative non-MI-players depend on MI's choice of favorites. Moreover, the collective meta-inductivists introduced in Section 8 provide a basis for game-theoretic effects in the narrow sense.

As far as I know, prediction games have so far not been studied in the philosophical literature. There are, however, two related approaches in related fields. In *formal learning theory* (cf. Kelly 1996) only *one* player,

an object-inductive scientist, plays against a stream of events, and it is investigated which cognitive tasks can reliably be achieved under which conditions on the stream of events. Concerning inductive prediction tasks the general result is negative, because of the possibility of 'demonic' streams of events which at every time n produce the opposite of the object-inductivist's prediction. This is just a variant of Hume's lesson.[1] In contrast, my prediction games consist of several prediction methods playing against each other, and my investigation does not focus on the reliability but on the optimality of methods. Even if for every meta-inductive prediction method there exist suitably chosen 'demonic' stream of events for which its predictive success is zero, such a method may still be optimal, provided one can prove that in all 'demonic' cases also all other methods which are accessible to the meta-inductive method must have zero success. This will be a consequence of my theorems.

A second field which comes close to my approach, although it has not been related to the problem of induction, is the theory of *universal prediction* (cf. Merhav and Feder 1998), more precisely its *nonprobabilistic variant*, which has been developed by mathematical learning theorists (for an overview, cf. Cesa-Bianchi and Lugosi 2006). In this account, one considers *online predictions based on expert advice*: a forecaster (who corresponds to our meta-inductivist) predicts an arbitrary event sequence based on the predictions of a set of experts (who correspond to our 'non-MI-players'). One speaks of 'universal' prediction theory because, as in our approach, the event-sequence may be arbitrary; and the setting is called 'online learning' because the players have simultaneously to learn from past events and to make new predictions. In Section 7 we will make use of a central theorem attained in this field.

4. The Simple Meta-inductivist and Its Limitation: Convergent Oscillations. I have investigated the prediction game by means of *mathematical analysis*, supported by *computer simulations* (programmed in Python24 together with Eckhart Arnold). The performance of a type of meta-inductivist has always two sides: (i) its *long-run* behavior, which is of central significance; and (ii) its *short-run* performance, which is also important: although one should be willing to buy *some* short-run losses of a prediction method for

1. In formal learning theory one considers mainly hypotheses evaluation tasks which are not considered here. Kelly's major result about prediction tasks is the following: an infinite stream of events is correctly predictable by a scientific method (a method whose predictions are a function of past events) after some finite time if and only if this infinite data stream is among a recursively enumerable set of possible data streams (1996, 260ff.).

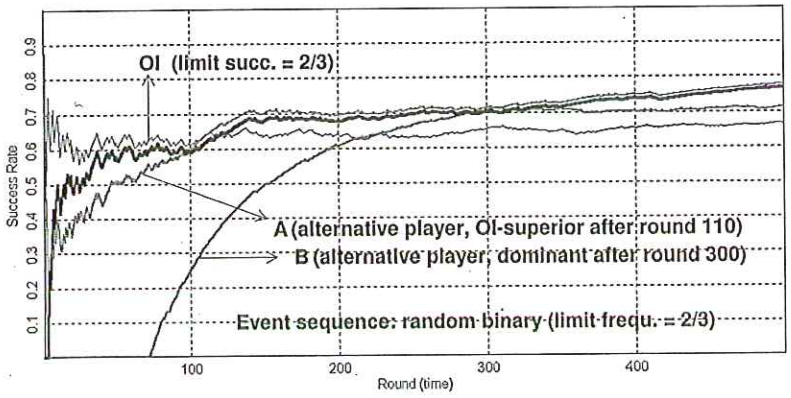

Figure 1. MI, OI, and two alternative players A and B.

the sake of its long-term optimality, these short-run losses should not be too large, and they should be under rational control.

In this section, I investigate the performance of the simple MI, which was defined in the previous section. MI belongs to the class of so-called *one-favorite* meta-inductivists, which choose at each time n a non-MI-player as their favorite for the next time and predict what their favorite predicts. For one-favorite meta-inductivists, binary prediction games are a *subclass* of real-valued prediction games. We obtain them by restricting (*a*) event sequences to those that contain only binary events, and (*b*) non-MI-players to those that predict only binary values. This implies that the meta-inductivist also delivers a binary prediction, because she predicts what her favorite predicts. Therefore, our theorems about one-favorite meta-inductivists (Theorems 1–3) apply to real-valued as well as to binary prediction games.

Figure 1 illustrates the behavior of MI at hand of a computer simulation of MI in a binary prediction game with OI and two alternative players: MI (boldface) goes always with the best player, which changes from OI to A to B.

From now on, maxsuc_n denotes the maximal success rate of the non-MI-players at time n. The set of non-MI-players is said to contain a (unique) *best* player $B \in \{P_1, \ldots, P_m\}$ iff there exists a time point n_B such that for all later times B's success rate is greater than the success rate of all other non-MI-players. n_B is called B's *winning* time. The central result about MI is Theorem 1.1, which tells us that MI predicts long-run optimal whenever there exists a best non-MI-player.

Theorem 1. For each prediction game $((e), \{P_1, \ldots, P_m, \mathrm{MI}\})$ whose player-set contains a best player B, the following hold:

1.1. Long run: MI's success rate approximates the maximal success of the non-MI-players (from below): $\lim_{n\to\infty}(\mathrm{maxsuc}_n - \mathrm{suc}_n(\mathrm{MI})) = 0$.

1.2. Short run: $(\forall n \geq 1:)\, \mathrm{suc}_n(\mathrm{MI}) \geq \mathrm{maxsuc}_n - (n_B/n)$, where n_B is B's winning time.

The proof of Theorem 1 is obvious and just explained informally: after the time point n_B MI's favorite will be B forever, but until time n_B MI's success may be zero in the worst case, due to switching favorites (see below). This yields Theorem 1.2, and Theorem 1.1 follows.

Theorem 1.2 informs us about the maximal short-run loss of MI. Since the time point n_B may come arbitrarily late, MI's cumulative short run loss may be arbitrarily high. Nevertheless, the result of Theorem 1.2 is at least *something*, because it shows that a high short-run loss of MI is caused by a late arrival of B's winning time. In conclusion, MI's optimality is restricted to prediction games that contain a best player whose winning time doesn't occur to late.

When I started my research on meta-induction, Brain Skyrms and Nicholas Rescher have pointed out to me that in determining her favorites, the meta-inductivist must buy some losses, compared to the best non-MI-method. These losses result from the fact that in order to predict for time $n + 1$, the meta-inductivist can only take into account the non-MI-players' success rates until time n. Whenever MI recognizes that her present favorite P_1 has lost one point compared to some new best player P_2, then MI has also lost this one point compared to P_2, before MI decides to switch to P_2. So for each switch of favorites MI loses a score of 1 in the binary prediction game, and a nonzero score ≤ 1 in the real-valued game, compared to the best non-MI-player. These losses may accumulate. The assumption of Theorem 1 excludes that MI can have more than finitely many losses due to switching favorites; so these losses must vanish in the limit.

The assumption of Theorem 1 is violated whenever the success rates of two or more leading non-MI-players oscillate endlessly around each other. There exist two sorts of success-oscillations: *convergent* oscillations (this section) and *nonconvergent* oscillations (next section). Convergent oscillations are given when two (or more) leading players oscillate in their success-rate around each other with constant period and with diminishing amplitude; that is, their success-difference converges against zero. Here MI loses one success point in every half oscillation period. In the *worst* case, two alternative players A and B oscillate around each other with

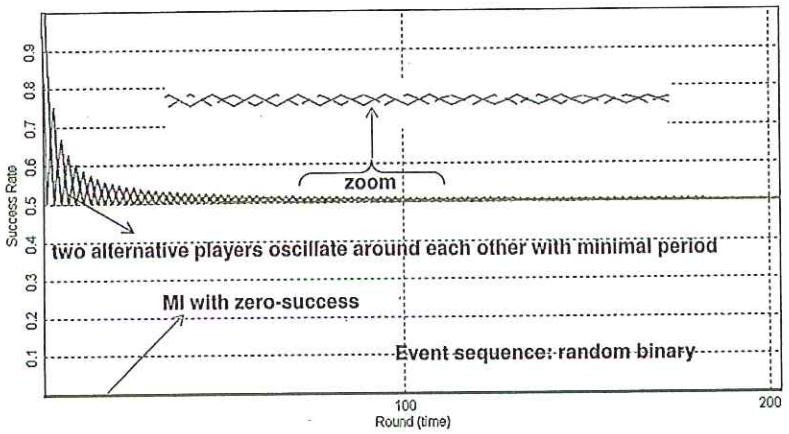

Figure 2. MI against two best alternative players in convergent oscillation.

the smallest possible period of four time units as follows, where MI's favorite is *underlined*:

A-scores iterated: . . . |0 0 1 1| . . .
B-scores iterated: . . . |1 1 0 0| . . .

This is a first example of a *deception* of the meta-inductivist by her favorites: the alternative players predict incorrectly exactly when they are in the position of being MI's favorite. In the result, the success rates of the two alternative players converges against .5, while the meta-inductivist's success remains zero for all time. A computer simulation of this scenario is shown in Figure 2. The object-inductivist OI has been omitted in this scenario, but its addition cannot avoid MI's breakdown.[2]

5. The ε-Meta-inductivist and Its Limitations: Systematic Deceivers. The meta-inductivist has a simple and robust defense strategy to permanent losses causes by convergent oscillations: don't switch favorites if their success difference is *practically insignificant*. I call this new type of meta-inductivist the *ε-meta-inductivist* εMI: εMI switches his favorite only if the success difference between his present favorite and a new better favorite exceeds a small threshold ε that is considered as practically insig-

2. If we add an OI with limit success .5 and assume that the limit success of the two oscillating players is slightly greater than .5, then MI's limit success will be slightly greater than zero. Alternatively, we may assume a 'demonic' event sequence which permanently deceives OI (recall Section 3); then the addition of OI wouldn't change anything in the result of Figure 2.

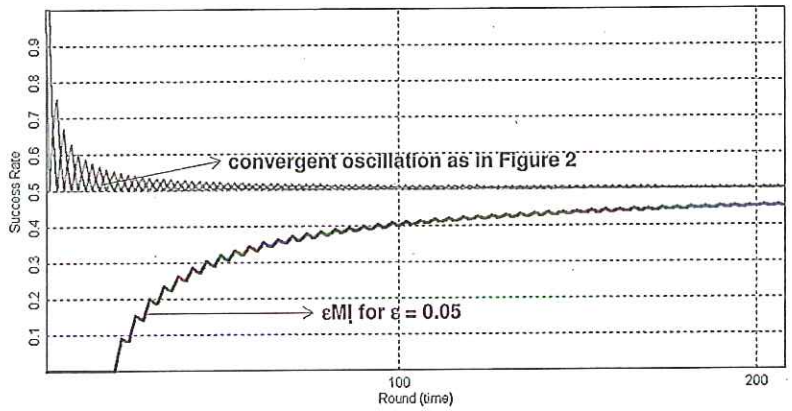

Figure 3. εMI in the convergent oscillation scenario of Figure 2.

nificant. Thus, $\text{fav}_{n+1}(\varepsilon\text{MI})$ = the leftmost player P in $\{P_1, \ldots, P_m\}$ such that $\text{suc}_n(P) > \text{fav}_n(\varepsilon\text{MI}) + \varepsilon$ if such a player exists; else $\text{fav}_{n+1}(\varepsilon\text{MI}) = \text{fav}_n(\varepsilon\text{MI})$. The performance of εMI is illustrated by the computer simulation in Figure 3, in which εMI plays against the two alternative players of the convergent oscillation scenario of Figure 2: when the success difference between the alternative players has become smaller than ε, εMI stops to switch but sticks to one player, with the result that εMI's success rate recovers and ε-approximates the maximum success of the two alternative players.

The move from MI to εMI gives rise to a stronger theorem than Theorem 1.1, namely Theorem 2.1. We say that a prediction game contains a subset $BP \subseteq \{P_1, \ldots, P_m\}$ of *ε-best* non-MI-players iff there exists a time n_{BP}, the *winning* time of BP, such that for all times $n \geq n_{BP}$, (*a*) each player in BP is more successful than each non-MI-player outside BP ($\forall P \in BP, \forall Q \in (\Pi - BP)$:$\text{suc}_n(P) > \text{suc}_n(Q)$), and (*b*) the successes of all BP-players are ε-close to each other ($\forall P \neq P' \in BP$:$|\text{suc}_n(P) - \text{suc}_n(P')| \leq \varepsilon$). Theorem 2.1 establishes that εMI ε-approximates the maximal success rate if there exists a subset of ε-best non-MI-players. Note that this condition does not imply, but is implied by the convergence of the non-MI-players' success rates towards a limit.

Theorem 2. For every prediction game $((e), \{P_1, \ldots, P_m, \varepsilon\text{MI}\})$ whose non-MI-players set contains a subset BP of ε-best players, the following hold:

2.1. Long run: εMI *ε-approximates* the maximal success of the non-MI-players (from below): $\lim_{n\to\infty}(\text{maxsuc}_n - \text{suc}_n(\varepsilon MI)) \leq \varepsilon$.

2.2 Short run: $(\forall n \geq 1{:})\ \mathrm{suc}_n(\mathrm{MI}) \geq \mathrm{maxsuc}_n - (n_{BP} + 1)/n - 2 \cdot \varepsilon$, where n_{BP} is *BP*'s winning time.

Proof (Theorem 2.1): After time n_{BP} at most *one* switch of εMI's favorite can occur. For if such a switch occurs at some time $s \geq n_{BP}$, then εMI's new favorite A will be an ε-best player in *BP*, and the definition of *BP* entails that after time n_{BP} the success-rate of no non-MI-player can exceed A's success rate by more than ε; whence εMI will favor A forever. It follows that εMI's success converges against the success of A from below (the initial losses vanish in the limit), and since A's success lies by at most ε below the maximal success of the other non-εMI-players, εMI's success ε-approximates this maximal success from below.

Proof (Theorem 2.2): Based on the argument for Theorem 2.1 we assume that s is the time of εMI's last switch of favorites and A is εMI's last favorite. There are two cases to consider:

Case 1, $s \leq n_{BP}$: The absolute success of εMI until time n_{BP} is zero in the worst case, because εMI's may be deceived by alternative players whose success rates oscillate around each other with amplitudes $> \varepsilon$ (see the example below). Thus $\mathrm{suc}_n(\varepsilon\mathrm{MI}) \geq \mathrm{suc}_n(A) - (n_{BP}/n)$, and (by definition of *BP*) $\forall n \geq n_{BP}{:}\mathrm{suc}_n(A) \geq \mathrm{maxsuc}_n - \varepsilon$. It follows that $\forall n \geq 1{:}\mathrm{suc}_n(\varepsilon\mathrm{MI}) \geq \mathrm{maxsuc}_n - (n_{BP}/n) - \varepsilon$.

Case 2, $s > n_{BP}$: Between times n_{BP} and s, εMI has only one favorite; we call him B. By reasoning as in Case 1, we obtain:

1. $\forall n$ with $n_{BP} < n \leq s{:}\mathrm{suc}_n(\varepsilon\mathrm{MI}) \geq \mathrm{suc}_n(B) - (n_{BP}/n)$, and hence
2. $\forall n$ with $n_{BP} < n \leq s{:}\mathrm{suc}_n(\varepsilon\mathrm{MI}) \geq \mathrm{maxsuc}_n - (n_{BP}/n) - \varepsilon$.

At time s, A becomes εMI's new favorite. Hence (i) $\mathrm{suc}_{s-1}(A) \leq \mathrm{suc}_{s-1}(B) + \varepsilon$, or in terms of absolute success, (ii) $a_{s-1}(A) \leq a_{s-1}(B) + \varepsilon \cdot (s - 1)$. At time s, A has earned at most score 1 (in the binary case 1) and B less than score 1 (in the binary case 0). Together with (ii) this gives us an upper bound for $a_s(A)$.

3. $a_s(A) \leq a_s(B) + \varepsilon \cdot (s - 1) + 1$.

It follows from 1 and 3 that

4. $\mathrm{suc}_s(\varepsilon\mathrm{MI}) \geq \mathrm{suc}_s(A) - ((n_{BP} + 1)/s) - \varepsilon \cdot ((s - 1)/s)$.

From time $s + 1$ onward, εMI earns the same scores as A. Therefore and by 4 we obtain

5. $\forall n > s$, $\mathrm{suc}_n(\varepsilon\mathrm{MI}) \geq \mathrm{suc}_n(A) - ((n_{BP} + 1)/n) - \varepsilon \cdot ((s - 1)/n)$.

Since $\forall n > s$, $\mathrm{suc}_n(A) \geq \mathrm{maxsuc}_n - \varepsilon$, 5 gives us

6. $\forall n > s$, $\text{suc}_n(\varepsilon\text{MI}) \geq \text{maxsuc}_n - ((n_{BP} + 1)/n) - \varepsilon \cdot ((n + s - 1)/n)$.

Because $(n + s - 1)/n$ is less than 2 in the worst case, it follows from 6 that

7. $\forall n > s$, $\text{suc}_n(\varepsilon\text{MI}) \geq \text{maxsuc}_n - ((n_{BP} + 1)/n) - 2 \cdot \varepsilon$.

From 2 and 7 of Case 2 and the result of Case 1 we obtain the claim of Theorem 2.2. QED

The worst-case bound of εMI's short-run loss provided by Theorem 2.2 (namely, $((n_{BP} + 1)/n) - 2 \cdot \varepsilon$) is not particularly good. At least, Theorem 2.2 tells us that εMI's short run loss decreases with an early arrival of the winning time n_{BP} of the ε-best players. The reason why Theorem 2.2 assigns to εMI an additional worst-case loss of at most $2 \cdot \varepsilon$ is that we wanted εMI's worst-case loss to be independent from εMI's last switch time s which may occur arbitrarily much later than n_{BP}.

The εMI is long-run optimal in a broader class of possible worlds than MI, on the cost that its optimality is not strict but *approximate*. Is approximate optimality good enough to count as a justification? I think: *yes*, although this justification is a weaker one. Of course, an approximately optimal strategy may be strictly dominated in many or even all prediction games, but this loss compared to the best strategy is *always small*—smaller than ε. Moreover, for almost all practical purposes there exists a choice of ε which is small enough to count as *practically insignificant*,[3] and Theorem 2.1 holds for *all* choices of ε. So I conclude that concerning the *long run*, the approximate versions of optimality and dominance (cf. Section 9.2) are 'almost as good' as their strict counterparts. However, there exists a *trade-off* in respect to the short-run performance, since a small ε goes usually hand in hand with a large n_{BP} in Theorem 2.2. So, the freedom to make ε very small is limited by the interest in keeping the short-run loss small.

The assumption of a subset of ε-best players is violated in prediction games with *nonconvergent* success-oscillations. It is here where we find the *worst cases* for meta-induction. If the success rates of two or more leading alternative players oscillate around each other in a *nonconvergent* manner with a nondiminishing amplitude of $\delta > \varepsilon$, then εMI will be deceived. The minimal periods of such nonconvergent oscillations must grow exponentially in time. The worst case are systematic deceivers, who are assumed to *know* (e.g., by clairvoyance) whether the meta-inductivist

3. Exceptions may be found in predictions of *chaotic* systems (as pointed out by the referee).

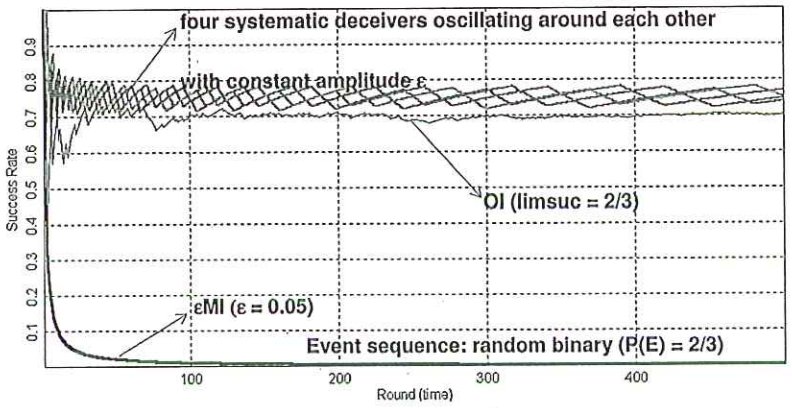

Figure 4. Systematic deception of εMI.

will choose them as favorite for the next time. They use this information to deceive the meta-inductivist by delivering a *worst* prediction (i.e., a prediction with minimal score) whenever the meta-inductivist chooses them as their favorite. The formal definition of systematic deceivers is simple: P is a *systematic deceiver* (of a meta-inductivist xMI) iff for all times $n \in |N^+$ the following holds: if P is xMI's favorite for time n, then P delivers a worst prediction for time n; otherwise, P predicts the right result for time n. Note that for natural loss functions, the worst prediction for time n is 0 if $e_n \geq .5$, and is 1 otherwise. Hence the score of the worst prediction is 0 in the binary prediction and a value between 0 and .5 in the real-valued prediction game; while the score of a correct prediction is always 1.

Two or more systematic deceivers will *always* oscillate around each other with an amplitude approximating ε from above, independent of εMI's choice of the significance threshold ε. Of course, a systematic deception strategy is partially at the cost of the deceiver's own success. But if εMI is playing against k deceivers, then at each time there will be $k - 1$ deceivers that predict correctly because they are not εMI's favorite. The computer simulation in Figure 4 shows a binary prediction game in which four alternative players deceive the ε-meta-inductivist. As long as a deceiver D_1 is εMI's favorite, D_1 predicts the wrong result until his success is more than ε below some deceiver D_2. At this time εMI switches his favorite from D_1 to D_2, D_1 starts to predict correctly and D_2 starts to predict wrong results, until the next switch of εMI occurs, etc. In this way, εMI's success rate is turned down to zero, while the mean success of the four deceivers per oscillation is .75.

The negative result of Figure 4 generalizes to all kinds of *one-favorite* meta-inductivists (recall their definition in Section 3): they must fail to be optimal whenever they play against $k \geq 2$ systematic deceivers (and use them as favorites), because in that case they have zero-success, while the deceivers will have a limit success of $(k-1)/k$, because in the long run each deceiver is εMI's favorite in 1 out of k times.

6. Deception Recording and the Avoidance Meta-inductivist. A natural reaction of the MI to systematic deception is deception-recording. The success-rate of a non-MI-player P *conditional* on times when P was xMI's favorite is abbreviated as $\mathrm{suc}_n(P|xMI)$ and defined the as the sum-of-scores of P's predictions for times $\leq n$ for which P was εMI's favorite, divided through the number of these times. As long the latter number is zero, we set by convention $\mathrm{suc}_n(P|\varepsilon MI) := 1$. A non-MI-player P (and the strategy played by P) is said to *deceive* (or to be a *deceiver*) *at time* n iff $\mathrm{suc}_n(P) - \mathrm{suc}_n(P|\varepsilon MI) > \varepsilon_d$, that is, difference between P's success-rate and P's favorite-conditional success-rate is practically insignificant. The value ε_d is the so-called *deception-threshold* ε_d. Our definition of a deceiver does not only capture systematic deceivers, whose favorite-conditional success rate is zero, but also more harmless kinds of deceivers whose favorite-conditional success rate drops below their unconditional success rate by more than ε_d. For the sake of the proof of the next theorem we have to require that $\varepsilon_d < \varepsilon$, that is, the deception threshold is strictly smaller than the switching threshold. Since ε_d can be as close to ε as one wants, this requirement is insignificant.

The avoidance meta-inductivist (aMI) is defined as follows: at any time, aMI chooses his favorite only among the nondeceivers at that time. aMI switches to a new favorite if either aMI's old favorite O has become a deceiver or O's success rate has dropped by more than ε below the maximal success rate of the nondeceivers (at the given time). To avoid unwanted deception recordings during the *initial phase* of the game when accidental success fluctuations are very high, we assume that the meta-inductivist starts the deception-recording of a non-MI-player P only when P has been a favorite for at least k_1 rounds, and a nonfavorite for at least k_2 rounds. In our computer simulations we have chosen $k_1 = k_2 = 5$. Note that even an object-strategy (such as OI) may become a deceiver, namely, when a 'demonic' stream of events deceives the object-strategy. Hence it is possible that the class of nondeceivers at a given time becomes empty; in such a case aMI predicts nothing and earns by convention score 0. The set of nondeceivers at time n is denoted by ND_n; and $\mathrm{maxsuc}(ND_n) := \max(\{suc_n(P): P \in ND_n\})$ is the maximal success rate among the nondeceivers at time n; by convention we set $\mathrm{maxsuc}(\varnothing) := 0$.

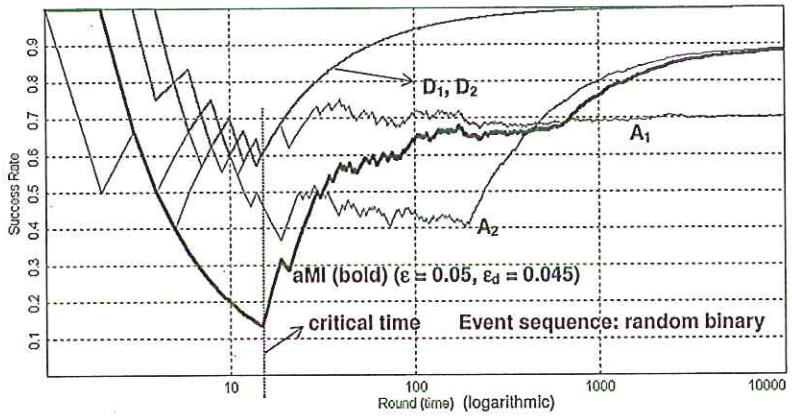

Figure 5. aMI against two deceivers (D_1, D_2) and two nondeceivers (A_1, A_2).

Of course, an avoidance meta-inductivist cannot predict optimally in regard to deceivers, because the deceivers will not be chosen as aMI-favorites and hence are free to predict better than the avoidance meta-inductivist and his favorites. What we expect is that aMI predicts optimally in regard to nondeceivers. This is confirmed by the computer simulation in Figure 5, in which aMI plays against two systematic deceivers D_1, D_2 and two permanent nondeceivers A_1, A_2 (logarithmic time scale). After the 'critical time' at which deception recording of aMI starts, the D_i are avoided as favorites by aMI, with the result that aMI goes along with the best player among the nondeceiving A_i's, and the D_i's success rate begins to approach the maximum value 1.

The result of the computer simulation in Figure 5 is generalized by Theorem 3, which establishes that in all prediction games aMI is indeed approximately long-run optimal with respect to nondeceivers.

Theorem 3.

3.1. Long run: For every prediction game $((e), \{P_1, \ldots, P_m, aMI\})$, aMI ε-approximates the maximal success on the nondeceiving non-MI-players (from below): $\lim_{n\to\infty}(\text{maxsuc}(ND_n) - \text{suc}_n(\text{aMI})) \leq \varepsilon$.

3.2. Short run: Let $n(i)$ be the latest time $\leq n$ such that $P_i = \text{fav}_{n(i)}(\text{aMI})$ (if it exists, else $n(i) := n$), and let $f_n(P)$ be the relative frequency of times $\leq n$ when P was aMI's favorite. Then $(\forall n \geq 1{:})\ \text{suc}_n(\text{aMI}) \geq (\sum_{1\leq i\leq m} \text{suc}_{n(i)}(P_i) \cdot f_n(P_i)) - \varepsilon_d - (m/n)$.

Proof (Theorem 3.1): The proof is an adaptation of the proof of Theorem 3 in Schurz 2008, Appendix. I abbreviate the latter theorem as th3*. The following adaptations have to be made: (*a*) The proof of th3* is formulated for binary prediction games, but it applies in the same way to the real-valued games, and (*b*) in the proof of th3*, it is assumed that the non-MI-players are ultimate non-deceivers (whence the claim of th3* applies to εMI instead of aMI). I found out that for aMI this restriction is unnecesessary. So the assumption of the existence of time point u in the first paragraph of the proof of th3* has to be dropped; all other proof steps remain intact.

Proof (Theorem 3.2): See the proof of Theorem 4.2 in Schurz 2008.

Different from εMI, aMI has also a good *short run* performance, because at any time aMI chooses her favorites only among nondeceivers. As Theorem 3.2 tells us, aMI's success rate is never more than $\varepsilon_d + (m/n)$ below the weighted average of the non-MI-players' success rates at the last time when they were aMI's favorite. Since the non-MI-players have an ε-approximately optimal success rate while being favorite, and since the term (m/n) quickly vanishes for $n \gg m$, this is a good short run result.

The fact that aMI does not predict optimally in regard to deceivers is certainly a drawback. For a player P who is recorded as a deceiver will be 'stigmatized' by aMI as a deceiver as long as P does not decrease his unconditional success rate (since P's aMI-conditional success rate is frozen as long as aMI does not favor P).

Let us take stock. So far we have discovered a variety of *one-favorite* meta-inductive strategies (MI, εMI, aMI) whose performance got successively improved. But none of these meta-inductivists' predictions are universally optimal because of the possibility of (systematic) deceivers. Are there meta-inductivist strategies which are indeed universally optimal?

Meta-induction is a research program: one may invent various kinds of meta-inductive strategies. For example, one may construct refined one-favorite meta-inductivists who use very badly predicting players P *inversely*, by predicting the opposite of what P predicts. But this move would be useless against systematic deceivers because whenever the meta-inductivist decides to use a systematic deceiver inversely, the deceiver would know this and deliver a correct prediction. This is just an instance of our general negative result that no one-favorite meta-inductivist can cope with systematic deceivers. If there exist better meta-inductive strategies, they must be found in the class of *multiple-favorite* meta-inductivists. In the

next section we investigate their most important variant: weighted-average meta-inductivists.

7. Weighted-Average Meta-induction for Real-Valued Prediction Games. A weighted-average meta-inductivist predicts a weighted average of the predictions of the non-MI-players, weighted by their 'attractiveness'. The weighted average of several predictions of zeros and ones is a real value between 0 and 1, rather than a yes-or-no prediction. Therefore, this method cannot be applied to binary prediction games, in which all predictions must be either '0' or '1'—it can only be applied to real-valued prediction games. In this form, the method of weighted-average prediction has been studied in the theory of (nonprobabilistic) *universal prediction,* which was mentioned in Section 3. The results in this literature have not at all been related to the problem of induction, but the problem setting is similar to my prediction games. The results established in the previous section are not covered by the theorems established in this area; rather, for the binary and nonprobabilistic setting this research provides a merely negative result (cf. Cesa-Bianchi and Lugosi 2006, 67).

The weighted-average meta-inductivist (wMI) is defined as follows. For every non-MI-player P we define $at_n(P) := \mathrm{suc}_n(P) - \mathrm{suc}_n(\mathrm{wMI})$ as P's *attractiveness* (as a favorite) at time n. Let $PP(n)$ be the set of all non-MI-players with *positive* attractiveness at time n. Then wMI's prediction for time 1 is set to .5, and for all times > 1 with nonempty $PP(n) \neq \varnothing$ it is defined as follows.

Definition. Weighted-Average Prediction.

$$\forall n \geq 1: p_{n+1}(\mathrm{wMI}) = \frac{\sum_{P \in PP(n)} \mathrm{at}_n(P) \cdot p_{n+1}(P)}{\sum_{P \in PP(n)} \mathrm{at}_n(P)}$$

In words, wMI's prediction for the next round is the attractiveness weighted average of the attractive players' predictions for the next round. Should it happen that $PP(n)$ gets empty, $p_{n+1}(\mathrm{wMI})$ is reset to .5.

Informally explained, the reason why wMI cannot be deceived is the following. A non-MI-player who tries to deceive wMI would be one who starts to predict incorrectly as soon as his attractiveness for wMI is higher than a certain threshold. The success rates of such wMI-adversaries must oscillate around each other. But wMI does not favor just one of them (who predicts incorrectly in turn), but wMI predicts according to an attractiveness weighted average of correctly and incorrectly predicting adversaries, and therefore wMI's long-run success must approximate the maximal long-run success of his adversaries. Note that a 'demonic' event sequence which minimizes wMI's score at each time cannot change this

result, because a low success rate of wMI goes hand in hand with a low success rates of the attractive non-MI-players.

Theorem 4 does not hold for arbitrary but only for those loss functions $l(p_n, e_n)$ which are *convex* in p_n. By definition, $l(p_n, e_n)$ is convex in its argument p_n iff for all values of $e_n \in [0,1]$ and for every weight $\gamma \in [0,1]$ and two possible predictions $a < b \in [0,1]$ the following holds: $l((\gamma \cdot a + (1-\gamma) \cdot b),\ e_n) \leq \gamma \cdot l(a, e_n) + (1-\gamma) \cdot l(b, e_n)$. In words, the loss of a weighted average of two predictions is smaller-equal than the weighted average of the losses of the two predictions. It is easy to see that the natural loss-function $l(p_n, e_n) := |p_n - e_n|$ is convex. There exist many other convex loss-functions, e.g., $|p_n - e_n|^q$ for $q \geq 1$, and Theorem 4 applies to all of them. Theorem 4.1 establishes that wMI is indeed a *universally* long-run optimal prediction strategy, even in the strict (and not approximate) sense. Also wMI's short-run performance is good, as Theorem 4.2 reveals. Since identically predicting non-MI-players can be identified in a prediction game, the number m corresponds to the number of alternative accessible prediction strategies, which is under complete control and not too high, and the worst-case short-run loss $\sqrt{m/n}$ will quickly vanish for times $n \gg m$.

Theorem 4. For every prediction game $((e), \{P_1, \ldots, P_m, \text{wMI}\})$ whose loss-function $l(p_n, e_n)$ is *convex* in the argument p_n, the following hold:

4.1. Long run: $\text{suc}_n(\text{wMI})$ (strictly) approximates the non-MI-players' maximal success: $\lim_{n\to\infty}(\text{maxsuc}_n - \text{suc}_n(\text{wMI})) = 0$.
4.2. Short run: $(\forall n \geq 1{:})\ \text{suc}_n(\text{wMI}) \geq \text{maxsuc}_n - \sqrt{m/n}$.

Proof: We identify wMI with the polynomially weighted average forecaster F described in Cesa-Bianchi and Lugosi (2006, 12) with its parameter p set to 2. The non-MI-players $\{P_1, \ldots, P_m\}$ are identified with the N 'experts' (i.e., $N = m$). Cesa-Bianchi and Lugosi define the predictions of F for the special case $p = 2$ as follows:

$$p_{n+1}(F) = \frac{\sum_{P \in PP(n)} (L_n(F) - L_n(P)) \cdot p_{n+1}(P)}{\sum_{P \in PP(n)} (L_n(F) - L_n(P))}, \qquad (1)$$

where $L_n(P) := \sum_{1 \leq i \leq n} l(p_i(P), e_i)$ is the *cumulative* loss of a player P until time n, which equals $n - a_n(P)$ (by our definitions in Section 1). Hence $a_n(P) = n - L_n(P)$, and so (1) implies

$$p_{n+1}(F) = \frac{\sum_{P \in PP(n)} (a_n(P) - a_n(F)) \cdot p_{n+1}(P)}{\sum_{P \in PP(n)} (a_n(P) - a_n(F))}$$

$$= \frac{n \cdot \sum_{P \in PP(n)} (\mathrm{suc}_n(P) - \mathrm{suc}_n(F)) \cdot p_n(P)}{n \cdot \sum_{P \in PP(n)} (\mathrm{suc}_n(P) - \mathrm{suc}_n(F))} \qquad (2)$$

$$= \frac{\sum_{P \in PP(n)} at_n(P) \cdot p_n(P)}{\sum_{P \in PP(n)} at_n(P)},$$

which means that F's predictions coincide with the predictions of wMI, and $\mathrm{suc}_n(F) = \mathrm{suc}_n(\mathrm{wMI})$. Corollary 2.1 of Cesa-Bianchi and Lugosi (2006, 12–13) establishes for F's short run loss the following constraint (setting $p := 2$):

$$L_n(F) - \min(\{L_n(P_i) : 1 \leq i \leq m\}) \leq \sqrt{m \cdot n}, \qquad (3)$$

hence

$$a_n(F) \geq \max(\{a_n(P_i) : 1 \leq i \leq m\}) - \sqrt{m \cdot n}, \qquad (4)$$

and so (by dividing through n)

$$\mathrm{suc}_n(F) = \mathrm{suc}_n(\mathrm{cwMI}) \geq \max(\{\mathrm{suc}_n(P_i) : 1 \leq i \leq m\}) - \sqrt{m/n}. \qquad (5)$$

(5) is the claim of Theorem 3.2; and Theorem 3.1 is an immediate consequence of it. QED

In prediction games satisfying the conditions of Theorem 1, the strategy wMI will soon converge to the simple MI-strategy, since after some time, only the best player will have positive attractiveness, hence wMI will predict as if she would favor this best player forever. Only if the success rates of the non-MI-players oscillate around each other forever, wMI will go on to average the predictions of several players forever.

8. Collective Weighted-Average Meta-induction for Binary Prediction Games. Theorem 4 does not apply to binary prediction games, because even under the assumption that the events and the non-MI-player's predictions are binary, wMI's predictions are not binary. One could introduce a 'digitalized' loss function over wMI's real-valued predictions which models the loss of binary predictions (e.g., by interpreting $p_n \geq 0.5$ as '1' and $p_n < 0.5$ as '0'), but this digitalized function would no longer be convex. The failure of Theorem 4 for binary prediction games can be recognized from the following example by Cesa-Bianchi and Lugosi (2006, 67): assume a meta-inductivist playing against two non-MI-players, one of them constantly predicting 1 and the other constantly predicting 0, and a 'demonic' event-sequence producing constantly the opposite of the meta-inductivist's predictions. Then whatever the meta-inductivist predicts, his success rate will be constantly zero, while the maximal success rate of the two non-MI-players must always be $\geq .5$. Thus, we have obtained a second

general negative result: a meta-inductive strategy which predicts long-run optimal for an *individual* player in arbitrary *binary* prediction games does not exist. In combination with Theorem 4 this is a philosophically deep result, insofar as it shows that a continuous (real-valued) nature is more friendly to the inductivist than a discrete (digitalized) nature.

Nevertheless I have found a way to apply Theorem 4 indirectly also to the prediction of binary events, namely by means of assuming a *collective* of k meta-inductivists, abbreviated as $\text{cwMI}_1, \ldots, \text{cwMI}_k$, and by considering their *mean success rate* ('cwMI_i' stands for 'collective weighted-average meta-inductivist number i'). I regard wMI's real-valued prediction as an *ideal* (though nonadmissible) prediction, which is approximated by the mean value of the k binary predictions of the collective of cwMI-meta-inductivists as follows: for $1 \leq i \leq k$, cwMI_i predicts 1 if $i \leq [p_n \cdot k]$, and else 0, where $[x]$ is the integer-rounding of the real number x. Thus, $[p_n \cdot k]$ cwMI's predict 1 and $k - [p_n \cdot k]$ cwMI's predict 0. In this way, one obtains a universal optimality result for the *mean success rate* of collective of cwMI's, abbreviated as $\overline{\text{suc}}_n(\text{cwMI})$, which is formulated in Theorem 5. Compared to Theorem 4, the upper bound of cwMI's mean success rate is decreased by the additional term $1/(2 \cdot k)$, which reflects the maximal loss due to approximation of the ideal prediction by k binary predictions. This additional loss can be made arbitrarily small by increasing the number of meta-inductivists.

Theorem 5. For every binary prediction game $((e), \{P_1, \ldots, P_m, \text{cwMI}_1, \ldots, \text{cwMI}_k\})$:

5.1. Long run: $\overline{\text{suc}}_n(\text{cwMI})$ $1/(2 \cdot k)$-approximates the non-MI-players' maximal success: $\lim_{n\to\infty}(\overline{\text{suc}}_n(\text{cwMI}) - \text{maxsuc}_n) \leq 1/(2 \cdot k)$.

5.2. Short run: $(\forall n \geq 1:)\ \text{suc}_n(\text{cwMI}) \geq \text{maxsuc}_n - \sqrt{m/n} - 1/(2 \cdot k)$.

Proof: At every time i $(1 \leq i \leq n)$ the actually achieved mean score is $[p_n \cdot k]/k$ if $e_i = 1$, and $1 - [p_n \cdot k]/k$ if $e_i = 0$; while the ideally achieved score is p_n if $e_i = 1$ and $1 - p_n$ if $e_i = 0$. Since $|[p_n \cdot k] - p_n \cdot k| \leq 0.5$, it follows that at every time, the actually achieved mean score deviates from the ideally achieved score by at most $1/(2 \cdot k)$. Hence we obtain for the mean success rate: $\overline{\text{suc}}_n(\text{cwMI}) \geq \text{suc}_n(\text{wMI}) - 1/(2 \cdot k)$. From this and Theorem 4, the claims of Theorem 5 follow. QED

Figure 6 shows a computer simulation of a collective of ten cwMI's playing against 4 specially designed cwMI-adversaries, who predict incorrectly as soon as their attractiveness gets higher than a variable threshold.

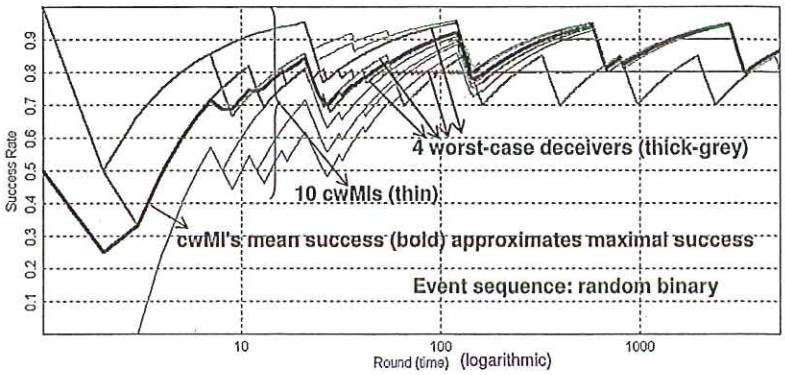

Figure 6. Ten collective weighted-average meta-inductivists against four cwMI-adversaries.

The relation of the cwMI-strategy to the situation of an *individual* meta-inductivist in binary prediction games is the following. The cwMI-adversaries may conspire against a particular individual, say against cwMI_3, and constantly deceive cwMI_3 (alternatively, a 'demonic' event sequence may constantly deceive cwMI_3). But the cwMI-adversaries cannot deceive the other cwMI's at the same time, and their anti-cwMI_3-conspiration will not affect the optimality result for the cwMI's mean success. Moreover, if we assume that the cwMI's *share* their mean success, we find here a basis for interactive utility effects in the narrow game-theoretical sense.

9. Epistemological Conclusions.

9.1. Refined Inductive Strategies. Besides the *simple* object-inductive and meta-inductive strategies, there are refined versions of them, which I call their *conditionalized* versions. They exploit correlations in nonrandom worlds which obtain between the events e_n and prior events, internal or external to the sequence (e), with help of Reichenbach's principle of the *narrowest reference class* (1949, sec. 72). Here is an illustration of this principle: assume $\{R_1, \ldots, R_r\}$ is a partition of events prior to time n, described in terms of nomological predicates, such that the given person X has reliable information about which cell R_i was realized before time n, and the cells are statistically relevant for the events of the sequence E, that is, $\bar{e}_n|R_i \neq \bar{e}_n|R_j$ for $j \neq i$, where $\bar{e}_n|R_i$ is the R_i-conditionalized mean value of e up to time n. Let $R{:}|N \rightarrow \{R_1, \ldots, R_r\}$ be the function which assigns to each time n the cell $R(n)$ of the partition which was realized before time n. Then the conditionalized OI-strategy transfers the condi-

tionalized mean value $\bar{e}_n|R(n)$, and in the binary case its integer-rounding $[\bar{e}_n|R(n)]$, to the next time (note that $[\bar{e}_n|R(n)]$ coincides with the conditional frequency $f_n(E|R(n))$. Provided that all involved mean values and frequencies converge to a limit, one can prove that conditionalizing to reference partitions may only *improve* the success, compared to the simple OI (cf. Good 1983, Chapter 17).

Also the conditionalized meta-inductivist works with a reference partition, but she conditionalizes the success frequencies of the other players to the cells of this partition. While a simple xMI ε-approximates the maximal success always from below, the success rate of a conditionalized xMI may be even strictly greater than the success rates of all other players. This fact does not affect the approximate optimality of the simple xMI, because we assume that refined meta-inductivistic techniques, if they are accessible, are among the methods of the alternative players (hence, with a 'non-xMI-player' we mean a 'non-simple-xMI-player'). However, the fact shows that a simple meta-inductivist can *improve* his results by getting access to refined meta-inductivist (or object-inductivist) techniques. The assumption that the simple xMI's may favor other meta-strategies is philosophically important, because it blocks an infinite regress of meta-levels.[4]

Of course, a scientific meta-inductivist will not be satisfied to have mere output-access to successful prediction strategies; she will also try to understand these strategies and gain *internal* access to them, because a meta-inductivist who can simulate a prediction strategy no longer depends on the accessibility of other players. Internal access to refined prediction strategies is typically provided by scientific *theories*. By way of comparison: a conditionalized object-inductivist behaves like a Kuhnian scientist, who is fixed on his own theory about efficient reference partitions. In contrast, a conditionalized meta-inductivist fits with the Popperian picture of the open-minded scientist: she, too, develops her own theories (object-inductivistic strategies), but she does not insist on her theory, but recommends another theory should it turn out to be more successful than her own theory.

9.2. Approximate Optimality and Dominance. In the standard decision-theoretic framework, an action A is called (strictly) *optimal* in a class of available actions **A** iff in every possible world w, the utility of A in w is greater-or-equal the utility of all other actions in **A**; and A is called *dominant* in **A** iff A but no other action B is optimal in **A** (cf. Weibull 1995, 13). In order to transfer these notions to prediction games, we must respect

4. A circular situation in which the xMI wants to favor another meta-player who wants to favor him simply leads to the effect that both players are not accessible to each other.

TABLE 1. LONG-RUN OPTIMALITY OF xMI W.R.T. Σ IN W.

xMI-Strategy	Kind of Optimality	Σ Comprises the Following Strategies	Worlds in W Contain Finitely Many xMI-Accessible Players Satisfying
MI (Theorem 1)	Strict	All (in world)	$\exists$ best non-MI-player
εMI (Theorem 2)	ε	All	$\exists$ set of ε-best non-MI-players
aMI (Theorem 3)	ε	Nondeceiving	No condition
wMI (Theorem 4)	Strict	All	Real-valued game
cwMI (Theorem 5)	$1/(2 \cdot k)$	All	Binary game

the following two differences: *first*, we explicate the more general notion of (approximate) long-run optimality, and *second*, the class of available actions **A** is the class of all prediction strategies which are *accessible* to a given person (or group of persons) X, and this class must be allowed to *vary* from world to world. In order to express *differentiated* results, we relativize the notion of approximate optimality to a class of strategies Σ w.r.t. that xMI is optimal, and to a class of worlds **W** in which all strategies in Σ and possibly some other strategies are played.

> **Definition.** Approximate Optimality: A prediction strategy S is δ-*optimal* for a person [or group of persons] X w.r.t. a set of strategies Σ in a class of worlds **W** iff for every world in **W** in which X play(s) S the following holds: the predictive success (the mean predictive success, respectively) of X δ-approximates the maximal predictive success of all players playing strategies in Σ.

We choose 'δ' instead of 'ε' in this definition because 'δ' has a variable meaning: it can be 0 (MI, wMI), ε (εMI, aMI) or $1/(2 \cdot k)$ (cwMI); when $\delta = 0$ we speak of *strict* optimality. The long-run optimality results of Theorems 1–5 are summarized in Table 1.

The (ε-)optimality of a prediction strategy is only a *weak* best-alternative justification of it, because it is compatible with the existence of other (ε-)optimal prediction strategies. A stronger justification would be (ε-)dominance, defined as follows:

> **Definition.** Approximate Dominance: A prediction strategy S is δ-*dominant* for a person (or group of persons) X w.r.t. a set of strategies Σ in a class of worlds **W** iff (*a*) S is δ-optimal for X w.r.t. Σ in **W**, and (*b*) no other prediction strategy S^* ($\neq S$) in Σ is δ-optimal for X w.r.t. Σ in **W**.

We cannot show that the meta-inductivists of Table 1 are generally (ε-)dominant in their respective classes of players and worlds, on two reasons. First, there exist the explained *conditionalized* xMI-strategies, who may even improve the simple xMI-strategies (though never more than ε given they are accessible). Second, one may introduce *mixed* strategies

which use noninductive clairvoyance strategies only in 'paranormal' worlds in which these strategies are superior, while they use object-inductive strategies in 'normal' worlds. Thereby, a *normal world* is defined as a world in which *only* object-inductive prediction strategies can have nonaccidental (long-run) predictive success. Also mixed strategies may predict (ε-)optimal. We can only show that meta-inductive strategies are (ε-)dominant w.r.t. all *noninductive* prediction strategies, which are by definition strategies which predict noninductive in at least one normal world. It follows from our definitions that noninductive prediction strategies will predict (ε-)suboptimal in at least one (normal) world, whence they cannot be (ε-)optimal. Thus we can summarize our result on δ-dominance as follows: if a meta-inductivist xMI is δ-optimal w.r.t. a strategy class $\mathbf{\Sigma}$ and in a world class $\mathbf{W}$ according to Table 1, then xMI is δ-dominant w.r.t. the subclass of noninductive strategies in $\mathbf{\Sigma}$ in the world-class $\mathbf{W}$.

This result about dominance may be criticized to be an almost trivial consequence of my definition of a 'normal world' and 'noninductive strategy'. I think the result is logically but not philosophically trivial. Independent from that question, Reichenbach (1949, 475–475) has pointed out that already the *optimality* argument may be considered as a sufficiently strong justification of meta-induction, insofar as meta-induction is the *only* prediction strategy for which optimality can be *rationally demonstrated.*

9.3. The Philosophical Nature of the Meta-inductivist: A Universal Learner. While one-favorite meta-inductive strategies are optimal only under certain restrictions, weighted-average meta-induction has turned out to be universally optimal. One may object that wMI is restricted to real-valued und cwMI to binary prediction games, but I think that playing a real-valued versus a binary prediction game is a matter of convention rather than a restriction on worlds. One may also criticize that for binary prediction games optimality is only guaranteed for the mean success of a collective of meta-inductivists cwMI, but I think that *collective optimality* is nevertheless an important kind of universal optimality.

In conclusion, I think the achieved optimality results on meta-induction are strong enough to show that a noncircular justification of (meta-)induction can be successful. This justification does not show that meta-induction must be successful (in a strict or probabilistic sense), but it *favors* the meta-inductivistic strategy against all other accessible competitors. This is sufficient for justificational purposes, without being in dissent with any of Hume's skeptical arguments. The given justification of meta-induction is mathematically-analytic (or 'a priori'), insofar it does not make any assumptions about the nature of the considered worlds except

from practically evident assumptions about prediction games, such that its players can perform calculations, can observe past events, and are free to decide. However, as we have explained in Section 2, this analytic justification of meta-induction implies an a posteriori justification of object-induction in our real word, because so far object-induction has turned out to be the most successful prediction strategy. This argument is *no longer circular*, *given* that we have a noncircular justification of meta-induction—and we have it.

The major advantage of the meta-inductivistic approach is its *radical openness* towards all kinds of possibilities. In my view, this radical openness is a sign of all *good* foundation-oriented (instead of 'foundationalistic') programs in epistemology. Unlike in Rescher's "initial justification" of induction (1980, 82), meta-induction does not exclude esoteric worldviews or prediction methods from the *start*. Such an a priori exclusion would prevent a constructive dialog between a scientific philosopher and an esoteric-minded person. Meta-induction takes all these possible worldviews initially seriously and argues: wherever the 'ultimate truth' lies, you should in any case employ meta-induction because it is universally optimal among all accessible prediction methods.

Many readers will still uphold skeptical reservations. They will ask: how can it *ever* be possible to prove that a strategy is optimal with respect to *every* other accessible strategy in *every* possible world—without assuming anything about the nature of alternative strategies and possible worlds? My heuristic answer to this *skeptical* challenge is as follows: this is possible for meta-inductive strategies because these strategies are *universal learners*: whenever they are confronted with a so far better strategy, they can imitate the better strategy (output-accessibility) or even learn to reproduce it (internal accessibility).

REFERENCES

Carnap, Rudolf (1947), "On the Application of Inductive Logic", *Philosophy and Phenomenological Research* 8: 133–147.

Cesa-Bianchi, Nicolo, and Gabor Lugosi (2006), *Prediction, Learning, and Games*. Cambridge: Cambridge University Press.

Good, Irving J. (1983), *Good Thinking*. Minneapolis: University of Minnesota Press.

Howson, Colin (2000), *Hume's Problem*. Oxford: Clarendon.

Kelly, Kevin T. (1996), *The Logic of Reliable Inquiry*. New York: Oxford University Press.

Merhav, Neri, and Meir Feder (1998), "Universal Prediction", *IEEE Transactions on Information Theory* 44: 2124–2147.

Reichenbach, Hans (1938), *Experience and Prediction*. Chicago: University of Chicago Press.

——— (1949), *The Theory of Probability*. Berkeley: University of California Press.

Rescher, Nicholas (1980), *Induction*. Pittsburgh: University of Pittsburgh Press.

Salmon, Wesley C. (1957), "Should We Attempt to Justify Induction?", *Philosophical Studies* 8: 45–47.

——— (1974), "The Pragmatic Justification of Induction", in Richard Swinburne (ed.), *The Justification of Induction*. Oxford: Oxford University Press, 85–97.

Schurz, Gerhard, (2008), "Meta-induction", in Clark Glymour, Wei Wang, and Dag Westerstahl (eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the Thirteenth International Congress*. London: King's College Publications.
Skyrms, Brian (1975), *Choice and Chance: An Introduction to Inductive Logic*. Encino, CA: Dickenson.
Weibull, Jörgen (1995), *Evolutionary Game Theory*. Cambridge, MA: MIT Press.