Three Nightmares of the Inductive Mind

Of all the disquieting riddles and paradoxes found in the arsenal of epistemological scepticism—understood as a *systematic* and *piecemeal* scrutiny into the methods and paradigms of the formation and justification of knowledge-claims—one problem has particularly proved, time and again, to be a never-ending source of cognitive vexation. With few notable exceptions, philosophers and philosophically-minded scientists and statisticians (e.g., Hume, Goodman, Putnam, Stegmüller, Boltzmann and De Finetti among others) have invariably either downplayed and deflected the seriousness of this problem and its variations, or have simply given up worrying about it in the hope that it may miraculously disappear. The said problem is nothing but David Hume's strong version of the problem of induction which, unbeknownst to Hume himself, was destined to become the superacid of methodological scepticism capable, in the blink of an eye, of eating away the foundation of any epistemic project built on naïve forms of empiricism and rationalism.

It is often the case that philosophers who pose sceptical problems recoil in fear once they realize the far-reaching implications of such problems, and Hume, with his problem of induction, is no exception. They rush into defusing their inadvertent exercise in scepticism. But systematic scepticism is something akin to an explosive chemical chain reaction. Once it is set off, with every passing minute it becomes more difficult to extinguish the flames. Pour on more water, and the fire spreads to areas you never imagined flammable. A genuine philosopher—regardless of her alliances—seeks to examine how far the fire spreads. Methodological scepticism is a scandal to be recognized and investigated, not ignored or swept under the carpet. It is only through systematic and rational scepticism that philosophy might be able to fundamentally shake the idleness of thought and entrenched beliefs. A philosophy—whether it is aligned to materialism or realism, empiricism or rationalism—that does not recognize the force of rigorous scepticism or take on its challenges is not worth its name. Accordingly, the aim here is not to dismiss the investigative power of systematic scepticism or to simply put up with its quandaries, but to embrace and exacerbate it as nothing but a rational critical challenge of the utmost conceptual severity.

Here I shall attempt to reappropriate Hume's problem as a broad and effective critique of inductivist and deductivist trends in philosophy of knowledge and philosophy of mind, including cognitive sciences and particularly the program of artificial general intelligence. By inductivism and deductivism, I broadly mean any approach to knowledge and mind that claims that either a purely inductive method or a purely deductive method *alone* is sufficient for the formation of knowledge claims or the realization of mind's structuring powers. To this extent, the aim of this renewed engagement with Hume's problem is twofold:

(*a*) Expanding the analysis of the problem of induction in its Humean form to its more recent reformulations by the likes of Nelson Goodman and Hilary Putnam. This will allow us not only to develop a more in-depth understanding of the nature of this problem, but more importantly, to differentiate and separately address three distinct predicaments which, in the Humean version of the problem of induction, are treated as one. These predicaments can be classified as quandaries of *retrodiction*, *prediction*, and *formalization* which respectively pose challenges to the epistemic status of inductive inferences on three different levels:

(1) **The reliability hypothesis of memory** which secures the accuracy or factuality of derived empirical data and the history of past observations.
(2) **The reliability hypothesis of law-like statements confirmed by evidence** which ensures the adequacy of the role of evidence in confirming hypotheses either in the context of the inductivist theory of confirmation where positive single instances together with projectable predicates count as sufficient criteria of confirmation, or in the context of the deductivist theory of corroboration in which hypotheses are selected to be tested against counterexamples or negative single instances.
(3) **The formal and epistemological completeness of inductive models** according to which a purely inductive agent or intelligence can provide a non-arbitrary description of not only the external world but also the (inductive) model of mind it inhabits.

(*b*) Concomitant with this analysis, we shall also focus on the import of the problem of induction for philosophy of mind and the project of artificial general intelligence. We shall see that the same predicaments that challenge the epistemic legitimacy of induction also threaten the coherency of those strains of AI in which intelligence is simply equated with predictive induction and prediction is defined by the information-theoretic concept of compression. It is often assumed that the formal-computational account of Occam's principle of simplicity as put forward by algorithmic information theory—specifically, Ray Solomonoff's account of induction which is couched in terms of the duality of regularity and compression—circumvents the epistemic quandaries of induction. However, in dispensing with the specificity of the theoretical-semantic *context* in which the principle of simplicity finds its significance as a *pragmatic* tool, the formal generalization of Occam's razor as the cornerstone of all existing computational models of induction not only finds itself faced with the predicaments harboured by the problem of induction but also results in a number of new complications such as arbitrariness and computational resource problems.

Modelling general intelligence on purely inductive inferences is seen by the current dominant trends in AGI research as an objective realist index of general intelligence. In the same vein, posthumanism treats the inductive models of general intelligence as an evidence against the exceptionalism of the conceptualizing human mind as a *sui generis* criterion that sets apart general intelligence as a qualitative dimension from quantitative intelligent problem-solving behaviours. Yet, as will be argued, the formal generalization of Occam's razor as a means of granting induction a privileged role capable of replacing all other epistemic activities, along with the equation of general intelligence with induction, turn out to be precisely the fruits of the human's experiential-cognitive biases.

All in all, my aim is to argue that the force of epistemological scepticism as expressed in the problem of induction can be understood not only in terms of a formidable challenge to entrenched philosophical dogmas and cognitive biases, but also in terms of a razor-sharp critique of purely inductive models of mind and inductivist trends in artificial general intelligence.

Irrespective of their specificity, all models of AGI are built on implicit models of rationality. From the early Carnapian learning machine to Solomonoff's prediction as the model of an optimal or universal learning machine to Marcus Hutter's equation of compression with general intelligence and, more recently, the Bayesian program of rational AGI as proposed by Eliezer Yudkowsky, inductivist models are no exception. In this respect, there is certainly a discussion to be had about the sociocultural and political dimension of such trends: What is it exactly in the inductivist models of rationality or approaches

to general intelligence that make them susceptible to appropriation by superintelligence folklores or, worse, by ideologies which champion instrumentalist or even social Darwinist conceptions of intelligence? Rather than answering this question, I intend to take a different approach: A socio-political critique by itself is not, by any means, adequate to challenge such trends in cognitive sciences; nor is a well-constructed rationalist critique, which often devolves into quibbles over whose model of general intelligence or rationality is better. These trends should instead be challenged in terms of their own assumptions and debunked as not only unfounded but also logically erroneous.

A Humean Provocation

To understand the exact nature of Hume's problem of induction, let us first reconstruct it in a more general form and then return to Hume's own exposition of the problem. But before we do so, it would be helpful to provide brief and rudimentary definitions of deduction and induction.

Deduction can be defined as a form of reasoning that links premises to conclusions so that, if the premises are true, following step-by-step logical rules, then the conclusion reached is also *necessarily* true. Deductive inferences are what Hume identifies as demonstrative inferences. In the Humean sense, a demonstrative inference is strictly speaking an inference where a pure logical consequence relation is obtained. Such a logical relation has two formal characteristics: (1) there is no increase in the content of the conclusion beyond the content of the premises. Therefore, demonstrative inferences in their Humean sense can be said to be non-ampliative inferences, i.e. they do not augment the content or add anything other than what is already known; (2) the truth of premises is carried over to the conclusion. Accordingly, demonstrative inferences are truth-preserving. It is important to note that non-ampliativity and truth-preservation are—pace some commentators e.g., Lorenzo Magnani[1]—two separate features and do not by any means entail one another. All that truth-preservation implies is the transferability of the truth of the premises to the conclusion. It does not say anything regarding the augmentation of the content or lack thereof, nor does it exclude the possibility that, if new premises are added, the truth of the conclusion may change. Therefore, Magnani's claim that non-monotonicity—as for example captured by substructural logics—stands in contrast to both the non-ampliative and the truth-preserving characters of demonstrative inferences is based on a confusion.

In contrast to deductive-demonstrative inferences, inductive inferences cannot be as neatly formulated. However, roughly speaking, induction is a form of inference in which premises provide strong support (whether in causal, statistical, or computational terms) for the *outcome*—as distinguished from the deductive *conclusion*—of the inference. Whereas the truth of the conclusion in deductive reasoning is logically certain, the truth of the outcome in inductive reasoning is only *probable* in proportion to the supporting evidence. Hence, as evidence piles up, the degree of supporting valid statements for a hypothesis indicate that false hypotheses are—as a matter of generalization—probably false and, in the same vein, that true hypotheses are—as a matter of generalization—probably true. But this dependency on evidence also means that inductive inference is contingent and non-monotonic. Non-monotonicity means that the addition of new premises can fundamentally change the truth of the conclusion, either

[1] L. Magnani, *Abductive Cognition* (Dordrecht: Springer, 2009).

drastically raising or lowering the degree of support already established for the inductive outcome. The significance of induction is that it permits the differentiation of laws from non-laws. This is precisely where the problem of induction surfaces.

Now, with these clarifications, Hume's problem of induction can be formulated quite generally without being narrowed down to a special class of non-demonstrative inferences (e.g., induction by enumeration) as follows:

A) Our knowledge of the world must be at least at some level based upon what we observe or perceive, insofar as purely logical reasoning by itself cannot arrive at knowledge. We shall call this the problem of synthetic content of knowledge about the world.

B) Despite A, we take our factual knowledge of the world to exceed what we have acquired through mere observation and sensory experience. However, here a problem arises. Let us call it *problem 1*: How can we justify that our knowledge of the unobserved is really knowledge? At this point, the central issue is the problem of justification rather than the problem of discovery in the sense that, for now, it does not matter how we have attained this supposed knowledge. Thus, the Kantian claim that Hume confuses *quid facti* and *quid juris*—the origination of knowledge claims and the justification of knowledge claims—and that the problem of induction applies only to the former, simply misses the point.

C) Since the justification of the alleged knowledge in B cannot be a matter of logical demonstration, it must then be obtained by way of arguments whose premises are observations of the kind cited in A and the conclusion must be a kind of knowledge which goes beyond observation, i.e. knowledge of the kind mentioned in B.

D) Two tentative solutions can be provided for the problem of justification characterized in C:

D-1) Justification by demonstrative arguments: But as argued earlier, demonstrative inferences do not augment the content (i.e. they are not ampliative). That is to say, if the premises are the observed, nothing beyond the content of the observed is yielded in the conclusion. The knowledge of the unobserved (B) is included in the observed premises (A). But this contradicts B, since the kind of knowledge it assumes must to go beyond mere observation. In other words, the content of the statements regarding the alleged knowledge of the unobserved is not contained in the content of the knowledge of the observed. The first solution, however, does not permit this because in demonstrative arguments, the content of the conclusion is necessarily included or contained in the content of the premises.

D-2) Justification by non-demonstrative arguments: If the non-ampliative nature of the demonstrative inference does not allow the transition from the observed to the unobserved, then the solution to the problem of justification would have to involve ampliative inferences. But is there such a thing as an ampliative inference? The answer is yes: take a logically invalid inference in which the conclusion has an augmented or stronger content in comparison to the conjunction of premises. However, this is not what the inductivist qua proponent of the alleged knowledge of the unobserved is looking for. In logic, invalid logical inferences may be of some interest, but when it comes to knowledge justification, they are merely absurd. Therefore, the ampliative inference *by itself* is not a sufficient condition for the kind of argument that can answer the problem of justification. The solution to the problem of justification would require an inference in which not only is the content of the conclusion

stronger than that of the premises, but also the tentative truth in the content of the premises—knowledge of the observed—can be transferred to the conclusion, i.e. the knowledge of the unobserved (cf. the earlier note on the distinction between ampliative and truth-preserving features). In short, the solution to the problem of justification would require a truth-preserving (of demonstrative arguments) and ampliative (of non-demonstrative arguments) inference. This is the generalized form of Hume's problem of induction: Is there such an inference—generally speaking and with no reference to specific inductive rules—one that is both truth-preserving and ampliative, demonstrative and non-demonstrative?

E) There is no such non-demonstrative inference as characterized in D-2. An inference is either ampliative such that the content is extended—but then there is no guarantee that the conclusion is true even if all premises are true; or it is truth-preserving—in which case the conclusion is true and the inference is valid but the content is not augmented. Said differently, the problem of justification concerning the inferential transition of the knowledge of the observed to the knowledge of the unobserved will end up with either content-augmenting but non-truth-preserving inferences, or truth-preserving but non-content-augmenting inferences. Neither case can justify the transition of the observed premises to the unobserved conclusion.

F) At this point, the inductivist might argue that in order to resolve the problem of justifying the transition from the observed to the unobserved, we must abandon the absurd idea of a truth-preserving ampliative inference and instead replace it with a probability inference. That is to say, we should understand inductive inferences strictly in terms of probability inferences. But this supposed solution does not work either. For either what is *probable* is construed in terms of frequency, so that the more probable is understood as what has so far occurred more often—in which case we are again confronted with the same problem: How do we know that a past frequency distribution will hold in the future? To arrive at and justify such knowledge, we must again look for a truth-preserving ampliative inference, of the very sort that the inductivist has claimed to be a futile enterprise. Or the probable is construed in a different sense, which then raises the question of why we should anticipate that the more probable—in whatever sense it has been formulated—will be realized rather than the improbable. In attempting to answer this question, we face the same problem we initially sought to resolve. Thus, the nature of the problem of induction turns out to be not limited to the truth-preserving schema of justification which some commentators have claimed to be 'the outcome of the strictures imposed by the deductive paradigm underlying the classical view of scientific demonstration'.[2]

G) Considering E and F, the nature of Hume's problem of induction is revealed to be—contrary to common interpretations—not about the origination of knowledge claims (*quid facti*) or even the justification of induction (problem 1), but about a far more serious predicament: we cannot extrapolate knowledge of the unobserved from knowledge of the observed.

If we extend the conclusion reached in G from what has been observed so far and what has not been observed yet to knowledge of the past and the future, Hume's problem of induction then boils down to the claim that *we cannot possibly gain knowledge of the future.* It should be clarified that this claim does not

[2] N. B. Goethe, 'Two Ways of Thinking About Induction', in *Induction, Algorithmic Learning Theory, and Philosophy* (Dordrecht: Springer, 2007), 238.

mean that our knowledge of the future can never be certain either in the deductive sense or in the sense of the probable—something that any inductivist would accept—but rather that our contention regarding the possibility of having such knowledge is irrational, i.e. not reasonable in any sense. To put it more bluntly, there simply cannot be any (inductive) knowledge of the future to be deemed certain or uncertain, determinate or wholly indeterminate in any sense. In a contemporary formulation, we can express such knowledge within a system of inductive logic where we interpret probability in the sense of degree of confirmation—the most fundamental concept of inductive logic which characterizes 'the status of any scientific hypothesis, e.g., a prediction or a law, with respect to given evidence'.[3] Such a purely logical concept of probability is distinguished from probability in the sense of 'the relative frequency of a kind of event in a long sequence of events'[4] which can be entirely couched in statistical terms. We can respectively call these two different but related senses of probability, the logical or confirmationist and the frequentist accounts of probability, or, using Carnap's classification, probability$_1$ and probability$_2$. In the sense of the logical concept of probability$_1$, knowledge of the unobserved or the future can be expressed by way of Carnap's formula $c(h, e)=r$, where

- c is a confirmation function (or alternatively, a belief function in the context of formal learning theory) whose arguments are the effective hypothesis h and the empirical data e. Theorems of inductive logic which hold for all regular c-functions can be such theorems as those of Bayes and classical probability theory.
- h is an effective or computable hypothesis expressing supposed universal laws or a singular statement or assertion about the future such that
    - h is expressible in a language L rich enough to support a description of space-time order and elementary number theory;
    - If it is a consequence of h that $M(x_i)$ is true (where M is a molecular predicate of L and x_i an individual constant running through the names of all the individuals), then it can be said that $h \supset M(x_i)$ is provable or alternatively, computable in L;
    - h is equivalent to a set of (computable) sentences of the forms $M(x_i)$ and $\neg M(x_i)$. For example, $M(x_1)$ can be read as Is-Green(*this emerald*). In this sense, 'x_1 is green' or 'x_1 is not green' means that the position x_1 is occupied (or is not occupied) by something green, or that green occurs (or does not occur) at x_1. More accurately, the predicate is attached to the description of the individual's arrangement (e.g., this emerald) in the space-time continuum.
    - The outcome of the inductive reasoning is concerned not so much with the acceptance, rejection or temporary suspension of h as with finding the numerical value of the probability of h on the basis of e. This means that even though a thought or a decision—or more generally judgements about h—are not explicitly framed as a probability statement, they can nevertheless be reconstructed as a probability statement.

  Now, if a hypothesis implies that each individual satisfies the molecular predicate $M_1(x) \supset M_2(x)$, then for each i, $M_1(x_i) \supset M_2(x_i)$ should be deducible from h in L, in order for h to qualify as an effective hypothesis.

[3] R. Carnap, *Logical Foundations of Probability* (Chicago: University of Chicago Press, 1950), viii.
[4] Ibid.

- e is a statement about the past or a report about the observed, evidence or empirical data.
- r is a real number denoting the quantitative degree of confirmation or the degree of belief such that $0 \leq r \leq 1$. It is represented by a measure function P i.e. an a priori probability distribution. In this sense, $c(h, e) \leq 1$, $c(h, e) \geq 0.9$ and $c(h, e) \geq 0.5$ mean that, depending on the particular system of inductive logic and the inductive theory of confirmation, the degree by which the statement expressed by h is *confirmed* by the statement expressed by e approaches 1, converges on or remains greater than 0.9 or, in the weakest condition of confirmation, becomes and remains greater than 0.5. Expressing the value of r in terms of a limit function as approaching or remaining greater means that either of the above scenarios admits of exceptions.
    - For an initial n-membered segment of a series with the molecular property M that has been observed m times, we can anticipate that the relative frequency of the members observed with the characteristic M—or the m-membered sub-class of M—in the entire series should correspond to $1 - m/n$. For example, in the case of 0.5, for every n we can expect to find an m such that, if the next m-membered individuals ($x_{n+1}, x_{n+2}, \ldots, x_{n+m}$) are all M, then the degree of confirmation of the effective hypothesis $M(x_{n+m+1})$ is greater than 1/2 irrespective of the characteristic of the first n-individuals. Suppose n=8, then it must be possible to find an m—let us say $m=10^7$—such that we can state that $x_9, x_{10}, \ldots, x_{10000000}$ being all green, then it is probable more than one-half that $x_{100000001}$ will also be green *even if* $x_1, x_2, \ldots, x_8$ are not green.
    - The criterion for the adequacy of the measure function is that for every true computable hypothesis h, the instance confirmation $P(M(x_{n+1})) \mid M(x_0), \ldots, M(x_n)$ should at the very least converge on and exceeds 0.5 after sufficiently many confirming or positive instances $x_0, \ldots, x_n$. Let us call this condition CP1.
    - In order for any measure function to satisfy the weak condition of effective computability so as to qualify as an explicit inductive method, it must satisfy the following condition CP2: For any true computable hypothesis $M(x_{n+m+1})$ and for every n, 'it must be possible to find i.e. compute' an m such that if $M(x_{n+1}), \ldots, M(x_{n+m})$ hold, then $P(M(x_{n+m})) \mid M(x_0), \ldots, M(x_{n+m})$ is greater than 0.5.
- $c(h, e)$ is a statement that can be *analytically* proved in L.

It is important to note that even though e is based on the observed relative (statistical) frequency—i.e. probability$_2$—and indeed contains empirical content, $c(h, e)=r$ or the logical probability$_1$ statement does not contain e nor is it derived from e. What probability$_1$ statement contains is a *reference* to the evidence e and its empirical content.[5]

Hume's problem of induction challenges the claim that there can ever be $c(h, e)$ as knowledge of the future. Why? Because, irrespective of the specificities of the system of inductive logic and the theory of confirmation, the c-function is either analytic or ampliative. If it is analytic, the possibility that it can provide us with knowledge about the future is precluded. This is because e (i.e. the non-analytic source of

[5] 'Thus our empirical knowledge does not constitute a part of the content of the probability$_1$ statement (which would make this statement empirical) but rather of the sentence e which is dealt with in the probability$_1$ statement. Thus the latter, although referring to empirical knowledge, remains itself purely logical.' Ibid. 32.

knowledge about the past or the observed) together with an analytic statement—that is, $c(h, e)=r$—cannot provide us with knowledge or information of any kind about future or unobserved (h). If on the other hand such information is indeed possible, then, in so far as e-statements are about the past and h about the future, $c(h, e)=r$ cannot then be an analytic statement and cannot be analytically proved in L. In other words, the h-statement that refers to the future cannot be considered reasonable or rational if the only factual information it is based on is the past. Once again, this raises the question of why should we expect information about the future to continue the trends of the past in any sense.

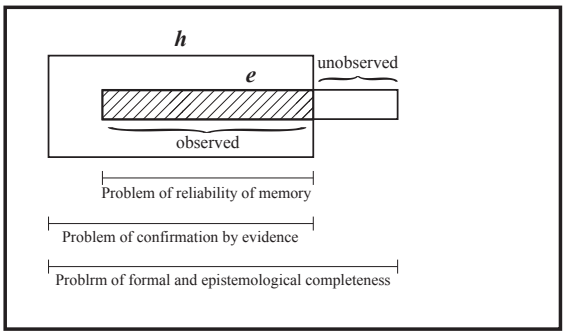
Furthermore, as Wolfgang Stegmüller has pointed out through a fictional conversation between Hume and Rudolf Carnap as a champion of inductivism, the choice of the particular c-function is quite arbitrary.[6] An inductivist like Carnap might say, 'It quite suffices for a rational procedure that there is in the long run a higher probability of success'.[7] That is, the inductive model of rationality—which for an inductivist is the *only* viable model of rationality—is based on the claim that, given a sufficiently long but finite time, it is reasonable or rational to believe that the inductive model (whether of the mind, intelligence, or a scientific theory) will be vindicated by a higher probability of success. Hume, however, would then challenge Carnap's claim by saying:

> Why should one be rational in your sense of the word 'rational'? The possible answer: 'Because it is just rational to be rational', would, of course, amount to a mere sophism; for in the latter part of the sentence you would be referring to your concept of rationality. The whole sentence then would be taken to mean that it is rational to accept your concept of rationality. But that is exactly the very thing that is in question. Finally, our common acquaintance will not be unaware of the fact that there are infinitely many different possibilities for defining the concept of confirmation and, hence, the concept of rationality.[8]

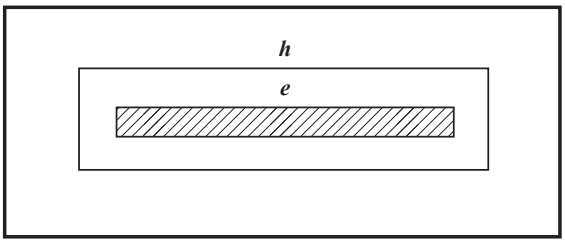
Now, if we take a system of inductive logic as a design for a leaning machine, i.e. a computational agent that can extrapolate empirical regularities from supplied e-statements, we can treat $c(h, e)=r$ as a principle of inductive logic upon which a Carnapian computational learning machine or an AGI can be designed. However, such an inductivist machine cannot in any sense be called rational or based on reasonable principles. The generalization of Hume's problem of induction would count as a clear refutation of such a purely inductive model of general intelligence or purported model of rationality or mind. Of course, a more astute inductivist might object that the c-functions 'correspond to "learning machines" of very low power'.[9] In other words, that we should give up the idea of modelling the mind or general intelligence on induction as the degree of confirmation and instead model it on the recursion-theoretical account of induction. As we shall see, this solution is not only plagued with the aforementioned quandaries of prediction, but also, in so far as the problem of induction is not limited to the predictive induction or c-functions, it is plagued with the problems of retrodiction and formalization which are in fact more serious to the extent that, in being less pronounced, they encroach upon more fundamental assumptions than the possibility of inductively inferring knowledge of the future from knowledge of the past.

---

[6] W. Stegmüller, *Collected Papers on Epistemology, Philosophy of Science and History of Philosophy* Vol. 2, 117–119.
[7] Ibid., 119.
[8] Ibid.
[9] H. Putnam, *Philosophical Papers* Vol. 1 (Cambridge, UK: Cambridge University Press, 1979), 297.



Induction



Deduction

Uniformity, Regularity, and Memory

Having examined Hume's problem of induction in its general form, we can now proceed to briefly look at Hume's own argument. According to Hume, inductive reasoning is grounded on the principle of the uniformity of nature as its premise—that is, unobserved events are similar to observed events, so that generalizations obtained from past observed occurrences can be applied to future unobserved occurrences. In Hume's words, 'that instances of which we have had no experience, must resemble those of which we have had experience, and that the course of nature continues always uniformly the same'.[10] But this principle itself is a conclusion reached by induction, and cannot be proved by the understanding or by deductive reasoning. It cannot be proved by deduction because anything that can be proved deductively is necessarily true. But the principle of uniformity is not necessarily true since the deductive framework admits, without logical contradiction, counterexamples for events which have not yet been experienced, in which a true antecedent (past patterns of events) is consistent with the denial of a consequent (future patterns of events not similar to the past).

Thus, if the principle of uniformity cannot be proved through deduction, and if therefore the validity of induction cannot be established deductively, then it must be proved by causal-probabilistic or inductive reasoning. Yet the validity of such reasoning is precisely what we sought to prove. To justify the principle of uniformity and induction by inductive reasoning is simply question-begging, i.e. a fallacy in which the conclusion is granted for the premises. Therefore, it follows that induction cannot be proved inductively either, because this would count as vicious circularity.

For this reason, Hume's problem of induction comes down to the idea that experience cannot provide the idea of an effect by way of understanding or reason (i.e. deductive and causal inferences), but only by way of the impression of its cause, which requires 'a certain association and relation of perceptions'.[11] Understanding cannot produce cause-effect relations or matters-of-fact since such relations are obtained via the inductive generalization of observations. Matters of fact rely on causal relations and causal relations can only be obtained inductively. But the validity of inductions themselves cannot be

[10] Hume, *A Treatise of Human Nature*, 390.
[11] Ibid.

corroborated deductively, nor can they be explained inductively. Therefore, what is problematic is not only the derivation of uncertain conclusions from premises by way of induction, but also, and more gravely, the very inductive principle by which such uncertain conclusions are reached.

Hume's problem of induction, accordingly, challenges the validity of our predictions—the validity of the connection between what has been observed in the past and what has not yet been observed. We cannot employ deductive reasoning to justify such a connection since there are no valid rules of deductive inference for predictive inferences. Hume's resolution to this predicament was that our observations of patterns of events—one kind of event following another kind of event—creates a habit of regularity in the mind. Our predictions are then reliable to the extent that they draw on reliable causal habits or regularities formed in the mind through the function of memory that allows us to correlate an impression with its reproduction and anticipation. For example, if we have the impression (or remember) that A resulted in B, and if we also witness at a later time and in another situation that 'an A of the same kind resulted in a B of the same kind', then we anticipate a nomological relation between A and B: B is the effect of A, as the cause of which we have an *impression*.

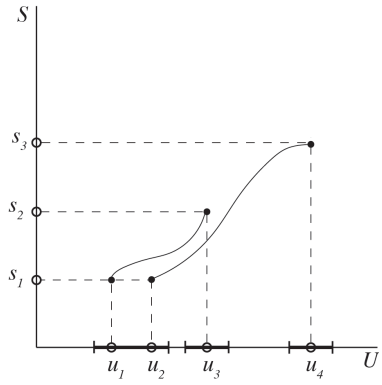
However, rather than settling the problem of induction, Hume's resolution—i.e. the reliability of habits of regularity accessible through memory—inadvertently reveals a more disquieting aspect of the problem: That the problem of induction challenges not only our predictive inductions about the future, but also our retrodictive or memory-driven knowledge of the past. In this sense, Hume's problem is as much about the derivation of the future-oriented *h*-statement from the *e*-statements as it is about the empirical reliability or factuality of the *e*-statements or information about the past. A proponent of the inductivist model of the mind, general intelligence, or scientific theories thinks that all he must do is to make sure that the evil Humean demon does not get through the door, not knowing that the demon is already in the basement. It is a dogmatic assumption to conclude that, so long as one can manage to take care of the reliability of predictive claims—either through a better theory of confirmation, a better Bayesian inference, or a more adequate formal-computational reformulation of inductive reasoning—one does not need to worry about the reliability of empirical reports referring to the past. In other words, a puritan inductivist who believes that general intelligence or the construction of theories can be sufficiently modelled on inductive inferences alone takes for granted the reliability of the information about the past, namely, *e*-statements. Yet for a so-called ideal inductive judge modelled on the alleged sufficiency of induction alone, the predicaments of induction hold as much for empirical data referring to the past as they do for assertions about the future

Such an ideal inductive judge would be particularly vulnerable to the problem of the unreliability of the knowledge of the past derived through memory or in Humean terms, the regularity-forming memory of impressions of causes. The strongest version of the problematic nature of the reliability hypothesis of memory has given by Russell's 'five minutes ago' paradox,

> There is no logical impossibility in the hypothesis that the world sprang into being five minutes ago, exactly as it then was, with a population that 'remembered' a wholly unreal past. There is no logically necessary connection between events at different times; therefore nothing that is happening now or will happen in the future can disprove the hypothesis that the world began five minutes ago. Hence the occurrences which are called knowledge of the past are logically

> independent of the past; they are wholly analysable into present contents, which might, theoretically, be just what they are even if no past had existed.[12]

As Meir Hemmo and Orly Shenker have elaborated, the 'five minutes ago' paradox can be conceptually reframed using the Boltzmannian notions of microstate (complexion) and macrostate (distribution of state):[13] Suppose at time t_1 an observer S remembers an event that took place at an earlier time t_0. Let us say, S observes at t_0 a partially deflated ball, and at the same time remembers that at an earlier time t_{-1} the ball was fully inflated. This memory is occasioned by the microstate s_1 of the nervous system of S at time t_0. The microstate memory s_1 is compatible with at least two microstates of the rest of the universe U, u_1 and u_2, at t_0. These two microstates are in the same *macrostate* (each macrostate is represented on the U-axis by a bracket). If we were to retrodict from the microstate s_1u_1 of $S\times U$ the microstates of S and U at t_1, then we would have been able to find them in the microstate s_2u_3 wherein the observer experiences a fully inflated ball and U is in a macrostate compatible with this experience. However, in the case of the microstate s_1u_2 at t_0 the scenario changes. For if we were to retrodict from it the microstates of S and U at t_1, we would have found them in the microstate s_3u_4—that is, where the observer experiences a fully deflated ball. The false retrodiction from s_1u_2 at t_0—as the consequence of many-to-one or possibly many-to-many correlations between the observer's memory states and the rest of the universe—is what is captured by the 'five minutes ago' paradox.



The gist of the 'five minutes ago' paradox consists of two parts: (1) memory-beliefs are constituted by what is happening now, not by the past time to which the said memory-beliefs appear to refer. In so far as everything that forms memory-beliefs is happening now, there is no *logical* or *a priori* necessity that what is being remembered (the reference of the memory-belief) should have actually occurred, or even that the past should have existed at all. (2) There is no logical reason to expect that memory states are in one-to-one correspondence with the rest of the universe. There can be both many-to-one and many-to-many correlations between memory states and external states of affairs. Therefore, what we remember as the impression of a cause, a past event, or an observation, may very well be a false memory—either a different memory or a memory of another impression of a cause. Accordingly, our knowledge of the past or of the impressions of causes can also be problematic at the level of logical plausibility and statistical improbability, which does not imply impossibility. Consequently, it is not only the justification of our

[12] B. Russell, *The Analysis of Mind* (London: George Allen & Unwin Ltd, 1921), 159–60.
[13] Hemmo and Shenker, *The Road to Maxwell's Demon*.

predictions regarding events not yet experienced or observed that faces difficulty, but also our memories of past impressions that have shaped our regularities and habits of mind.

The Dissolution of Hume's Problem and Its Rebirth

The Humean problem of induction undergoes a radical change first at the hands of Nelson Goodman in the context of the new riddle of induction, and subsequently those of Hilary Putnam in the context of Gödel's incompleteness theorems.[14]

Goodman observes that Hume's version of the problem of induction is, at its core, not about the justification of induction but rather about how evidence can inductively confirm or deductively corroborate law-like generalizations. Before moving forward, let us first formulate the Hempelian confirmation problem that motivates Goodman's problem of induction: A positive instance which describes the same set of observations can always generate conflicting or incompatible hypotheses. To overcome this problem, the positive instances must be combined with projectable hypotheses, i.e. hypotheses supported by positive instances and capable of forming law-like generalizations. But then a new riddle emerges: How can projectable hypotheses be distinguished from non-projectable hypotheses which are not confirmed by their positive instances? This new riddle of induction has come to be known as Goodman's grue paradox.

Let us imagine that before time *t* (e.g., a hypothetical future time such as 2050), we have observed many emeralds recovered from a local mine to be green, and no emerald to be of another colour. We thus have the following statements based on successful observations,

Emerald *a* is green, emerald *b* is green, etc.

Such evidence statements then afford generalizations of the kind supported by evidence,

All emeralds are green (not just in the local mine but everywhere).

Here the predicate *green* can be said to be a projectable predicate or a predicate that is confirmed by its instances (emerald *a*, emerald *b*, etc.), and can be used in law-like generalization for the purposes of prediction.

Now let us introduce the predicate *grue*. An emerald is grue provided it is green and observed *or* (disjunction) blue and unobserved before the year 2050, i.e. if and only if it is green before time *t* and blue thereafter. Here, the predicate grue does not imply that emeralds have changed their colour, nor does it suggest that, in order for emeralds to be grue, there must be confirmation or successful observation of its instances. We call such a predicate a non-projectable or grue-type predicate.

In the case of grue emeralds, we then have non-projectable generalizations,

Emerald *a* is grue, emerald *b* is grue, etc.

[14] See N. Goodman, 'The New Riddle of Induction' in *Fact, Fiction, and Forecast* (Cambridge, MA: Harvard University Press, 1979), 59–83; and H. Putnam, *Representation and Reality* (Cambridge, MA: MIT Press, 1991).

The generalizations 'All emeralds are green' and 'All emeralds are grue' are both confirmed by observations of *green* emeralds made before 2050. Before 2050, no grue emeralds can be observationally—i.e. inductively—distinguished from any green emeralds. Hence, the same observations support incompatible hypotheses about emeralds to be observed after *t*—that they are green and that they are blue. This is called Goodman's grue paradox. The paradox shows that there can be generalizations of appropriate form which, however, are not supported by their instances. So now the question is: What exactly is the difference between supposedly innocent generalizations such as 'All ravens are black' which are supported by their instances, and grue-type generalizations ('All ravens are blite', i.e. black before time *t* and white thereafter) which cannot be supported by their instances, but are nevertheless equally sound? Or, how can we differentiate between healthy law-like generalizations based on projectable predicates supported by positive instances and grue-like (or not law-like) generalizations based on non-projectable predicates not supported by positive instances? This is Goodman's new riddle of induction, which asks why it is that we assume that, after time *t*, we will find green emeralds but not grue emeralds, given that both green and grue-type inductions are true and false under the same set of conditions such that,

a) Based on the observations of many emeralds qua positive single instances, a miner using our common language will inductively reason that all emeralds are green. The miner forms the belief that all emeralds to be found in the mine or elsewhere are and will be green before and after time *t*.
b) Based on the same set of observations of green emeralds, a miner using the predicate 'grue' will inductively reason that all emeralds observed after time *t* will be blue, even though thus far only green emeralds have been observed.

Goodman's response to the paradox is as follows: the predicate green is not essentially simpler than the predicate grue since, if we had been brought up to use the predicate grue instead, it could very well be the case that grue would no longer count as nonsensical or as more complex than the predicate green by virtue of being green and blue. In that case, we could use predicates grue and bleen (i.e. blue before time *t*, or green subsequently) just as we now use the predicates green and blue. An objection can be made that, unlike green, grue is artificially defined disjunctively, and that therefore the natural predicate green should be preferred. Per Goodman's response, there is no need to think of grue and bleen-type predicates as disjunctive predicates. They can easily be thought as primitive predicates such that the so-called natural or simple predicate green can be defined as *grue if observed before time* t *or bleen thereafter*. Hence even the predicate green can be shown to be disjunctive. To this extent, the hypotheses we favour do not enjoy a special status because they are confirmed by their instances, but only because they are rooted in predicates that are entrenched in our languages, as in the case of green. If grue and bleen were entrenched, we would have favoured hypotheses of their kinds. Moreover, it should be noted that Goodman's argument applies not only to positive instances but also to negative ones (counterexamples), and as such also includes the deductivist theory of corroboration which is based on a reliable way of choosing a candidate element among rival hypotheses for the purpose of testing against counterexamples.[15]

[15] On this point see Lawrence Foster's response to Paul Feyerabend: L. Foster, 'Feyerabend's Solution of the Goodman Paradox', *British Journal for the Philosophy of Science* 20:3 (1969), 259–60.

If projectable and non-projectable predicates are equally valid, then what kinds of constraints can we impose on a system of inductive reasoning that will exclude grue-type non-law-like generalizations? Goodman's response is that that no purely formal or syntactical constraints can be sufficient to distinguish projectable from non-projectable predicates. In this sense, a machine equipped with a formal model of induction runs into the problem of distinguishing law-like from non-law-like generalizations. The only way to tell apart healthy green-like from grue-like properties is in terms of the history of past inductive inferences. The reason we use green and not grue is because we have used green in our past inductions. But equally, we could have been using the predicate grue rather than green so that we would now have justified reasons to use grue and not green.

In his radical version of the new problem of induction, utilizing Gödel's incompleteness theorems, Putnam adopted and refined this argument to show that inductive reasoning cannot be formalized—i.e., that there are no syntactical or formal features of a formalized inductive logic that can be used to make the aforementioned distinction. Putnam's use of incompleteness theorems, however, targets not just formal-computational accounts of induction but *any* computational description of the human mind or general intelligence. For this reason, I choose to limit Putnam's argument to a computational and *purely* inductive model of general intelligence. This is an agent or ideal inductive judge who is only in possession of an inductive model either constructed based on the (recursion-theoretic) computational theory of inductive learning or on Solomonoff's duality of regularity and compression (anything that can compress data is a type of regularity, and any regularity can compress the data).[16] The reason for this choice is that I would like to retain the main conclusions reached by Putnam's argument for at least the special case of an artificial agent restricted to one epistemic modality (i.e. computational induction), thereby avoiding the justified objections raised by, for example, Jeff Buechner against the overgeneralized scope of Putnam's argument.[17]

A formal system *F* is complete if, for every sentence of the language of that system, either the sentence or its negation can be proved (in the sense of derivability rather than proof in the absolute sense) within the system. *F* is consistent if no sentence can be found such that both the sentence and its negation are provable within the system. According to the first incompleteness theorem, any consistent *F* that contains a small fragment of arithmetic is incomplete—that is, there are sentences (Gödel-sentences) which cannot be proved or disproved in *F*. According to the second incompleteness theorem, for a consistent system *F* that allows a certain amount of elementary arithmetic (but more than the first theorem) to be carried out within it, the consistency of *F* cannot be proved in *F*. Then *F* can be said to be Gödel-susceptible.

An artificial general intelligence, or even the human mind modelled purely on a computational inductive model, is always Gödel-susceptible. Put differently, such a computational agent or purely inductive mind can never know the truth (in the formal derivability sense) of its Gödel-sentences in the epistemic modality under which it enquires into the world. This computational inductive agent can never know the model it inhabits. It cannot know whether the model it inhabits is standard, in which case its Gödel-sentence is true, or is non-standard, in which case its Gödel-sentence is false. For this agent, knowing the

[16] R. Solomonoff, 'A Formal Theory of Inductive Inference parts 1 and 2', *Information and Control* 7:1 (1964), 224–54.
[17] J. Buechner, *Gödel, Putnam, and Functionalism* (Cambridge, MA: MIT Press, 2008).

model it occupies and under which it conducts inquiry into the world is not just underdetermined. It is rather completely indeterminate in so far as, within such a system, the only possible information that can lead to the determination of the truth of the model's Gödel-sentences can only be obtained by finitary derivation. And within such an agent's model, finitary derivation cannot establish the truth of the Gödel-sentences unless the agent's inductive model is updated to a new computational system—in which case the question of the model the agent occupies and its indeterminacy will be simply carried over to the new system.

In its general form, Putnam's argument in *Representation and Reality* rejects the possibility that inductive inferences can be computationally formalized. This is because either Bayesian reasoning (i.e. prior probability metrics) cannot be arithmetically formalized, or projectable predicates cannot be formalized. A purely inductive computational model of the mind or general intelligence is Gödel-susceptible, which means that the description of such a model is indeterminate and hence arbitrary. Whereas Goodman's argument challenges the distinction between rival hypotheses or law-like and non-law-like generalizations based on formal-syntactic constraints, Putnam extends Goodman's argument to the description of the mind itself: In so far as inductive inferences cannot be arithmetically formalized due to Gödel-susceptibility, no computational model of a purely inductive mind or an inductive model of general intelligence 'can prove it is correct or prove its Gödel sentences in the characteristic epistemic modality of the proof procedure of the formal system formalizing those methods'.[18]

This problem, however, could have been avoided had the model of general intelligence accommodated epistemic multimodality (inductive, deductive, and abductive methods, syntactic complexity as well as semantic complexity). But the inductivist proponent of artificial general intelligence is too greedy to settle for a complex set of issues which require that we expand the model of mind and rationality. Not only does he want to claim that the problem of constructing AGI is the problem of finding the best model of induction (based on the assumption of the sufficiency of induction for realizing the diverse qualitative abilities which characterize general intelligence); he also seeks to lay out this omnipotent inductive model in purely syntactic-axiomatic terms without resorting to any semantic criterion of cognition (i.e. conceptual rationality). Instead, what the inductivist gets is the worst of all possible words. He ends up with both the reliability quandaries harboured by the problems of induction, old and new, *and* the problems of the computational formalization of induction.

In addition, Putnam's argument as formulated in his essay '"Degree of Confirmation" and Inductive Logic' can be understood as a general argument against the possibility of the construction of a universal learning machine.[19] Such a machine is essentially a measure function *P* that is effectively computable and which, given sufficient time, would be able to detect any pattern that is *effectively* computable.[20] Since the

[18] Ibid., 73.
[19] H. Putnam, '"Degree of Confirmation" and Inductive Logic', in *The Philosophy of Rudolf Carnap* (La Salle, IL: Open Court, 1963), 761–83.
[20] 'When considering the kinds of problems dealt with in any branch of logic, deductive or inductive, one distinction is of fundamental importance. For some problems there is an effective procedure of solution, but for others there can be no such procedure. A procedure is called *effective* if it is based on rules which determine uniquely each step of the procedure and if in every case of application the procedure leads to the solution in a finite number of steps. A *procedure of decision* ("Entscheidungsverfahren") for a class of sentences is an effective procedure either, in

ideal of any inductive system is to satisfy the previously mentioned conditions CP1 and CP2, and furthermore, since a universal learning machine should be effectively computable, such a machine must satisfy two additional general conditions which correspond respectively to CP1 and CP2: For an inductive method *D*,

CP1′: *D* converges on any true computable hypothesis *h*.

CP2′: *D* is computable.

Putnam has demonstrated that the effectively computable *P* (i.e. the universal learning machine) is diagonalizable such that CP1′ and CP2′ violate one another. Stated differently, no inductive method can simultaneously fulfil the condition of being able to detect every true effective computable pattern *and* the condition of the effective computability of the method itself, and so qualify as a universal learning machine: For a candidate computable measure function *P*, a computable hypothesis *h* can be constructed in such a way that *P* fails to converge on *h*:

1) Let **C** be an infinite class of integers n_1, n_2, n_3, … having the following property: the degree of confirmation (*r*) of $M(x_{n_1})$ exceeds 0.5 if all preceding individuals are *M*. For $M(x_{n_2})$, *r* exceeds 0.5 if all preceding individuals after x_{n_1} are *M*. Or generally, the degree of confirmation $M(x_{n_j})$ is greater than 0.5 if all the preceding individuals *after* $x_{n_{j-1}}$ are *M*.
2) The predicate *M* belongs to the arithmetical hierarchy, i.e. it can be defined in terms of polynomials and quantifiers.
3) **C** is a recursive class and as such the extension of the arithmetic predicate *M.* It is recursive in the sense that there exists a mechanizable procedure to determine whether an integer can be found in this class. **C** is the direct result of the *effective* (computability) interpretation of CP2 i.e. 'it must be possible to find an *m*'.
4) Beginning with the first individual x_0, compute $P(M(x_0))$ and let $h(x_0)$ be $\neg M(x_0)$ iff $P(M(x_0))>0.5$.
5) For every new individual x_{n+1}, continue the previous procedure: compute $P(M(x_{n+1}) \mid h(x_0), \ldots, h(x_{n+1}))$ and let $h(x_{n+1})$ be $\neg M(x_{n+1})$ iff the probability of $P(M(x_{n+1}))$ exceeds 0.5.
6) Even though *h* is computable, nevertheless because of the construction of instance confirmation given by the measure function *P*, it never remains above or exceeds 0.5.
7) Thus if an inductive method *D* is to satisfy CP1 and CP2, then it cannot be reconstructed as a measure function. Or alternatively, if *D* is supposed to converge to any true computable hypothesis (CP1′) and to also be computable itself (CP2′), then it would be impossible to reconstruct it as a measure function or a universal learning machine with the aforementioned characteristics.

---

semantics, for determining for any sentence of that class whether it is true or not (the procedure is usually applied to L-determinate sentences and hence the question is whether the sentence is L-true or L-false), or, in syntax, for determining for any sentence of that class whether it is provable in a given calculus (cf. Hilbert and Bernays [Grundlagen], Vol. II, § 3). A concept is called *effective* or *definite* if there is a procedure of decision for any given case of its application (Carnap [Syntax] § 15; [Formalization] § 29). An effective arithmetical function is also called *computable* (A. M. Turing, *Proc. London Math. Soc*, Vol. 42 [1937]).' Carnap, *Logical Foundations of Probability*, 193.

Moreover, Tom Sterkenburg has shown that even a Solomonoff optimal learning machine falls under Putnam's diagonal argument.[21] An optimal learning machine can be defined as a pool of competing learning machines or inductive experts with no assumption about the origin of data and for which the criterion of reliability (i.e. guaranteed convergence on the true hypothesis) has been replaced with the more moderate criterion of optimality (i.e. it is guaranteed to converge on the true hypothesis if *any* learning machine does).

Bluffing Your Way Through Simplicity

Faced with the different ramifications of the problem of induction, at this point, an inductivist invokes the magic word 'simplicity', or some variation of it: either elegance, which is concerned with the formulation of a hypothesis, or parsimony, which deals with the entities postulated by a hypothesis. In either case, simplicity is taken as a magic remedy against the plights of induction. As long as there is the principle of simplicity, there is a way out of the predicaments of induction (e.g., differentiating projectable from non-projectable predicates). For an inductivist proponent of theory-formation and theory-comparison, simplicity is what enables us to separate good hypotheses from bad ones, or distinguish true theories when dealing with competing, incompatible, or rival theories. At first glance, this claim regarding the significance of the principle of simplicity does indeed appear sound, for the principle of simplicity is a tool that imposes helpful and necessary *pragmatic* constraints on our epistemic inquiries. But the inductivist is not interested in simplicity as a pragmatic tool whose application requires access to semantic information about the context of its application. When the inductivist speaks of simplicity, he does not refer to simplicity or Occam's razor as a contextual pragmatic tool, but to simplicity as an objective epistemic principle.

When comparing incompatible or rival theories T_1 and T_2, solely based on a general and context-independent objective notion of epistemic simplicity, one of the theories (the simpler one) can be characterized as true. But among two incompatible and rival theories where one of the theories is actually false, the appeal to the principle of simplicity cannot be indiscriminately made, since in one or more contexts, the false theory may be simpler than the true one, and may accommodate well-formulated questions which are ill-posed in the other theory.[22]

A more up-to-date inductivist can claim that such an idealized objectivist notion of epistemic simplicity does indeed exist: the formal-computational account of Occam's razor, where simplicity is equated with compression, and compression is couched in terms of the effectiveness of Solomonoff's prediction. It is precisely this absolute and objective notion of epistemic simplicity—understood in terms of the formal

[21] T. F. Sterkenburg, 'Putnam's Diagonal Argument and the Impossibility of a Universal Learning Machine' (2017), <http://philsci-archive.pitt.edu/12733/>.

[22] 'Even in the case of Ptolemy's and Copernicus's theories, there are well-posed questions in the one theory, which are ill-posed in the other by respectively resting on presuppositions that are declared to be false in the other: e.g., Ptolemy can ask how long it takes for the sun to go around the earth, but Copernicus cannot; and Copernicus can ask how long it takes the earth to go around the sun, but Ptolemy cannot.' Grünbaum, *Is Simplicity Evidence of Truth?*, 271.

duality of regularity and compression—that lies at the heart of inductivist trends in artificial general intelligence.

According to algorithmic information theory, a data object such as the specification of a hypothesis is simpler when it is more compressible, i.e. when it can be captured by a shorter description. This idea can be made formally precise using the theory of computability, resulting in Kolmogorov's measure of complexity for a data object as the length of its shortest description or the program that generates it. The length of the program is essentially the number of bits it contains. Solomonoff's induction (or method of prediction) uses this complexity measure to give higher probability to simpler extrapolations of past data: For a monotone machine that has been repeatedly fed random bits through the tossing of a fair coin where the probability of either 0 or 1 is 0.5,[23] the output sequence σ of any length receives greater algorithmic probability if it has shorter descriptions of the input sequence ρ that has been given to the machine in just that manner. The probability that that we end up in this manner feeding the machine a sequence that starts with ρ entirely depends on the length $|\rho|$.[24] Once the machine processes ρ, it outputs a sequence. For an output σ of any length that starts with this sequence, ρ can be said to have been a guide or program for the machine to produce the sequence σ that enjoys a greater algorithmic probability. In other words, ρ is effectively the machine description of σ.

It has been formally proved by Solomonoff that the aforementioned method of prediction is reliable in the sense that it leads to the truth. Essentially, Solomonoff's induction is based on the definition of a type of predictor with a preference for simplicity, along with a proof that a predictor of this type is reliable in that it is guaranteed to converge on the truth. Accordingly, Solomonoff's induction is a formal argument that justifies Occam's razor. In Solomonoff's theory, simplicity is characterized in terms of the weighted sum of program lengths, which depends on the choice of the monotone universal Turing machine. The choice of the machine which determines the length of the program or description corresponds to the argument from parsimony, while the length of the program itself corresponds to the argument from the perspective of elegance.

[23] A monotone machine can be characterized as a true on-line machine which, at the same time as processing a stream of input bits, can produce a potentially infinite stream of output bits. Since in Solomonoff's system the choice of machine is restricted to universal Turing machines, and furthermore, since in the classical Church-Turing paradigm of computability, the machine cannot accept new input bits during the operation, this criterion is satisfied by the addition of a specialized oracle. It is called monotone since the monotonicity constraint permits to directly infer from the machine U a specific probabilistic source. A function $M(y,t)$ can be called monotonic when for a later time t' of the time t and extensions of y' of the descriptions y, we can derive from M a data object which is the extension of $M(y,t)$. Essentially, the monotonic function is a transformation such that it returns for each finite binary string, the probability that the string is generated by the machine U, once U is fed repeatedly a stream of uniformly random input produced by bets that the probability of either 0 or 1 is 0.5. This allows us to define monotone descriptional complexity of a data object in terms of U with almost the shortest description and without reference to the hidden information in the length of either ρ or σ.

[24] Solomonoff has demonstrated that this probability is $2^{-|\rho|}$.

However, a closer examination of Solomonoff's Carnap-influenced formal theory of induction reveals that this objective notion of simplicity is circular.[25] The argument, as advanced by Solomonoff and further detailed by Vitányi and Hutter, can be briefly formulated as follows:[26]

Given two classes of predictors **Q** and **R** which respectively specify the class of algorithmic probability predictors via all universal monotone Turing machines and the class of effective mixture predictors via all effective priors which embody inductive assumptions:

1. Predictors in class **Q** have distinctive *simplicity qua compressibility bias*. Or equally, predictors in the class **R** operate under the inductive assumption of *effectiveness* in the context of sequential prediction.
2. Predictors in **Q** are reliable in every case. Or, predictors in **R** are consistent.
3. Therefore, predictors with a simplicity qua compressibility bias are reliable in essentially every case. Or, predictors operating under the inductive assumption of effectiveness are consistent.

However, by making explicit the property of consistency in the second step of the argument (i.e. the consistency property of Bayesian predictors as applied to the class of effective predictors),[27] it can be shown that the argument essentially runs as follows:

1. Predictors in **R** operate under the assumption of effectiveness.
2. Predictors in **R** are reliable under the assumption of effectiveness.

In other words, a vicious circularity in the definition of simplicity qua compressibility bias emerges: predictors operating under the assumption of effectiveness are reliable under the assumption of effectiveness. The meaningful application of the formal notion of simplicity-as-compressibility to infinite data streams is ultimately predicated on the inductive assumption of effectiveness. But this assumption only offers a weak notion of simplicity in so far as any inductive assumption can be taken as a specification of simplicity, which then requires a new inductive argument to specify which assumption of effectiveness is preferable or which notion of simplicity is more strongly objective. Adding such an argument would again require further inductive arguments to establish the ideal effectiveness as the simplicity stipulation. Without these additional arguments, the notion of simplicity ends up being

[25] Solomonoff has explicitly referred to Carnap's claim that predictive induction is the most powerful and general form of induction as well as to his theory of inductive logic as the degree of confirmation, see Solomonoff, *A Formal Theory of Inductive Inference* parts 1 and 2 (1964).

[26] See R. Solomonoff, 'Complexity-based Induction Systems: Comparisons and Convergence Theorems', *IEEE Transactions on Information Theory* (1978), 422–32; and for the elaboration of Solomonoff's system in connection with Occam's razor and artificial intelligence, see M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications* (Dordrecht: Springer, 2008); and M. Hutter, *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability* (Dordrecht: Springer, 2004).

[27] Bayesian consistency means that posterior distribution concentrates on the true model—that is, for every measurable set of hypotheses, the posterior distribution goes to 1 if it contains truth and 0 if it does not: Thus a prior p_0 on the parameter space Θ is consistent at $\theta \in \Theta$ if according to the chance hypothesis θ, the chance of a sequence of outcomes arising that together with p_0 would generate a sequence $(p_1, p_2, \ldots)$ of posteriors that did not concentrate in the neighbourhood of θ is zero. A consistent prior is 'essentially guaranteed to lead to the truth, in the sense that no matter which chance hypothesis is true, any nonpathological stream of data generated by that hypothesis would lead an agent with that prior to pile up more and more credence on smaller and smaller neighborhoods of the true hypothesis'. See G. Belot, 'Bayesian Orgulity', *Philosophy of Science* 80:4 (2013), 490.

viciously circular, and its connection to reliability cannot be established. But with the addition of an inductive argument that specifies effectiveness, a potentially infinite series of arguments will be required. Thus, ironically, the formal definition of simplicity requires a program that can no longer be identified as simple (elegant or parsimonious) in any sense. Moreover, pace Vitányi and Hutter, there is nothing in the definition of Solomonoff's universal induction nor in the definition of any inductive-predictive method that warrants our interpreting effectiveness as a metaphysical constraint on the world rather than as an epistemic constraint (what is calculable?).

Foregoing the metaphysical conceptions of simplicity and effectiveness would require us to abandon the more ambitious claims regarding the sufficiency of inductive-predictive methods, the possibility of a universal learning machine, and the inductive nature of general intelligence in favour of far more modest pragmatic-epistemic claims—which may indeed be significant in the context of our own methods of inquiry and only in conjunction with other epistemic modalities.

This is the predicament of simplicity-qua-compressibility as an objective epistemic notion: its criteria are underdetermined if not wholly indeterminate, and its definition is circular. In idealizing or overgeneralizing the notion of simplicity in terms of compressibility and identifying general intelligence with compression, the inductivist robs himself of exactly the semantic-conceptual resources which might serve not only to determine the criteria for the application of the principle of simplicity, but also to define general intelligence in terms not of compression but of the selective application of compression. Once again, the inductivist proponent of general intelligence finds himself confronted with old and new predicaments, albeit this time within the context of the formal-computational models of induction.

Ultimately, the pessimism weighing against the possibility of artificial general intelligence in philosophy of mind and the over-optimism of proponents of the inductivist models of general intelligence, in a sense, originate from their choice of model of rationality. They choose either a thick concept of rationality that does not admit of the artificial realization of mind, or a notion of rationality so thin that not only is artificial general intelligence inevitable, it inevitably takes the shape of an omnipotent omniscient inductive superintelligence. The popularity of these factions is not so much a matter of theoretical sophistication or technological achievement as the result of the dominance of such impoverished concepts of rationality. In their pessimism and over-optimism, they are both beholden to paradigms of justification derived from narrow conception of rationality and mind. To truly begin to examine the prospects of the artificial realization of general intelligence, one ought to start from the position of systematic scepticism with regard to any paradigm of rationality built on a method of theoretical inquiry claiming to be a sufficient replacement for every other method (e.g., over-confident—as in contrast to modest—Bayesian or statistical methods) and to any inflationary model of mind that collapses the qualitative distinction between different faculties and the requirements for their realization.