Incompatibility and the pessimistic induction: a challenge for selective realism

Florian J. Boge¹

Received: 21 April 2020 / Accepted: 26 March 2021 / Published online: 4 May 2021

© The Author(s) 2021

Abstract

Two powerful arguments have famously dominated the realism debate in philosophy of science: The No Miracles Argument (NMA) and the Pessimistic Meta-Induction (PMI). A standard response to the PMI is selective scientific realism (SSR), wherein only the working posits of a theory are considered worthy of doxastic commitment. Building on the recent debate over the NMA and the connections between the NMA and the PMI, I here consider a stronger inductive argument that poses a direct challenge for SSR: Because it is sometimes exactly the working posits which contradict each other, i.e., that which is directly responsible for empirical success, SSR cannot deliver a general explanation of scientific success.

Keywords Pessimistic induction · Base rate fallacy · Selective realism · Theory change

1 Introduction

Two powerful arguments have famously dominated the realism debate in philosophy of science: The No Miracles Argument (NMA) and the Pessimistic Meta-Induction (PMI).¹ The intuition underpinning the NMA has been colorfully stated by Worrall (2011, p. 11) as follows:

¹ Following Magnus (2010), I take Stanford's (2006) 'new induction' to be a separate, maybe deeper, issue, and will hence not comment on it any further below.

✉ Florian J. Boge
fjboge@uni-wuppertal.de

¹ Interdisciplinary Centre for Science and Technology Studies (IZWT), Wuppertal University, Gaußstr. 20, room S.11.19, 42119 Wuppertal, Germany

‘How on earth’, it seems unavoidable to ask, ‘could a theory score a dramatic predictive success [...] unless its claims about the reality ‘underlying’ the phenomena [...] are at least approximately in tune with the real underlying structure of the universe?’

In Putnam’s (1975a, p. 73) more straightforward words, scientific realism (SR) is “the only philosophy that doesn’t make the success of science a miracle.” But surely there are no miracles. So SR must be true: There *is* a mind-independent reality with definite nature and structure (metaphysical stance), which scientific theories *are capable* of making reference to (semantic stance), and of which our best scientific theories *are in fact* at least approximately true (epistemic stance).²

On the other hand, the PMI rests on an intuition that seems just as strong. In the (equally colorful) words of Poincaré (1905, p. 160):

The ephemeral nature of scientific theories takes by surprise the man of the world. Their brief period of prosperity ended, he sees them abandoned one after another; he sees ruins piled upon ruins; he predicts that the theories in fashion to-day will in a short time succumb in their turn, and he concludes that they are absolutely in vain.

A much discussed list of theories apparently in line with this intuition, “which could be extended ad nauseam”, has famously been offered by Laudan (p. 33, 1981). Like Laudan, Putnam (1978, pp. 24–5) took this kind of “historical gambit” (Laudan 1984, p. 157; *emph. omitted*) to have an impact on the purported *reference* of scientific terms. More precisely, he took it to support an inductive argument of the following form: “as no term used in the science of more than fifty (or whatever) years ago referred, so it will turn out that no term used now [...] refers.” (Putnam 1978, p. 25)

However, reference is a subtle issue. Schurz (2009, 2011), for instance, identifies the assimilation and release of phlogiston (‘(de)phlogistication’) as “[t]he theoretical expressions of phlogiston theory [...] which did all the empirically relevant work” (Schurz 2009, p. 108), and proves a structural correspondence theorem according to which these³ correspond “and hence *indirectly refer*” (*ibid.*, p. 109; *emph. added*) to the acceptance or donation of electrons by (ionized) atoms under specific conditions.

Now even if we accept this account of ‘indirect reference’, it is clear that the move from phlogiston to oxygen theory implies a radical shift in *ontology*. This is not denied by Schurz (2009, p. 120), who points out that his theorem saves the

²See Putnam (1975b), Boyd (1980), or Psillos (1999). I take it for granted here that all one can hope for is *approximate* truth; something endorsed, to my knowledge, by almost everyone in the debate these days. However, I refrain from offering (or buying into) a positive account of approximate truth (or its close cousin ‘verisimilitude’) here, because existing conceptions are vast: They range from, say, the ratio of true to total claims in a (conjunctive) theory (Cevolani and Schurz 2017) over the Kulback-Leibler divergence between model probabilities and the ‘true chance function’ (Sober 2002) to the fact that a theory’s numerical predictions break down at a certain scale (Williams 2019)—and I am not convinced that any single one is useful for my present purposes. See also Schurz (2017) for a recent overview.

³The theorem is fairly general though, and may be applied to other cases as well (see Schurz 2009, Section 5).

‘outer structure’ of a theoretical term φ like ‘dephlogistication’, i.e., “the causal relations between the complex entity φ and the empirical phenomena,” but not its ‘inner structure’, so “for example, that ‘dephlogistication’ is a process in which a special substance different from ordinary matter, called ‘phlogiston’, leaves the combusted substance.”

The general thinking expressed here is well in line with a standard response to the PMI-intuition, advanced, in different versions, by e.g. Psillos (1999), Kitcher (1993), Worrall (1989), and even Poincaré (1907, cf. p. 139): The endorsement of a *selective* attitude towards scientific realism (SSR), wherein only those parts responsible for empirical success—and in some way or other stable throughout theory change—are considered as representative of something real.

According to SSR, the approximate truth of a theory must be located in some select set of its posits, *directly* connected to empirical success. Furthermore, as these posits are closely connected to certain posits in *successor* theories, usually through some kind of weakening or generalization, it is possible for the SSRist to uphold a view of ‘convergence to the truth’; if only upon exclusion of a lot of theoretical ‘fluff’, if you will.

The question now arises whether *all* pessimistic inductive moves based on theory-failure thus stand refuted. I believe that they do not: In an extension of a recent argument against the NMA (Boge 2020), I will show that there is a pessimistic induction that threatens even SSR. The main result of this paper is the establishment of a thorough connection between the recent debate over the NMA (Henderson 2017; Dawid and Hartmann 2018; Boge 2020) and the ensuing debate about the PMI, which results in a stronger (though not in the same sense historical) challenge that needs to be answered even by the SSRist. This is an improvement over Boge (2020), where I neither acknowledged the connection between the framework developed therein and the PMI nor made the implications for SSR explicit.

In a first step, I will investigate the common structure of the NMA and the PMI, when recast in a broadly Bayesian framework.⁴ I will then turn to the formal framework introduced in Boge (2020) and, following up on the issues raised in this introduction, demonstrate that it gives rise to a pessimistic inductive inference that poses a challenge even for SSR.

A drawback to this line of reasoning is that it only provides a *local* argument against SSR. That is, fields of research in which the relevant situation occurs, i.e., in which there are several strictly incompatible classes of theories that persist together and encounter a stalemate regarding empirical success, are presumably sparse. However, localizing debates to individual fields of research is well in line with the recent trend in philosophy of science (e.g. Magnus and Callender 2004; Ruhmkorff 2013; Park 2019; Asay 2019; Vickers 2019; Dawid and Hartmann 2018), and I will here actually also introduce another field of research (psychometry), next to the one

⁴Note that there is no unified notion of ‘Bayesianism’: Good (1983, p. 20), for instance, constructs some 46,656 variants, based on a few fundamental differences, and indicates how a continuum could be constructed as well. By ‘Bayesianism’ I here mostly mean that one is willing to model credences (of idealized agents under idealized conditions) by appeal to probability functions, and that these can be updated in accord with Bayes’ rule when new evidence is obtained.

(nuclear physics) extensively discussed by Morrison (2011) and myself (Boge 2020), to which the argument can be applied.

However, more importantly, I will also argue that the mere *existence* of local bases for such a stronger PMI gives rise to a *general* challenge for SSR. In short: if it is possible to connect the success of a theory to a selected set of its elements that, for all we know, *cannot* be pieced together with the success-inducing elements of another, rival theory, then this undermines the very motivation for taking a selectivist stance in the first place.

2 The (common) structure of both arguments

2.1 The PMI: inductive or deductive?

The history of the PMI is complicated. As we have seen in the introduction, Poincaré (1905, p. 160), Laudan (1981), and Putnam (1978) each point to the historical record of abandoned scientific theories to make a point about SR, but in each case the point is different. Very often, Laudan (1981) is named as the *locus classicus* for the PMI. However, a careful analysis by Wray (2015, p. 65; *emph. added*) shows that:

Laudan's aim is to show that [...] a theory having genuinely referring theoretical terms is *neither necessary nor sufficient* for the theory being successful, and that a theory's being true (or approximately true) is *neither necessary nor sufficient* for a theory's being successful. [...] In this respect, the famous list of failed theories that Laudan does provide in 'Confutation' is unnecessary and constitutes overkill.

So while most have taken Laudan to offer a negative inductive argument about scientific theories' ability to refer and to get at the truth, his aims were more modest. Furthermore, for establishing (as Laudan in fact does) that success neither implies (approximate) truth nor reference, nor vice versa in each case, it would have sufficed to provide no more than four counter-examples. In effect, Laudan's classic paper has rather supplied a (disputable; e.g. Vickers 2013) *basis* for a pessimistic inductive argument against SR than an actual *argument*.

On the other hand, "Putnam", who himself "does not endorse the argument[,] presents the clearest example of the [PMI] advanced as an argument against scientific realism." (Wray 2015, p. 62) But Putnam's version, quoted also in the introduction to this paper, uses an unnecessarily strong (and thus highly defeasible) formulation:

Putnam's formulation [...] suggested that [...] '*all* the theoretical entities postulated by one generation ... invariably "don't exist"'. (Wray 2015, p. 62; *orig. emph.*)

The appeal to *reference* in Putnam's formulation came out problematic too: many terms that might be hastily discarded as *non-referential* could be rehabilitated as *indirectly* referential after all. However, given the drastic ontological changes accompanying, say, the move from phlogiston to oxygen theory or from Maxwell's equations for classical fields to quantum electrodynamics, it still seems possible to

dispute the status of these former theories as ‘even remotely true’. A pessimistic inductive inference to this effect, as discussed for instance by Chakravartty (2017, Sect. 3.3.), could hence be given in the following form:

- (Pr) Most past scientific theories have turned out to be radically false, despite their empirical success.
 (C) Therefore, by induction, most present and future scientific theories will (probably) turn out to be radically false as well, despite their empirical success.

This sort of argument is the proper target of SSR. Maxwellian electrodynamics, for instance, *was* right, the SSRst may hold, in important interpretation-invariant respects, such as the relation $\partial_\mu F^{\mu\nu} = j^\nu$, i.e., regardless of whether $F^{\mu\nu}$ and j^ν represent classical fields or quantum operators. Hence, many (if not most) discarded, successful theories did get something quite right, and it is that something which we should be realists about (whether it be ‘only structure’ or even some amount of content).

Surprisingly, however, the above formulation of the PMI has not been in the focus of the debate: “another variation [...] common in the philosophical literature” (Wray 2015, p. 64), and popularized also by Psillos (1999), displays the PMI as a *reductio*, i.e., as a deductive argument. Here is one such formulation, due to Lewis (2001, p. 373):

1. Assume that the success of a theory is a reliable test for its truth.
2. Most current scientific theories are successful.
3. So most current scientific theories are true.
4. Then most past scientific theories are false, since they differ from current theories in significant ways.
5. Many of these false past theories were successful.
6. So the success of a theory is not a reliable test for its truth.

I believe that this formulation (or the similar one endorsed by Psillos) has given rise to much confusion. For example, Mizrahi (2013) considers the possibility of explicating ‘significant difference’ in Lewis’s premise 4 as ‘incompatibility’. Following Devitt (2011, p. 288), however, he believes that “it does not follow that most past theories are false. For that conclusion to follow, pessimists have to assume that current theories are true.” (Mizrahi 2013, pp. 3215–6)⁵

But even assuming that ‘significant difference’ could be somehow interpreted as incompatibility—as it, presumably, *must*; for why should one abandon a theory on the ontological level if it was *fully* compatible with the next one?—this conclusion seems completely unwarranted: If we stick to the formulation of the PMI as an induction, why should we not assume the theories consecutively produced by science to approximate a sequence of mutually exclusive but by far not *exhaustive* alternatives (van Fraassen 1989; Stanford 2006)?

⁵See very similarly Psillos (1996, pp. S307–8): “if we hold current theories to be truthlike, then past theories are bound not to be truthlike since they posited entities that are no longer believed to exist, and posited laws and theoretical mechanisms that have now been abandoned. Without this premise the pessimistic conclusion does not follow.”

Ruhmkorff (2013) points out another subtlety associated with such a construal of ‘difference’ as ‘incompatibility’. According to Ruhmkorff (2013, p. 411), “[b]asic statistical inference requires that the sample be selected randomly and that the relevant properties of the individuals in the sample and population be logically independent”. Hence, due to the logical connections between past and present theories, the “PMI [...] fails to satisfy the criterion of independence.”

Ruhmkorff’s point obviously only makes sense if we think of the PMI as relying on *independent and identically distributed* trials, and it is no secret that “observational data rarely satisfy such conditions.” (Spanos 2019, p. 73) Hence, considerations of (in)compatibility between theories can (and should) enter our inductive arguments perfectly well; a point that will become important below.

2.2 The NMA and its connection to the PMI

Now consider the NMA again. As shown in the introduction, The NMA could be reconstructed as a deductive argument as well, for embracing that only if SR is true is scientific success no miracle, and that there indeed are no miracles, the truth of SR follows. But this is not the standard way the NMA has been conceived of:

NMA is intended to be an instance of inference to the best explanation (henceforth IBE, or abduction). What needs to be explained, the explanandum, is the overall empirical success of science. NMA intends to conclude that the main theses associated with scientific realism, especially the thesis that successful theories are approximately true, offer the best explanation of the explanandum. Hence, they must be accepted precisely on this ground. (Psillos 1999, p. 69)

IBE, or abduction more generally, is a class of inferences, leading (in a non-truth-preserving way) from an explanandum to an explanans that, if true, would render the explanandum expected (e.g. Douven 2017). An informal reconstruction of the NMA as an abductive argument could hence be given as follows:

- (Pr1) Scientific theories are remarkably successful in predicting, retrodicting, and explaining observable phenomena.
- (Pr2) If SR were true, i.e., if our best scientific theories were at least approximately true of a mind-independent reality which generates the observable phenomena, then this success would be explainable (but not otherwise).
- (C) Therefore, quite probably, SR is true.

Certainly, at lot is debatable here. For instance, an important part of van Fraassen’s empiricist program (e.g. 1980, 2008, 2010) can be considered a response to the bracketed clause in premise 2. However, I here want to focus on something quite different, namely the commonality between both arguments (the NMA and the PMI).

First of all, note that induction and abduction are both ‘ampliative’, i.e., add content. Furthermore, there is at least some consensus⁶ that “Bayesianism provides a

⁶See, for example, Poole (2000), Lipton (2004), and Romeijn (2013).

useful framework for studying abduction and induction as forms of ampliative reasoning” (Niiniluoto 2004, p. 64). It should hence come as no surprise that both arguments have been cast in a very similar formal mold that allows to make the connections very explicit.

One classic treatment of the NMA in a Bayesian mold certainly is Howson (2000). Following the ‘Harvard medical school test’, in which most students erroneously offer a high estimate for the probability of a patient having contracted a disease based on the fact that (a) the test is positive, it (b) has few false positives, and (c) also few false negatives, Howson argues that the NMA is fallacious: It commits to the same mistake as involved in the students’ erroneous judgments, i.e., in what is sometimes called ‘base rate neglect’ (e.g. Magnus and Callender, 2004) but should better be called *prior neglect*.

More precisely, assume that

- (i) $p(\neg s|t) \ll 1$,
- (ii) $p(s|\neg t) \ll 1$,

where s means that a given theory T succeeds in some empirical test, and t that T is approximately true. This means that, whatever procedures used to test T empirically, we can assume that if T is true and these procedures are suitable for testing T at all, then it is very unlikely that it will not succeed (‘few false negatives’) and that, if it was not true, it is very unlikely that it will succeed (‘few false positives’). However, using Bayes’ theorem, it is elementary to see that nothing follows from this about the probability $p(t|s)$ unless substantial assumptions are being made about the prior $p(t)$, sometimes confusingly called a ‘base rate’ (which is not generally well-defined; see Psillos 2009, p. 62 vs. Howson 2013). And isn’t $p(t|s)$ what we are really interested in?

Now apparently unaware of Howson’s argument (see Magnus and Callender 2004, p. 322), Lewis (2001) has argued that the PMI *also* suffers from prior neglect. In essence, Lewis turns the argument upside down: First off, using the law of total probability, we have

$$p(s) = p(s|t)p(t) + p(s|\neg t)p(\neg t) = \dots = p(t)(1 - p(\neg s|t) - p(s|\neg t)) + p(s|\neg t), \quad (1)$$

so that, under assumptions (i) and (ii),

$$p(t) = \frac{p(s) - p(s|\neg t)}{(1 - p(\neg s|t) - p(s|\neg t))} \approx p(s). \quad (2)$$

In words: if testing is ‘reliable’, in the sense of satisfying (i) and (ii), then, deductively, an increase in the probability of success implies an increase in the probability of truth. Hence, if we ground our credences in proportions and the proportion of successful theories increases over time, we should increase our confidence in the presence of (approximately) true theories accordingly, under the additional assumption of (i) and (ii). This formalizes, and somewhat justifies, the ‘convergentism’ disputed by Laudan.

Furthermore—and this is Lewis’s main point—*nothing* can be concluded about whether testing *is* ‘reliable’ (in Lewis’s sense) from the bare observation that many

past successful theories were false: If the proportion of approximately true theories among all existing ones is low initially but increases over time, this observation is to be expected in the sample of past theories often presented; even if the fraction of successful theories is constantly higher among the true than the false ones (see Saatsi 2005, p. 1095, for a helpful illustration).

Several things are problematic in Lewis's account though. For instance, it is not easy to recognize the sort of *inductive* argument that we were initially concerned with. This is because Lewis is arguing against the conclusion of his *reductio*-reconstruction: That success is not a reliable guide to truth. That conclusion is, in a sense, undermined by Lewis's argument, but it is not immediately clear how it bears on the proper PMI expounded on in Section 2.1.

To see that it need not immediately do so, take the standard Bayesian model of induction, which means conditioning on pieces of evidence e_i ($i \in \{1, \dots, n\}$) to update one's credence for a certain hypothesis h , where the e_i are consequences of h . Relevant pieces of evidence e_i are thus of the form $s_i \wedge \neg t_i$, where the i indicates reference to a particular, past scientific theory T_i . Following Section 2.1, the targeted h may now be assumed to be $\forall i : s_i \wedge \neg t_i$, meaning that theories which succeed empirically are, and will continue to be, quite far from the truth. Hence, the probability that is supposed to be sufficiently large on account of the historical record is

$$p(\forall i : s_i \wedge \neg t_i \mid \bigwedge_{j=1}^n (s_j \wedge \neg t_j)), \quad (3)$$

which has the implication that, for any present theory T_m (with $m > n$), $p(s_m \wedge \neg t_m)$ must be assumed even larger.⁷

Now if, for any given theory at any point in time, success is not a guide to truth, then either (i) or (ii) must be false. The more 'realistic' target is (ii): The accusation in the PMI is that radically false theories have regularly succeeded in empirical testing, not that true theories cannot even be found out empirically (which would constitute a whole other challenge). But since $p(\neg t_m) \leq 1$, $p(s_m \mid \neg t_m)$ is of course just as large or again even larger than Eq. 3, which establishes the desired conclusion.

However, recall how Bayesian-style arguments of this kind tend to be undermined by prior neglect: Lewis's point exactly is that $p(t)$ is undetermined, and that the proportion of severe falsity among past scientific theories could be vastly greater than among (and thus unrepresentative for) *all* scientific theories, including present ones (Magnus and Callender 2004, p. 326). Hence, relying on the historical record as evidence, "it would be fallacious to conclude that a test is unreliable on the basis of a large number of false positives alone." (Lewis 2001, p. 376)

As we already know, this argument can be turned on its head: Lewis's notion of 'reliability' is certainly non-standard in statistics circles,⁸ and Wray (2013, p. 1725)

⁷ Proof: Abbreviate the large conjunction by n , the quantified statement by $\forall$, and that about $T_{m>n}$ by m . Then by elementary probability theory $p(m) \geq p(\forall)$, since m is a deductive consequence of $\forall$. Since n is also a deductive consequence of $\forall$, it equally follows that $p(\forall \mid n) = p(\forall)/p(n)$. Now assume $p(\forall \mid n) \geq p(m)$. Adding both inequalities up and subtracting $p(m)$ on both sides, we obtain $p(\forall) \geq p(\forall)p(n)$, which is false unless $p(n) = 1$, in which case equality holds. That latter case is, however, false by assumption in any sensible application of the Bayesian scheme to inductive reasoning.

⁸ Kyriacou (2001), e.g., refers to (i) and (ii) as establishing a test's *validity*, which he sharply distinguishes from reliability.

points out that $p(t) = p(\neg s|t) = p(s|\neg t) = 10\%$ suffices to obtain $p(t|s) = .5$. And isn't $p(t|s)$ what we are really interested in?⁹

On the other hand, in the properly inductive formulation of the PMI, there are also priors that strongly influence its impact, namely $p(\bigwedge_{j=1}^n (s_j \wedge \neg t_j))$ and $p(\forall i : s_i \wedge \neg t_i)$; for given that $\forall i : s_i \wedge \neg t_i$ entails $\bigwedge_{j=1}^n (s_j \wedge \neg t_j)$, Eq. 3 boils down to the ratio between these.

Now most realists have argued exactly that $p(\bigwedge_{j=1}^n (s_j \wedge \neg t_j))$ is reasonably high: For instance, they have argued that, in contrast to present theories, past theories were 'premature' and, insofar as successful at all, likely successful for rather different reasons than present ones (e.g. Devitt 1984; Boyd 1984). Thus, it is no wonder that they were also far from the truth. On the other hand, $\forall i : s_i \wedge \neg t_i$ is a very strong claim and so $p(\forall i : s_i \wedge \neg t_i)$ may be assumed very low. Hence, there is no good reason for assuming that the PMI has established that Eq. 3 must be reasonably high. Or, stated differently: It is a theorem that, in the large n limit, Eq. 3 will converge to 1. But "[t]he result does not [...] tell us the precise point beyond which further predictions of the hypothesis are sufficiently probable not to be worth examining" (Howson and Urbach 2006, p. 94).

Anti-realists, on the other hand, could doubt the notion of maturity appealed to by the realist. They could maintain that the prior for the general hypothesis should at least be some 10%, for we cannot know *a priori* whether scientific theorizing *can* guide us to the truth, and whether empirical testing does not reveal a rather different quality (such as mere empirical adequacy). Furthermore, the discovery of the many past successes despite what is taken as severe falsity could be counted as rather unexpected, and we quickly end up with a posterior close to 1.

The discussion could be continued *ad nauseam*—the very point of Magnus and Callender (2004). This level of formal precision has hence not yet brought us very far past the clash of intuitions indicated in the introduction.

3 Towards a stronger induction

3.1 The frequency-based NMA

In an attempt to improve on the above situation, Dawid and Hartmann (2018) have offered a new rendering of the NMA, closely related to some informal arguments given by Henderson (2017), in which it comes out as a valid argument, not a fallacy. In short, assuming conditions (i) and (ii), Dawid and Hartmann (2018) show that, if one conditions $p(s)$ on the datum, o , that the rate of successful theories in some field $\mathcal{R}$ is r , and (absent any other relevant information) sets $p(s|o) = r$ as an estimate, it follows that

$$p(t|s \wedge o) > 1/2, \quad (4)$$

so long as $p(s|o) > 2p(s|\neg t \wedge o)$, which provides a sensible formulation of a positive, but still rather weak truth-success connection.

⁹Note that a similar conclusion is established also by the properly inductive Bayesian argument above, by repeating the ultimate step with $p(s_m) \leq 1$ and $p(\neg t_m|s_m)$.

In principle, I see nothing wrong with using this sort of frequency estimate to ground (or replace) a prior. This is a standard procedure in many statistical analyses, even though care must be taken to get the frequencies right (e.g. Schurz 2014, p. 132 ff.; or Sesardic 2007, for an impressive example). Hence, $p(s|o) = r$ does not seem objectionable in itself. However, in a response-paper (Boge 2020), I showed that there are reasons to be skeptical that this is compatible with (i) and (ii). I will briefly outline the argument below and then show how it can be extended into a pessimistic argument about the prospects of SSR.

3.2 The incompatibilist response

The heart of my earlier argument is to consider yet another datum, c , namely that there are k classes of *successful* theories in $\mathcal{R}$ within which theories are compatible, or incompatible only in a weak sense, and among which they are *strictly* incompatible. These notions were defined in a rough and ready way (Boge 2020, p. 4347–8; orig. emph.) as follows:

Call [it] an instance of *weak compatibility* [if] each theory within a class C either *is the* limit of another theory in C , or has another theory in C as a limit, or that two theories T' and T'' from C do not make any contradicting claims [...]. Call it a *strict incompatibility*, in contrast, when two theories rest on mutually contradicting assumptions and/or have mutually contradicting implications, and none of the two is the limit of the other in any sense.

‘Limits’ between *theories* are not something that is exactly defined, and the definitions may not be general enough (for instance, some theories might not be formulated mathematically, and so limiting theorems might not exist). I will offer improved definitions of the relevant notions below, when making contact with SSR and the PMI. But for now note that, under these assumptions, a new, or newly considered, theory

T will be approximately true only if it falls into the one out of the k classes which happens to contain all the approximately true theories, assuming there are any. It should be obvious that only compatible theories can be simultaneously true, and since only approximate truth is at stake, weakly compatible ones may well all be more or less approximately true at the same time. If T happens to fall into none of these classes however, it will define a new class, and hence only one out of $k + 1$ classes will contain approximately true theories. (Boge 2020, 4348; emph. omitted)

Together with the possibility of facing a ‘bad lot’ (van Fraassen 1989), this observation was used to stipulate that

$$p(t|c) \leq \frac{1}{k}, \quad (*)$$

i.e., that the probability of approximate truth cannot exceed the equivocating probability an objective Bayesians would initially assign to mutually exclusive cases. It is then shown in the paper that Eq. *, when conditioned also on o , is inconsistent with

the conjunction of Dawid and Hartmann's frequency-based estimate for the probability of truth and conditions (i) and (ii) (few false positives and negatives in the testing of mature scientific theories).

I will here not go further into the details of my earlier argument, but rather note a number of general observations. First of all, I indicated above that the PMI really targets the number of false positives, not false negatives. Hence, I will focus entirely on the impact that my earlier proposal has on false positives below. Furthermore, consider that, if we operate under the most realism-friendly assumption (equality in Eq. *),¹⁰ as shall be done for the remainder of the discussion, we get from the law of total probability:

$$p(s|c) = p(s|t \wedge c)p(t|c) + p(s|\neg t \wedge c)p(\neg t|c). \quad (5)$$

Conditioning also the probabilities for false positives and negatives on c , we may set

$$\begin{aligned} \text{(i}^*) \quad & p(s|t \wedge c) = 1 - p(\neg s|t \wedge c) = g, \\ \text{(ii}^*) \quad & p(s|\neg t \wedge c) = f, \end{aligned}$$

where an empirical test with high g is often called *sensitive*, one with low f *specific*. Then

$$p(s|c) = \frac{g + f(k-1)}{k}. \quad (6)$$

In case $k = 1$ (no strictly incompatible theories), this collapses into $p(s|c) = g$, so the compatibility assumptions become trivial and have no bearing on the relevant probabilities. Hence, assume that $k \geq 2$, from now on. Let also e an estimate of $p(s|c)$; an alternative to Dawid and Hartmann's $p(s|o) = r$. Then, rearranging Eq. 6, we may write

$$f = f_k(e, g) = \frac{ke - g}{k-1}, \quad (7)$$

giving us an expectation for false positives as a function of expected success and assumed sensitivity, parameterized by the number of compatibility-classes.

The members of $\{f_k\}_{k \in \mathbb{N} \geq 2}$ —to which I shall generically refer as ' f_k '—are smooth over $[0, 1]^2$, meaning that an agent's beliefs about false positives shouldn't rapidly jump at any value of (e, g) , and has a constant slope in both directions, i.e.: $\partial f_k(e, g)/\partial e = 1/(1 - (1/k))$ and $\partial f_k(e, g)/\partial g = 1/(1 - k)$.

As we can see, f_k scales negatively with g already for $k = 2$. So an *increase* in g leads to a *decrease* in f_k . But this effect is suppressed with growing k , and f_k increases almost *exactly* as e for fields with high k .¹¹ So as soon as there are sufficiently many compatibility classes, high expectations regarding success essentially enforce a low credence for specificity. The intuitive content is that if we accept that only few tested theories can be even approximately true, but are confident in their

¹⁰This assumption is 'most realism friendly' because it not only assumes $\mathcal{R}$ to be no bad lot but also that not *all* our previous theories in $\mathcal{R}$ have been far from the truth (the $k + 1$ -case).

¹¹ $1/(k) \approx 0$ for large k , and so $1/(1 - (1/k)) \approx 1$, whereas $1/(1 - k) \approx -(1/k) \approx 0$ in that same case. Moreover, already for $k = 2$, $\partial f_2/\partial e = 2$, whereas $\partial f_2/\partial g = -1$; so the increase of false positives to be expected with higher expectations of success is *twice* that of the decrease with increased expectations of sensitivity.

overall empirical success, we simply cannot take this success to be a good indicator for truth.

3.3 ...Absent any further information...

As pointed out above, Dawid and Hartmann (2018) stipulate their success estimate in the absence of further relevant information, as did I in Boge (2020) with regard to my truth-bound. Obviously, however, certain pieces of information might also skew our expectations regarding truth against Eq. *, even upon accepting c . For instance, extra-empirical virtues could tip us off to consider one rival as more promising; or knowledge of theoretical developments could spark the expectation of *unification* on the horizon. I here want to assess the quantitative impact of these observations on my argument by asking: How much skewing would defy the consequences outlined above?

To formalize the question, consider some relevant datum o' about $\mathcal{R}$, and let accordingly $p(t|c \wedge o') = \frac{1}{k} + \gamma$, for some $\gamma \in \mathbb{R}^+$ s.t. $\frac{1}{k} + \gamma < 1$ (i.e., $\gamma < \frac{k-1}{k}$). Then¹²

$$f_{k,\gamma}(e, g) = \frac{ke - g(1 + k\gamma)}{k(1 - \gamma) - 1}. \quad (8)$$

A sensible condition for realism is $f_{k,\gamma} < 1/2$. Then we get

$$\gamma(1 - 2g) < \frac{k-1}{k} + \frac{2g}{k} - 2e. \quad (9)$$

Moreover, assume as before that $g \approx 1$ and that we are initially rather neutral regarding T 's success ($e \approx 1/2$). This allows us to set

$$\gamma \gtrsim \frac{1-k}{k} - \frac{k-2}{k} = \frac{1}{k}, \quad (10)$$

which means that $p(t|c \wedge o') \approx 1$ for $k = 2$, and more generally $p(t|c \wedge o') \gtrsim \frac{2}{k}$.

Given that all compatibility classes were assumed successful, the former seems like a terribly strong increase. It could only be justified exactly in cases where T would deliver unification. In general, only higher values of k would allow a sensibly low increase as compared to $1/k$. But it is exactly in this case that we should also be skeptical about T 's truth. Hence, only special circumstances (unification) will allow to constrain $f_{k,\gamma}$ significantly, even under modest expectations of success.

3.4 The inductive character of the argument

What, if anything, is the connection of the above incompatibility-based argument and the PMI? According to Boge (2020, p. 4349), the connection is frail, but I believe that this verdict was premature.

The first important observation, also made in Boge (2020), is that Dawid and Hartmann's proposal is obviously an *inductive* fix of the NMA: The detailed Bayesian

¹²For simplicity, we use all the same letters as before, with obvious meanings.

analysis left open by them would mean conditioning on past success to infer something about success in general, and hence quite likely involve the use of the Bayesian formalism in the properly inductive sense appealed to above. However, even setting $p(s|o) = r$, where o means that a fraction r of theories in $\mathcal{R}$ was successful is a kind of inductive inference, sometime called “inductive-statistical specialization” (see Schurz 2014, p. 51; similarly Carnap 1950, p. 207).¹³

(Pr1) $r\%$ of all F s are G s.

(Pr2) This individual is an F .

(C) Therefore, we can be $r\%$ certain that this individual is a G .

If $F(T)$ means “ T is a theory in $\mathcal{R}$ ” and $G(T)$ means “ T is successful”, this immediately reproduces Dawid and Hartmann’s $p(s|o) = r$.

A second observation is that, while the traditional PMI is indeed concerned with the historical record (as we saw above), induction is, of course, not restricted to cases past, but rather to cases successively *observed* or *considered* (just think of the notorious “all swans are white”-example). Hence, that my earlier proposal is not a version of *Laudan’s* PMI because I consider “incompatible theories [...] that *persist together*” (Boge 2020, p. 4349; orig. emph.) has no bearing on whether this is an inductive argument or not.

Now reconsider condition (*) and let $F(C; T)$ mean “the class C into which T falls within $\mathcal{R}$ is one of k compatibility classes”, and $G(C; T)$ “the class C into which T falls contains approximately true members” (with C being that which something is predicated of, and T constituting an identifying parameter in each case). Given that the frequency of truth among these classes cannot be any higher than $1/k$, on account of c , the inductive-statistical specialization involved here is the following:

(Pr1) No more than $1/k$ of the k compatibility classes C_i into which theories T fall within $\mathcal{R}$ contain approximately true theories.

(Pr2) The compatibility class C_i into which this theory, T' , falls is one of these k classes or increases the number to $k + 1$.

(C) Therefore, we can be no more than $1/k$ certain that T' is even approximately true.

The last part strictly speaking involves a deductive step. As we can see, the induction base is constituted by the different *classes* into which theories fall, rather than theories directly. However, since T' is approximately true only if it falls into (or defines) the ‘right’ C_i , it can obviously be approximately true with a probability no greater than that of its class containing the approximate truth.

¹³We here also see explicitly in what sense Lewis’s reconstruction of the PMI has an inductive component at all: It uses a sample frequency as an estimate for the credences to be invested in the next example, and then (fallaciously, according to Lewis) infers something further about the (likely) truth-success connection. Hence, it depends on a similar inductive move in its premises.

The statistical part here is the discovery of k relevant classes, together with the recognition that the class relevant to the next theory will either fall under this disjunctive collection or even extend it.¹⁴ The inductive move then is the use of the corresponding statistical quantity for making a claim about the next example, i.e., that it is as unlikely to contain the approximate truth. Hence, Eq. * is clearly inductively grounded.

Note that this inference constitutes an *elementary* step in the reasoning chain, and a deeper Bayesian analysis would add virtually nothing. Furthermore, note the connection to Lewis's 'deductive' reconstruction of the PMI: if $p(t)$ must be assumed to *remain* rather small well into the present, while $p(s)$ also remains large or even increases, then this *does* defy the assumption that testing is reliable in the sense of (i) and (ii) (see eq. Eq. 2). Hence, the above induction properly facilitates the key move attested to Laudan by Lewis, without committing to a fallacy.

3.5 Creating a pessimistic challenge for SSR

Now initially I claimed that SSR was one important response to the traditional PMI. Hence, what (if anything) is the bearing of the present inductive argument on SSR? To understand this properly, a number of clarifications are in order. First off, recall that I had (vaguely) spelled out the connection between theories within a compatibility class C as provided by limiting theorems, securing that the successful empirical predictions of one theory, T , are preserved in another one, T' . However, as was pointed out above, limiting theorems may not exist, even though success is preserved.

Recall, for instance, Schurz's assessment of phlogiston and oxygen theory. It is unclear in what sense the former should be considered the 'limit' of the latter, even though the structural correspondence theorem demonstrates how empirical success can be preserved in a conservative way in the transition from one to the other. The common feature of both assessments is that the later theory manages to recover the successes of the former by 'imitating' part of its behavior while neglecting some of its characteristic features. For instance, by introducing corresponding atomic reactions while simultaneously eliminating the phlogiston from phlogistication, oxygen theory removes the appeal to a specific substance responsible for combustion while preserving phlogiston theory's successes. Similarly, by extending m to $\gamma(v)m_0$, relativity abandons the generic frame-independence of mass while still recovering the Newtonian predictions for all speeds that are small relative to that of light.

Since oxygen theory 'works better' empirically than phlogiston theory, just as, say, special relativity works better than Newtonian mechanics, a family of such theories constitutes what I here want to call a *success-continuum of theories* (SCT):

Definition 1 (SCT) Let $\mathcal{T}_t$ a time-indexed family of theories in the same field of research $\mathcal{R}$. Then at any given t , $\mathcal{T}_t$ is a *success-continuum* if, and only if, there is a

¹⁴The compatibility class into which a theory falls is at first not necessarily immediately obvious. For instance, in nuclear physics, there are hybrid models that seem to cut across the classes, but only upon close inspection can be seen to properly fall into one particular class (cf. Boge 2020, p. 4357).

partial ordering $\preceq_s$ such that $T \preceq_s T'$ or $T' \preceq_s T$ for all $T, T' \in \mathcal{T}_t$, where $T \preceq_s T'$ if, and only if, T' recaptures all empirical successes of T .

Note that the time-index refers to the family as a whole, not the individual theories. The reason for time-indexing at all is that new members may be augmented at some time $t' > t$ and that would extend (but not shuffle) the partial ordering.¹⁵

SCTs look like a generalization of compatibility classes. However, as I was aware already in Boge (2020), it cannot be excluded that the given field $\mathcal{R}$ features theories that have strictly disjoint areas of application. Hence, a compatibility class C more generally corresponds to a *disjoint union* $\sqcup_i \mathcal{T}_t^{(i)}$ of $\mathcal{T}_t^{(i)}$ from $\mathcal{R}$ that do not make contradicting claims.

The crucial point now is that the SSRist can be a realist about whatever survives in the succession $T, T', T'' \dots$, as it is that which she can hold responsible for the prevailing success. For instance, she can be a realist about the relation represented by ‘dephlogistication’ in the phlogiston theory, even if it does not include a substance called ‘phlogiston’ after all: At least there is a sense in which the claims to dephlogistication were *approximately* true, and the modern account in terms of electron donation and absorption may just be somewhat closer to the truth (given the increased empirical success accompanying it).

However, recall, on the other hand, what the change from T to T' to T'' entails:

[M]ost successful theories [...] share two important features: (i) the ability to provide a quantitatively accurate description of a reasonably wide range of phenomena; (ii) a *conceptual cohesion* that makes it possible to express its fundamental tenets in a reasonably simple mathematical form. [...] Even though in many situations [a] new theory might be practically equivalent to the old, there could be a *conceptual gulf* between them, as illustrated by [...] the Newtonian theory of gravity and [...] Einstein’s general theory of relativity. (Duncan and Janssen 2019, p. 35; *emph. added*)

Strike out ‘quantitatively’ and ‘mathematical’, and this is a fairly general account of theory change. In particular, it captures both, continuity and a fundamental breach. The remarkable point is that the fundamental breach is located in the ‘conceptual cohesion’ of the old theory, very much in line with Schurz’s acknowledgement of the loss of the ‘internal structure’ of dephlogistication with the abandonment of phlogiston.

Now Lakatos (1976) taught us that *a lot* about a theory can be changed in the face of new evidence, without inducing the need to abandon the theory completely. For Lakatos (1976, p. 243), “if and when the programme ceases to anticipate novel facts, its hard core might have to be abandoned”. However, as we saw, it is usually necessary to also abandon the *core concepts* of the old theory, in order to extend the scope of theoretically predicted, (use-) novel facts. Hence, when we say that ‘phlogiston theory is obsolete’, we mean that some of its *fundamental* assumptions

¹⁵Of course, there may also be transitory phases where $\preceq_s$ is reversed; for instance, this may have been the case w.r.t. quantum physics and classical mechanics when, e.g., decoherence effects were not fully understood yet. However, if the situation turns out to be non-transitory, this usually means that the new member will soon be abandoned entirely.

had to be given up (the existence of phlogiston); something that impairs the theory's "conceptual cohesion".

Accordingly, I take it that 'fundamentality' has to do with the concepts a theory employs, and is sensibly defined as follows in this context:

Definition 2 (Fundamentality) A posit, p , underlying theory T is *fundamental* for T if, and only if, removing p from T makes T *conceptually incoherent*.

I don't think it is possible to define 'conceptual cohesion' explicitly, or to tell by means of a single definition when it 'collapses'. But I do think this can both be made fairly clear by means of examples. For instance, postulating a phlogistication-process *without* radically changing its meaning (which would amount to embracing the successor theory) while also *abandoning* the central entity which drives it (phlogiston) is conceptually incoherent: It means believing in a process in which something is absorbed which doesn't exist or cannot even in any way be defined. Hence, assuming that there is no special substance ('phlogiston') leaves no other option than to abandon phlogiston theory, even if certain elements of it can be reinterpreted in the context of the next theory. Similarly, employing (without radical meaning-change) the equations of motion that derive from a non-relativistic Hamiltonian while embracing that material bodies cannot be faithfully described by reference to a time-dependent position variable (their center of mass coordinate) in *three-dimensional* space is conceptually incoherent. (And so forth.)

In this sense, 'phlogiston' and Galilei (as opposed to Poincaré or diffeomorphism) invariance are fundamental concepts of phlogiston theory and classical mechanics, respectively, and simply removing either without a suitable replacement means embracing incoherence—while replacing either by a different concept changes the meanings of all other terms (and usually the shape of most theoretical posits), and thus amounts to embracing a different theory.

Now I take the main point of SSR to be that, while maybe *conceptually* fundamental in this sense, these concepts or the posits containing them were not *essential* for empirical *success*, i.e., (roughly) did not do any real work in the derivation of empirically adequate predictions: Just as it is not the phlogiston that does the work in predicting observed properties of combustion, it is not the (exact) Euclideaness of space that does the work in predicting the trajectories obeyed by the center of mass-coordinates of macroscopic objects at low velocities.

However, what are the *general* conditions under which a posit (or a corresponding concept) does do 'real' work in the derivation of a successful prediction? That I believe to be an open question. What can be clearly said so far is this:

sometimes it is clear that a used assumption [...] did work to generate the success *solely in virtue of the fact that it entails some other proposition which itself is sufficient for the derivational step in question*. (Vickers 2017, p. 3227; orig. emph.)

Vickers illustrates this claim with Bohr's derivation of the spectrum of ionised helium from an assumption of 'allowed' orbital trajectories. But of course the same spectral lines follow from the (logically weaker) assumption of discrete

energy-levels. However, according to Vickers (2017, 3227), “the realist does not need to claim that [even the weaker hypothesis] merits realist commitment”, for “[p]erhaps the derivation can go through with an assumption still weaker” (ibid.).

Indeed it can: spectral lines have non-negligible width, and this can only be understood by considering interactions, which leads to a Hamiltonian with mixed rather than a discrete spectrum (Basdevant and Dalibard 2006, p. 349). Hence, the relevant posit is not that the spectrum *is* discrete, but that it *contains* the discrete energies and has a distribution that peaks around them.

Now, perhaps the derivation of observed lines can go through with *even* weaker assumptions, and so the realist does not need to claim the existence of the mixed spectrum. But maybe she does not even need to claim the existence of Hamiltonians (or whatever they represent in quantum theory); of wave functions (or whatever the hell *they* represent) ... of a differentiable spacetime manifold...?¹⁶ Then what *does* the realist need to claim?

I submit that this recurrence to ever weaker posits (relational or not) threatens to collapse the realist position into an empiricist one, or at least verges on an *immunization* strategy (see similarly Peters 2014). For instance, following Asay (2013, pp. 11–2; 2012, p. 389–90), we can take the kind of *truthmaker* realists and anti-realists are willing to admit, respectively, to make for the crucial difference between them: If some posit was made true solely in virtue of relations (however complex) pertaining between observable entities only, it would certainly not be made true by the ‘right kind’ of ontology. But for sufficiently weak theoretical concepts, it is at best unclear whether they are capable of representing anything more than involved short-cuts for referring to complex (relations between) observables.¹⁷

Defining a sensible, non-trivial, and non-immunized endpoint to the successive weakening of posits by the SSRist would amount to defining what sorts of posits are ‘absolutely essential’ in a way that does justice to the above intuitions. This is obviously highly non-trivial, and may not be possible at all. Luckily, it is also not necessary for my purposes. Rather, I’d like to define *relative (predictive) essentiality* with reference to an SCT as follows:¹⁸

¹⁶Just consider how Lagrangians dominate quantum field theory, but are not generally connected to Hamiltonians via Legendre transformation therein (e.g. Schwartz 2014, p. 261); or how ‘wave functions’ might be interpreted as mere ‘generalized probabilities’, in turn connected to subjective degrees of belief (e.g. Leifer 2007); or how quantum gravity threatens the existence of spacetime in any meaningful sense (e.g. De Haro and de Regt 2018).

¹⁷To make this just a little more plausible, consider the “air of schizophrenia” van Fraassen (2006, p. 295; *emph. added*) finds in Worrall’s SSR, because “all the support given for the [...] claim [...] that [...] scientific knowledge is cumulative in some important respect [...] is explicitly and admittedly concerned only with an accumulation of *empirical* knowledge.” Or consider Worrall’s (2020, 200; *orig. emph.*) own defense against the infamous Newman-objection, that “the real Ramsey sentence *turns only theoretical and not observational predicates* into variables and then existentially quantifies over them.” If ramseyfied in that way, fully non-observable structures are defined only up to isomorphism, and classes of isomorphic structures are, by themselves, pretty much incapable of conveying any real knowledge at all (exactly Newman’s point).

¹⁸Further conditions could be added to this definition that would safeguard against obviously auxiliary assumptions entering. But then again, certain auxiliary assumptions might be considered essential after all (Peters 2014), and so I shall be satisfied with this definition.

Definition 3 (Relative essentiality) Let T_i an SCT. Then a posit, p , is *essential relative to T_i* if, and only if, p (i) figures non-trivially in the *derivation of empirically successful predictions* and (ii) is the *weakest* such posit over all of $T \in T_i$.

This definition allows that, with the augmentation of new members at $t' > t$, weaker posits may be introduced that have the same empirical consequences and render older ones non-essential (as in the Bohr / Schrödinger / interacting quantum mechanics-transition indicated above). It hence does justice to the core intuitions of SSR. However, with this notion of relative essentiality, it is also possible to make precise the notion of strict incompatibility:

Definition 4 (Strict incompatibility) Two theories, T_1 and T_2 , from the same field of research $\mathcal{R}$ are *strictly incompatible* if, and only if, they contradict each other w.r.t. *essential* posits. Any two theories that are not strictly incompatible are *at least weakly compatible*.

The important point here is that, within the SCTs defining the compatibility classes, most theories are certainly also *incompatible* in a sense: Quite usually, they will contradict each other in (conceptually) *fundamental* respects, as we have seen. But so long as they don't contradict each other in relatively *essential* respects, that poses no problem for the SSRist. Furthermore, so long as there is only one such class, this also saves the heart of the 'convergence to truth'- and no-miracles-intuitions along the way: Scientific progress reveals in *what respects* previous theories were more or less approximately true, and *that* explains their success.

Note that no reference to mathematical limits was made in these definitions, and so the present account generalizes my earlier one. However, for mathematical sciences it may still be *useful* to consider the (non-)existence of such limits, for seeing whether there is weak compatibility. I take it, in other words, that the rough and ready 'definition' I offered in Boge (2020) really provides a *criterion*, not a definition:

(LC) Two mathematically formulated theories, T_1 and T_2 , may be judged to be strictly incompatible in case none of the following conditions holds: (i) T_1 is the limit of T_2 ; (ii) T_2 is the limit of T_1 ; (iii) T_1 and T_2 do not make any contradicting claims.

As pointed out above, that criterion presupposes the notion of a 'limit between theories', which shall be defined as follows:

Definition 5 (Limits between theories) A theory T_1 is said to be the *limit* of another, T_2 , if for each successful empirical prediction of T_1 there is a mathematical limiting procedure such that, when the limit is taken, T_2 recovers the prediction.

Note the weak ordering of the quantifiers, which makes the definition quite liberal.

To feed flesh to these bones, we shall briefly reconsider the main case study (nuclear physics) that has been used for this context, and on top of that introduce a second one. For now, however, consider the overall structure of the argument so far.

The selectivist response to the traditional PMI goes roughly as follows: ‘Maybe, dear pessimist, you have shown that theories are often superseded by others that contradict them in many ways—even in ways that lead to the withdrawal of fundamental concepts. But that is not so bad after all, for there is continuity regarding the posits essential for their success. Hence, we can hold fast to these continuous (or even invariant) parts as being explanatory of the increasing overall success in science, and consider *them* as approximating the truth to an ever and ever higher degree.’

The response of the stronger induction now is this: ‘That assessment may be fine in several cases, dear realist, but it cannot be upheld in general. For there is a bunch of theories that contradict each other in *essential* ways, and all these still harvest success. Overall, your acclaimed truth-success connection is flawed, and many of the success-generating posits cannot even be approximately true. For they flatly contradict *other* success-generating posits, and in ways that appear to defy any sort of (non-instrumental) weakening capable of preserving success.’

In this connection, I have been confronted with an interesting objection by an anonymous referee: Assume that there were two sets of theories as to the number, N , of planets in the solar system, and that one of them could claim success on account of the posit ‘ $N = 7$ ’, one of them on account of the posit ‘ $N = 9$ ’. Then both these posits could be approximately true if, indeed, ‘ $N = 8$ ’. And, furthermore, a theory which would recover all the successes of the ‘ $N = 7$ ’- and ‘ $N = 9$ ’-theories on account of a ‘ $N = 8$ ’-posit, consistent with the weakening $7 \leq N \leq 9$, could clearly count as a unification in this field of astronomy.

However, if interpolating in this way between both theories would apparently always spoil the success entirely, we would here face the sort of situation I have in mind: Assuming a theory according to which either ‘ $N = 7$ ’ or ‘ $N = 9$ ’, we would in at least one of these cases be fairly wrong, given the large logical gap apparently created by embracing one of these posits, respectively, relative to all existing background knowledge. Furthermore, simply botching together a ‘ $N = 7 \vee N = 9$ ’-theory in order to preserve all the empirical success would be a clear example of ‘weakening gone wrong’: Which is it? Seven or nine?

Nevertheless, the question arises if the probability for the approximate truth of any given theory should here still be less than $1/k$. I would say ‘yes’. Think about it: According to SSR, the *reason* to invest doxastic commitment in a certain posit is exactly that it creates the notable successes. Now if accepting ‘ $N = 7$ ’ spoils the successes harvested with accepting the ‘ $N = 9$ ’ theory and there is neither any way to recapture these from the ‘ $N = 7$ ’ theory nor any way to combine these into, say, a ‘ $N = 8$ ’ theory while saving the successes, then the SSRist faces a choice.

Here, it is inessential that the posits themselves both lie in numerical proximity to one another. As the situation is set up, they can’t both lie in proximity to a posit that, according to SSR’s criterion for doxastic commitment, is the best candidate for the literal truth: The only candidate of that sort would¹⁹ be ‘ $N = 8$ ’. And by assumption, ‘ $N = 8$ ’ does not create empirical successes. So there is no good reason to assume

¹⁹...without loss of generality; we can easily extend this to, say, $N = 10$ and $N = 6$ also not generating successes.

that both theories, the ' $N = 7$ ' and ' $N = 9$ ' one, are (equally) approximately true, nor even a good reason to assume that both *posits* are equally approximately true. Rather, the conclusion must be that moving even one integer away in the quantitative posit means moving away big time from the truth (if any).

So the way the situation is set up, the success-wise incompatibility between ' $N = 7$ ' and ' $N = 9$ ' enforces a choice regarding doxastic commitment, as permitted by the γ introduced in Section 3.3. Furthermore, upon committing to what seems to be the most promising candidate (say ' $N = 7$ '), the SSRist may hold out for unification under a compatible successor. This would justify updating $p(n = 7) = (1/k) + \gamma \mapsto q(n = 7) = 1$ (with p and q conditioned on the relevant observations). But so long as the stalemate persists, the low inductive prior remains justified.

To make all this just a little more plausible, consider the following two extended toy models: Assume that, in reality, there be several 'dark' planets with exotic gravitational properties hidden in the solar system, and that observations are not unambiguous as to the actual number of 'bright' planets. Furthermore, one of the dark ones could be such that it can be detected under special circumstances, fueling the ' $N = 9$ ' theory. Then the actual number might be 20, but the gravitational effects might cancel out in such ways that, neglecting *either* all (and only the) 11 'fully dark' planets *or* the fully dark ones, one bright one, *and* the 'semi-dark' one (and only these) may get several observable properties of the entire system (use-novelly) right (different ones in each case). But we need not assume that the neglect of additional bright planets or the inclusion of any small number of dark planets has the same effect.

The homelier case would be one in which there are indeed, say, seven planets, but electromagnetic and gravitational effects within the solar system create the occasional illusion of one or two additional planets. Then if the laws or relations that could ultimately predict all observable effects, including the illusion of additional planets, would happen to require that, *ceteris paribus*, the number of planets in this particular system cannot be moved even an epsilon away from seven, this would make the posit ' $N = 9$ ' quite radically false.

Because we cannot strictly exclude scenarios of this general flavor, the observation of a possible numerical proximity between essential posits has—*absent any further information*—no bearing on the $1/k$ upper bound. I believe that this makes the relevant intuitions fairly clear. But to also make it tenable that a relevantly similar situation really persists, we need to look into actual scientific cases.

4 Flesh to the bones

4.1 Nuclear physics: Recap of the main case study

The most impressive case in favor of the strong induction is that of rival theories of the atomic nucleus (Morrison 2011; Boge 2020). The two most salient rivals here are shell- and liquid drop-models, which both claim a number of predictive successes and nevertheless disagree essentially, namely on the so called 'mean free path' (MFP), the average distance a particle can travel inside the nucleus without interactions. Shell

models cannot get off the ground (conceptually *or* prediction-wise) if one assumes interactions to be frequent (and the MFP to be small accordingly). Liquid drop models, in turn, need to assume tightly bound states in order to predict e.g. collective oscillations, and so a small MFP.

Strictly speaking, the two kinds of models are not just incompatible, but in a sense *incomparable*: one starts from the level of nuclei, the other of nucleons; and so the basic domains over which they are defined operate on different levels. However, it is perfectly *conceivable* that one could in principle coarse grain the shell model in such a way that it reproduces known, successful drop models under relevant conditions.

Given these considerations, the case for strict incompatibility is straightforward: $\text{MFP}_{\text{shell}} \rightarrow \text{MFP}_{\text{drop}}$ (or, insofar as sensible, vice versa) will in fact *not* reproduce any known drop model, and spoil the successful predictions.²⁰ The case has been studied in considerable detail by Morrison (2011) and Boge (2020), so it is unnecessary to recover everything in full glory here. However, it may be helpful to reconsider at least some details.

First off, it is not correct, as Morrison (2011, p. 347) claims, that it is “necessary” to assume “that the nucleus is a classical object [...] for generating accurate predictions” from the drop model. There is a quantized Hamiltonian that follows from the basic assumptions of the drop model (most importantly, the neglect of free nucleon states) and successfully predicts collective excitations, as well as accurate values for what is called the ‘surface diffuseness’ of nuclei (Boge 2020, p. 4352 and references therein). Similarly, the so called ‘mass formula’, which is at the heart of many empirical applications (such as the charge-squared/mass ratio for which nuclei become unstable) was derived by Weizsäcker (1935) using non-relativistic quantum mechanics.

On the other hand, with the right kind of quantum Hamiltonian for free nucleons, subject only to an ultra-short remnant of the strong force, the shell model can predict, e.g., certain regularities concerning nuclear spin or, most importantly, the ‘magic numbers’, i.e., “numbers of protons or neutrons respectively for which a nucleus exhibits extraordinary stability.” (Boge 2020, p. 4354) The difference is hence not quantum vs. classical, but rather ‘by and large free’ vs. ‘tightly bound’ nucleons.

But this is clearly an *essential* difference:

The *existence* of orbits in the nucleus with well-defined quantum numbers is possible *only* if the nucleon is able to complete several ‘revolutions’ in this orbit before being perturbed by its neighbors. Only then is the width of the level smaller than the distance to other levels[...]. (Blatt and Weisskopf 1979, pp. 777–8; *emph. added*)

The concept of Liquid-drop model originated in Bohr’s assumption of compound nucleus in nuclear reactions. When an incident particle is captured by a nucleus its energy is quickly shared by all the nucleons. The mean free path of

²⁰Moreover, assuming something about the average distance over which interactions can be neglected is a considerably weak assumption, so it is hard to see how to weaken it further without spoiling the possibility of (non-instrumental) theory-formation.

the captured particle is much smaller than the nuclear size. In order to explain such a behaviour, the interactions between nucleons have to be strong. Consequently, the particles *cannot* move independently with negligible collision cross sections with their neighbours. (Kamal 2014, p. 380–1; *emph. added*)

In other words: the two essential assumptions, expressible in terms of a requirement for the MFP, are that nucleons interact either negligibly or strongly. If either of these assumptions is dropped, it becomes immediately impossible to formulate the respective Hamiltonian, and so to get any of the aforementioned predictions going.

These are not all the rival SCTs in nuclear physics; in Boge (2020, Section 3.2), I counted some 15 of them. However, 15 seems like an awfully large number, so let us assume, for the sake of argument, that one can reduce these classes of models to “the four foundational models on which most developments in nuclear physics are based” after all, i.e., “the collective [=liquid drop] model, the shell model, the pair-coupling models, and the mean-field [...] models.” (Rowe and Wood 2010, p. ix)

Allusions to *unification* can be found especially in the shell model-literature, e.g. in Caurier et al. (2005) and Rowe and Wood (2010). However, in Boge (2020, Sects. 3.1 and 4), I discussed the (*acknowledged*) limitations of Caurier et al. (2005), and Rowe and Wood (2010, p. 581) express their expectations as follows:

With some confidence, we would say that “We are now at the end of the beginning and are moving to the beginning of the middle: a period that will be measured in decades, where a truly unified view of nuclear structure is within reach.”

The discussion of details is deferred to a forthcoming volume (cf. *ibid.*, p. 579), still missing some 10 years in. The hope for unification seems a faint one at best.

At the same time, there is also continuing interest in generalized drop models; for instance within the “hot research topic” of “[p]article emission intermediate in size between alpha and fission fragments” (Santhosh and Jose 2019, p. 1). Moreover, appreciating that sub-nuclear details cannot be fully neglected leads to so called ‘shell corrections’, which can be computed in different ways. But it turns out that “the parameters [of the drop model] marginally vary with the change of shell corrections,” and “the absolute and relative uncertainties are barely altered” (Cauchois et al. 2018, p. 7). In other words: The success of drop models remains almost entirely unaffected by how one conceives of sub-nuclear interactions. Hence we can safely assume $2 \leq k \leq 4$.

Let us now indulge a little numerical study of six possible cases, presented in Table 1. For concreteness, we assume that the only serious candidate for unification is some future shell model. As we can see, reasonably low values for $f_{k,\gamma}$ are possible only in the last two cases. In the fifth, however, one is overly confident in the model ($(1/k) + \gamma = 0.8$), given the above discussion on unification. In the sixth, one is in turn unnecessarily skeptical about success ($e = 0.6$), given the many predictive successes of the various SCTs (Greiner and Maruhn 1996).

Since we lack strong reasons for assuming unification on the horizon, but may take the expert witnesses to provide *some* indication for this, evidence suggests something

Table 1 Numerical study of values taken on by $f_{k,\gamma}(e, g)$ for different cases

$f_{k,\gamma}(e, g)$	k	γ	e	g	Case
0.30	4	0.50	0.75	0.90	Predictions of other models not covered by shell model; high confidence anyway
0.69	4	0.05	0.75	0.90	Predictions of other models not covered by shell model; according low confidence
0.53	2	0.10	0.75	0.90	Predictions of all but drop model covered by shell model; modest confidence
0.40	2	0.20	0.75	0.90	Predictions of all but drop model covered by shell model; intermediate confidence
0.15	2	0.30	0.75	0.90	Predictions of all but drop model covered by shell model; high confidence
0.14	2	0.15	0.60	0.85	Predictions of all but drop model covered by shell model; modest skepticism about success

like the second case. But then we are forced to assume *with about 70% confidence* that empirical testing in nuclear physics regularly gives back a positive response for severely false theories.

4.2 Psychometry: a second case?

The above case study is certainly interesting and, to my knowledge, has so far not been addressed in detail by realists. I believe, however, that one can make out at least one further example, this time from the social and mind sciences: Theories of psychological measurement. There are (at least) two SCTs that pose a strict incompatibility challenge, namely Item Response Theory (IRT) and Rasch Measurement Theory (RMT). A seminal source pointing to a fundamental conflict here is Andrich (2004), who displays the disagreement as a conflict between Kuhnian paradigms. We need not go down this Kuhnian road though, but should rather ask whether the above criteria for strict incompatibility apply.

The main point is that RMT models are, by design, equipped with a kind of separability between parameters, introduced by Rasch (cf. 1961, Section 5J) as necessary condition for what even counts as a ‘measurement’. That specific separability reflects a certain kind objectivity in comparison and is actually unique to RMT models (Andrich and Marais 2019, p. 329). Hence, it cannot be appropriately recaptured by IRT.

This so far only establishes *weak* incompatibility: RMT and IRT fundamentally disagree about the nature of measurement. But there are reasons to think that this disagreement actually amounts to an essential incompatibility, as will be seen below.

In a first approach, consider the following models for the psychological measurements of latent capacities (such as intellectual abilities) of a given test subject

in a test with dichotomous responses (yes or no), which responses can be right or wrong:

$$p(X_{ni} = 1 | \beta_n, \delta_i) = e^{\beta_n - \delta_i} / (1 + e^{\beta_n - \delta_i}), \quad (\text{RMT})$$

$$p(X_{ni} = 1 | \alpha_i, \beta_n, \delta_i) = e^{\alpha_i(\beta_n - \delta_i)} / (1 + e^{\alpha_i(\beta_n - \delta_i)}) \equiv p_{ni}, \quad (2P)$$

$$p(X_{ni} = 1 | \alpha_i, \beta_n, \gamma_i, \delta_i) = \gamma_i + (1 - \gamma_i)p_{ni}. \quad (3P)$$

Here, X_{ni} characterizes the response of the tested subject as either correct (1) or incorrect (0), and parameters with sole index i characterize properties of the item the subject is questioned on, whereas the β_n characterizes something about the subject indicative of the latent property to be measured by the test.

At this level of presentation, it may seem as though essential incompatibility can be dismissed right out of hand, because the two-parameter (2P) model obviously results from the three parameter (3P) model for $\gamma_i \rightarrow 0$, and the RMT model from that for $\alpha_i \rightarrow 1$. Hence, the acclaimed incompatible rivals appear to defy (**LC**). However, this impression is misguided, because it neglects the need to recover *successful predictions* (see Definition 5).

To understand why, or in what sense, successful predictions are spoiled when these connections between parameters are assumed, we first need to pay closer attention to the parameters' meanings. In particular, β_n symbolizes the 'proficiency' of the tested subject, i.e., a parameter characterizing the ability to perform in the given sort of test, given the relevant latent capacity, and δ_i the difficulty of the question (ignore α_i and γ_i for the moment). The restriction to these two parameters when choosing among the above models has profound implications. For instance, the RMT model implies that

$$p(X_{in} = 1 \wedge X_{jn} = 0 | (X_{in} = 1 \wedge X_{jn} = 0) \vee (X_{in} = 0 \wedge X_{jn} = 1)) = \frac{1}{1 + e^{\delta_i - \delta_j}}, \quad (11)$$

for any n and $i \neq j$, meaning that the probability for any particular succession of a correct and a false response, under the condition that some correct/false succession occurs on items i, j , depends solely on the relative difficulty of items i and j , not on the proficiency of the subject tested. Holding the index i fixed instead, a similar argument can be made for the comparison between performances of two subjects on the same item (cf. Andrich and Marais 2019, pp. 91–2).

This result generalizes to the total test score $r_n = \sum_{i=1}^k x_{in}$ being a 'sufficient statistic' for β_n , where k is the number of test-items and x_{in} the value taken on by X_{in} for subject n at question i ; i.e., $p(\bigwedge_{i=1}^k X_{in} = x_{in} | r_n)$ is independent of β_n . This generalizes the above result because conditioning on a particular test score is equivalent to conditioning on a large disjunction over all sequences giving rise to that same score. The meaning of this is that, relative to the 'baseline ability' of a given subject (as quantified by the score), the difficulty of the items can be characterized fully independently of that subject.

Rasch (1961) actually based his theory of measurement on a requirement of ‘specific objectivity’, which roughly says that comparisons between two test subjects (or items, respectively) must be independent of the item (or subject, respectively) on which they are compared. In other words, Rasch took it that, given two fixed test subjects, a test in which a comparison between them varies as a function of the parameter characterizing item-difficulty, not just as a function of their proficiencies, *simply does not measure one and the same latent capacity across all test items*.²¹

A consequence of specific objectivity is the separability between item difficulty and proficiency exemplified in formula (11) and the general result discussed thereafter. Moreover, it can be shown that, under a number of seemingly benign mathematical assumptions, *only* models that generalize formula (RMT) by introducing further *additive* terms or an overall multiplicative *constant* in the exponent (and nothing else), satisfy specific objectivity.²²

Since α_i varies with item, the (2P) and (3P) models are obviously not of the required form. For instance, trying to recapture formula (11) from the (2P) model, one would end up with $1/[1 + \exp(\alpha_i\delta_i - \alpha_j\delta_j + \beta_n(\alpha_j - \alpha_i))]$, which does not feature the same neat separation in probability between test items and subjects. Moreover, in the (2P) model, it can happen that the curves representing the probability of a correct response as a function of the subject’s proficiency *cross* for a number of different test-items (see Andrich and Marais 2019, p. 217). But this means that, with increased proficiency, a subject becomes better at responding to one item rather than another one, whereas for lower proficiency the order is reversed. The interpretation within RMT is that the test was ill-designed to test for one particular capacity in the first place.

Notably, it is exactly the positing/forbidding of additional parameters α_i , γ_i for the description of the item-subject relation that establishes / forestalls the success. Consider, for instance, the following sequence of events recaptured by Andrich (2004, p. 13). In an early application, Rasch encountered a mismatch between his model and the results of a Danish military intelligence test, but instead of revising the model, Rasch re-assessed the data and discovered “that the test seemed to be composed of different *kinds* of items” (ibid.; *emph. added*). Hence, they could not possibly all be used to assess the same latent capacity (or the same ‘facet’ thereof) in the sense of RMT.

When four further tests without this feature were designed by the military, these were expected to fit the model perfectly but again did not. Rasch then assessed the testing conditions and discovered that the test had been conducted under stronger time constraints. Correcting for items that were not answered due to a lack of time, rather than answered incorrectly, the expected fit was finally achieved.

A better fit could have probably been achieved immediately by introducing a more flexible models, like those of IRT. But that would have meant abandoning the underlying assumptions about objectivity, and would certainly *not* have lead to the

²¹ See also the foreword to Bond and Fox (2015).

²² See Fischer (1995a, b), Rasch (1961, 1977). Also Fischer (1995a, b) for proofs from different objectivity-conditions (meaning that the model uniquely represents a whole range of ways to codify objectivity) and the derivation of a non-dichotomous model.

discovery of the confounding variables. The point is that, by insisting on the aforementioned standards for objectivity in comparison, RMT *predicts* that any divergence from the resulting exponential curve can be accounted for in terms of confounding variables characterizing *test design* and *execution*, not properties of the variables defining the response of tested individuals to test-items. Given that these variables were repeatedly found by Rasch—and that upon correcting for them, his model achieved a very good fit—this constitutes a predictive success of RMT.

This is an isolated instance, and it remains unclear whether RMT is able to account for all suitable psychological data in similar ways. That is acknowledged especially by critics (e.g. von Davier 2016, p. 45), who in turn can also claim various empirical successes on the basis of IRT models, or rather, the parameters they make available.²³

Now it is of course conceivable that (2P) and (RMT) happen to agree on some data set—i.e., that $\alpha_i = 1, \forall i$ happens to attain a good fit. Indeed, IRTists usually include a ‘1P’ model in their list which formally corresponds to model (RMT); something lamented as creating confusion by advocates of RMT (e.g. Bond and Fox 2015; Andrich and Marais 2019). But this cannot be seen as establishing the relevant sense of continuity: Recall that (LC) requires one theory to recover *each* prediction of the other by means of some limiting procedure. Hence, so long as there is no well-defined domain of tests where RMT models must be applied for some good reason (even if initially not fit to data, as in the danish military example), whereas (2P) or (3P) achieve a good fit *more generally*, there is no ‘ $\alpha_i \rightarrow 1$ -limit’ which satisfies success-preservation. The ability to let the additional parameters that violate Rasch-objectivity vanish *locally* must be construed rather as a ‘lucky coincidence’.

Since offering grounds for the restriction to / introduction of certain parameters are considerably weak posits, and there appear to be no weaker posits in use that could be ‘doing the real work’ here, this establishes strict incompatibility in the sense of Definition 1. This in turn fully sanctions setting $k \geq 2$ for psychometry. With the prospects of unification at best unclear, this leaves us in something like the third row of Table 1. But that means that we should be at least more certain than not that successful empirical use of theories in the RMT and IRT families leads us severely astray about the nature of measurement and latent human capacities.

Now I expect some controversy to arise here as follows: Is it really a success if data can be made to conform *to the model*? I think the case discussed by Andrich (2004) clearly is a predictive success, as Rasch made a genuine discovery about distorting peculiarities of test design and execution. On the other hand, is it really a success if a multi-parameter model like 3P can be *fitted to data*? By itself, the presence of parameters cannot be counted against the possibility of predictive success: The famous Standard Model of particle physics has not four but 19 free parameters, and nevertheless counts as immensely successful by most standards.

The possibility of predictive success despite parameters appears to depend on the *meaning* of these. In the Standard Model, they are, e.g., couplings and particle masses; things that are likely ‘just empirical’. In the 3P model, α_i and γ_i can be

²³Lord (1980, p. 16), for example, is a classic source in this tradition which presents some data from verbal SATs in pretty good agreement with model (3P).

interpreted meaningfully as quantifying the discrimination-strength of items and the impact of guessing respectively; things likely just empirical as well.

There is another issue to be addressed here, namely that the Rasch-prediction of confounding variables is *wholesale*, not specific or even quantitatively precise. As Vickers (2019) points out, such predictions are not very ‘risky’ (i.e., not so easily brought into conflict with experience), and hence maybe not worthy of doxastic commitment at all.

I admit that this makes the case weaker, but I don’t think it invalidates it completely. For it is safe to say that supplementary assumptions in the vein of, say, a model of the testing conditions in the RMT case are needed for most (if not all) predictive successes in science.²⁴ For instance, referring to the celebrated Standard Model once more, it is hardly possible to draw *any* inferences to its validity from LHC data without including phenomenological physics models of hadron formation, additional scattering events, and even detector responses.

Whether this spoils the genuine *success* of the given theory must be decided case by case, and I don’t think it is problematic to assume that, if RMT is correct, any confounding variables must be modeled as extraneous to the actual measurement—so long as this not only restores the fit between data and model but also involves independently testable quantities (like time constraints or the comparability of test-items). Hence, I submit that psychometry provides another worthwhile example to be examined by philosophers under the header of the overall argument presented here.

4.3 Beyond individual fields

Let us now consider why $p(t|c) \leq (1/k)$ and $k \geq 2$, together with the evidence presented above, does not merely establish something about nuclear physics or psychometry. As pointed out throughout this paper, SSR is, in essence, a response to the traditional PMI. As Poincaré (1905, p. 163) put it:

When a physicist finds a contradiction between two theories [...] one of them at least should be considered false. But this is no longer the case if we only seek in them what should be sought. It is quite possible that they both express true relations, and that the contradictions only exist in the images we have formed to ourselves of reality.

As was pointed out in Section 3.4, the stronger PMI developed here is meant as a further attack on the truth-success connection claimed by the SSRist: That it is hardly possible to explain scientific success if not in terms of the approximate truth of at least the essential posits. The central point of the present argument hence is this: Given that it is sometimes the very working posits that contradict each other in one and the same domain, piecing even these together as an account of what the respective field of research might be getting right about reality means embracing outright inconsistencies.

²⁴See, e.g., Morrison’s (2009) careful investigation of models in measurement, or Duhem’s (1914) reasons for claiming confirmational holism and underdetermination.

Hence, for a non-negligible number of successful scientific theories, the prevalence of success among them *must* be explained in some other way. As a *general* response to the traditional PMI, the reference to ‘working posits’ therefore becomes unsatisfactory. Or in yet other words: Despite having tackled the problem of fundamental incompatibilities between past and present successful theories, SSRists now face the subsequent challenge of explaining why working posits should *sometimes* deserve doxastic commitment, whereas sometimes they clearly *don’t*.

5 Conclusions

In this paper, I have extended the recent debate over the NMA into a stronger PMI that provides a direct challenge for SSR. This was done by fleshing out local incompatibilities between (classes of) successful theories, as recognized in Boge (2020), in terms of posits that are *essential* relative to those classes, and hence the proper target of doxastic commitment according to SSR.

A foreseeable response by the SSRist could be disputing the relevance of the case studies, because the theories in question are ‘not sufficiently mature’. However, this issue was addressed explicitly for nuclear physics in Boge (2020, 4351), and for very similar reasons seems implausible also for psychometry. Another response might be that these are highly isolated cases that can be dissolved in a suitable way, as has been done (though maybe still somewhat tentatively) with other cases that were in general considered highly problematic for realism (see, in particular, Vickers 2020).

So long as such a dissolution is missing, however, the present argument makes it considerably less attractive to stick to the NMA intuition, and considerably more attractive to try and explain scientific success in some alternative way. I can here only gesture at the possibility of doing so: For instance, the prevalence of success in science could be possibly explained by some variant of van Fraassen’s (1980) Darwinian account, recently also appraised and extended by K. B. Wray (2018). On the other hand, the apparent presence of so many false positives that the present argument forces upon us might be understood in terms of Frost-Arnold’s (2019) recent account of *misleading* evidence.

In any case, given that the SSRist can equally only point to a limited number of examples where the relevant working posits have been preserved in a suitable way throughout theory change, the strong, local PMI developed here must be counted as a serious challenge to SSR.

Acknowledgements I thank two anonymous referees for valuable comments.

Funding Open Access funding enabled and organized by Projekt DEAL. Some research for this paper was done during my employment with the research unit *The Epistemology of the Large Hadron Collider*, funded by the German Research Foundation (DFG; grant FOR 2063).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the

material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, 42(1), I7–I16.
- Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: measuring in the educational, social and health sciences. Springer Texts in Education*. Singapore: Springer.
- Asay, J. (2012). A truthmaking account of realism and anti-realism. *Pacific Philosophical Quarterly*, 93(3), 373–394.
- Asay, J. (2013). Three paradigms of scientific realism: a truthmaking account. *International Studies in the Philosophy of Science*, 27(1), 1–21.
- Asay, J. (2019). Going local: a defense of methodological localism about scientific realism. *Synthese*, 196(2), 587–609.
- Basdevant, J.L., & Dalibard, J. (2006). *Quantum mechanics*, 2nd edn. Berlin: Springer.
- Blatt, J.M., & Weisskopf, V.F. (1979). *Theoretical nuclear physics*. New York: Springer.
- Boge, F.J. (2020). An argument against global no miracles arguments. *Synthese*, 197, 4341–4363.
- Bond, T., & Fox, C.M. (2015). *Applying the Rasch model: fundamental measurement in the human sciences*. Taylor & Francis (3rd edn).
- Boyd, R. (1980). Scientific realism and naturalistic epistemology. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1980(2), 613–662.
- Boyd, R. (1984). The current status of the realism debate. *Scientific Realism* 41–82.
- Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago Press.
- Cauchois, B., Lü, H., Boilley, D., Royer, G. (2018). Uncertainty analysis of the nuclear liquid drop model. *Physical Review C*, 98(2).
- Caurier, E., Martinez-Pinedo, G., Nowacki, F., Poves, A., Zuker, A.P. (2005). The shell model as a unified view of nuclear structure. *Reviews of Modern Physics*, 77(2), 427.
- Cevolani, G., & Schurz, G. (2017). Probability, approximate truth, and truthlikeness: more ways out of the preface paradox. *Australasian Journal of Philosophy*, 95(2), 209–225.
- Chakravartty, A. (2017). Scientific realism. In E.N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition.
- Dawid, R., & Hartmann, S. (2018). The no miracles argument without the base rate fallacy. *Synthese*, 195(9), 4063–4079.
- De Haro, S., & de Regt, H. (2018). Interpreting theories without a spacetime. *European Journal for Philosophy of Science*, 8, 631–670.
- Devitt, M. (1984). *Realism and truth*. Blackwel.
- Devitt, M. (2011). Are unconceived alternatives a problem for scientific realism? *Journal for General Philosophy of Science*, 42(2), 285–293.
- Douven, I. (2017). Abduction. In E.N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition.
- Duhem, P. (1954[1914]). *The aim and structure of physical theory* (2nd edn). Princeton University Press. translated by Philip P. Wiener.
- Duncan, A., & Janssen, M. (2019). *Constructing quantum mechanics: volume 1: the Scaffold: 1900–1923*. Oxford: Oxford University Press.
- Fischer, G.H. (1995a). The derivation of polytomous rasch models. In G.H. Fischer, & I.W. Molenaar (Eds.) *Rasch models* (pp. 293–306). Berlin: Springer.
- Fischer, G.H. (1995b). Derivations of the Rasch model. In G.H. Fischer, & I.W. Molenaar (Eds.) *Rasch models* (pp. 15–38). Berlin: Springer.
- Frost-Arnold, G. (2019). How to be a historically motivated antirealist: the problem of misleading evidence. *Philosophy of Science*, 86(5), 906–917.
- Good, I. (1983). *Good thinking: the foundations of probability and its applications*. Mineapolis: University of Minnesota Press.
- Greiner, W., & Maruhn, J.A. (1996). *Nuclear models*. Berlin: Springer.

- Henderson, L. (2017). The no miracles argument and the base rate fallacy. *Synthese*, 194(4), 1295–1302.
- Howson, C. (2000). *Hume's problem: induction and the justification of belief*. Oxford: Oxford University Press.
- Howson, C. (2013). Exhuming the no-miracles argument. *Analysis*, 73(2), 205–211.
- Howson, C., & Urbach, P. (2006). *Scientific reasoning. The Bayesian Approach* (3rd edn). Open Court: Chicago.
- Kamal, A. (2014). *Nuclear physics*. Berlin: Springer.
- Kitcher, P. (1993). *The advancement of science: science without legend objectivity without illusions*. Oxford: Oxford University Press.
- Kyriacou, D.N. (2001). Reliability and validity of diagnostic tests. *Academic Emergency Medicine*, 8(4), 404–405.
- Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S.G. Harding (Ed.) *Can theories be refuted? Essays on the Duhem-Quine Thesis* (pp. 205–259). Netherlands: Springer.
- Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, 48(1), 19–49.
- Laudan, L. (1984). Realism without the real. *Philosophy of Science*, 51(1), 156–162.
- Leifer, M.S. (2007). Conditional density operators and the subjectivity of quantum operations. *AIP Conference Proceedings*, 889(1), 172–186.
- Lewis, P.J. (2001). Why the pessimistic induction is a fallacy. *Synthese*, 129(3), 371–380.
- Lipton, P. (2004). *Inference to the best explanation*. Taylor & Francis.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Routledge.
- Magnus, P. (2010). Inductions, red herrings, and the best explanation for the mixed record of science. *The British Journal for the Philosophy of Science*, 61(4), 803–819.
- Magnus, P.D., & Callender, C. (2004). Realist ennui and the base rate fallacy. *Philosophy of Science*, 71(3), 320–338.
- Mizrahi, M. (2013). The pessimistic induction: a bad argument gone too far. *Synthese*, 190(15), 3209–3226.
- Morrison, M. (2009). Models, measurement and computer simulation: the changing face of experimentation. *Philosophical Studies*, 143(1), 33–57.
- Morrison, M. (2011). One phenomenon, many models: inconsistency and complementarity. *Studies in History and Philosophy of Science Part A*, 42(2), 342–351.
- Niiniluoto, I. (2004). Truth-seeking by abduction. In F. Stadler (Ed.) *Induction and deduction in the sciences* (pp. 57–82). Berlin: Springer.
- Park, S. (2019). Localism vs. individualism for the scientific realism debate. *Philosophical Papers*, 48(3), 359–377.
- Peters, D. (2014). What elements of successful scientific theories are the correct targets for ‘selective’ scientific realism *Philosophy of Science*, 81(3), 377–397.
- Poincaré, H. (1905). *Science and hypothesis*. Walter Scott Publishing Co., Ltd.
- Poincaré, H. (1907). *The value of science*. The Science Press.
- Poole, D. (2000). Learning, bayesian probability, graphical models, and abduction. In P.A. Flach, A.C. Kakas, D. Poole (Eds.) *Abduction and induction* (pp. 153–168). Berlin: Springer.
- Psillos, S. (1996). Scientific realism and the ‘pessimistic induction’. *Philosophy of Science*, 63, S306–S314.
- Psillos, S. (1999). *Scientific realism. How science tracks truth*. London: Routledge.
- Psillos, S. (2009). *Knowing the structure of nature: essays on realism and explanation*. Basingstoke: Palgrave Macmillan.
- Putnam, H. (1975a). *Mathematics, matter and method. Philosophical papers, Volume I*. Cambridge: Cambridge University Press.
- Putnam, H. (1975b). What is “realism”? *Proceedings of the Aristotelian Society*, 76, 177–194.
- Putnam, H. (1978). *Meaning and the moral sciences*. New York: Routledge.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, (Vol. 4 pp. 321–333).
- Rasch, G. (1977). On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. In M. Blegvad (Ed.) *The Danish yearbook of philosophy*. Copenhagen: Munksgaard.
- Romeijn, J.-W. (2013). Abducted by bayesians? *Journal of Applied Logic*, 11(4), 430–439.
- Rowe, D., & Wood, J. (2010). *Fundamentals of nuclear models: foundational models*. Singapore: World Scientific.

- Ruhmkorff, S. (2013). Global and local pessimistic meta-inductions. *International Studies in the Philosophy of Science*, 27(4), 409–428.
- Saatsi, J.T. (2005). On the pessimistic induction and two fallacies. *Philosophy of Science*, 72(5), 1088–1098.
- Santhosh, K., & Jose, T.A. (2019). Alpha and cluster decay using modified generalized liquid drop model with biso-spin dependent pre-formation factor. *Nuclear Physics A*. <https://doi.org/10.1016/j.nuclphysa.2019.121626>.
- Schurz, G. (2009). When empirical success implies theoretical reference: a structural correspondence theorem. *The British Journal for the Philosophy of Science*, 60(1), 101–133.
- Schurz, G. (2011). Structural correspondence, indirect reference, and partial truth: phlogiston theory and Newtonian mechanics. *Synthese*, 180(2), 103–120.
- Schurz, G. (2014). *Philosophy of science. A unified approach*. New York: Routledge.
- Schurz, G. (2017). Truthlikeness and approximate truth. In J. Saatsi, & G. Schurz (Eds.) *The Routledge handbook of scientific realism* (pp. 133–148). Routledge.
- Schwartz, M. (2014). *Quantum field theory and the standard model*. Cambridge: Cambridge University Press.
- Sesardic, N. (2007). Sudden infant death or murder? A royal confusion about probabilities. *The British Journal for the Philosophy of Science*, 58(2), 299–329.
- Sober, E. (2002). Instrumentalism, parsimony, and the akaike framework. *Philosophy of Science*, 69(S3), S112–S123.
- Spanos, A. (2019). *Probability theory and statistical inference*. Cambridge: Cambridge University Press.
- Stanford, P.K. (2006). *Exceeding our grasp: science, history and the problem of unconceived alternatives*. Oxford: Oxford University Press.
- van Fraassen, B.C. (1980). *The scientific image*. Oxford: Oxford University Press.
- van Fraassen, B.C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.
- van Fraassen, B.C. (2006). Structure: its shadow and substance. *The British Journal for the Philosophy of Science*, 57(2), 275–307.
- van Fraassen, B.C. (2008). *The empirical stance*. Yale University Press.
- van Fraassen, B. (2010). *Scientific representation: paradoxes of perspective*. Oxford: Oxford University Press.
- Vickers, P. (2013). A confrontation of convergent realism. *Philosophy of Science*, 80(2), 189–211.
- Vickers, P. (2017). Understanding the selective realist defence against the PMI. *Synthese*, 194, 3221–3232.
- Vickers, P. (2019). Towards a realistic success-to-truth inference for scientific realism. *Synthese*, 196(2), 571–585.
- Vickers, P. (2020). Disarming the ultimate historical challenge to scientific realism. *The British Journal for the Philosophy of Science*, 71(3), 987–1012.
- von Davier, M. (2016). Rasch model. In W.J. van der Linden (Ed.) *Handbook of item response theory* (Vol. 1, pp. 59–76). Chapman and Hall/CRC.
- Weizsäcker, C.V. (1935). Zur Theorie der Kernmassen. *Zeitschrift für Physik A: Hadrons and Nuclei*, 96(7), 431–458.
- Williams, P. (2019). Scientific realism made effective. *The British Journal for the Philosophy of Science*, 70(1), 209–237.
- Worrall, J. (1989). Structural realism: the best of both worlds? *Dialectica*, 43(1–2), 99–124.
- Worrall, J. (2011). The no miracles intuition and the no miracles argument. In D. Dieks, W. J. Gonzalez, S. Hartmann, T. Uebel, M. Weber (Eds.) *Explanation, prediction, and confirmation* (pp. 11–21). Berlin: Springer.
- Worrall, J. (2020). Structural realism: The only defensible realist game in town? In W.J. Gonzales (Ed.) *New approaches to scientific realism* (pp. 169–205). de Gruyter.
- Wray, K.B. (2013). Success and truth in the realism/anti-realism debate. *Synthese*, 190(9), 1719–1729.
- Wray, K.B. (2015). Pessimistic inductions: four varieties. *International Studies in the Philosophy of Science*, 29(1), 61–73.
- Wray, K.B. (2018). *Resisting scientific realism*. Cambridge: Cambridge University Press.