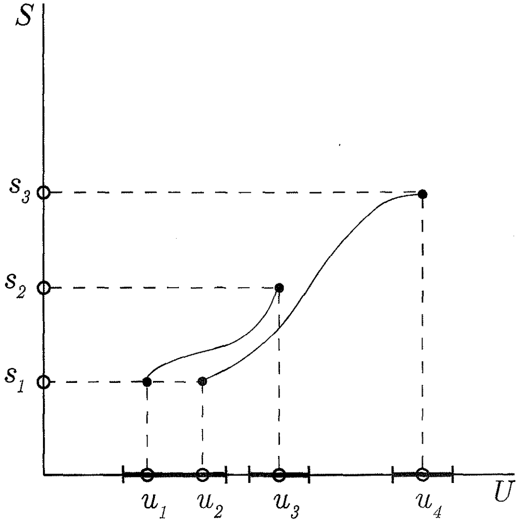
memory is occasioned by the microstate s_1 of the nervous system of S at time t_0. The microstate memory s_1 is compatible with at least two microstates of the rest of the universe U, u_1 and u_2, at t_0. These two microstates are in the same *macrostate* (each macrostate is represented on the U-axis 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.



Hemmo and Shenker's macrostate-microstate view of 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 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.[374]

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

[374] See Goodman, 'The New Riddle of Induction', in *Fact, Fiction, and Forecast*, 59–83; and Putnam, *Representation and Reality* (Cambridge, MA: MIT Press, 1988).

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 nonprojectable 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 generalizations 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 or that grue-type emeralds are date-dependent.[375] We call such a predicate a nonprojectable (i.e., an unnatural projection) or grue-type predicate.

[375] For a discussion on the common confusions around the grue paradox see D.M. Armstrong, *What is a Law of Nature?* (Cambridge: Cambridge University Press. 1983).

In the case of grue emeralds, we then have nonprojectable generalizations,

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

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 nonprojectable predicates not supported by positive instances? Another way to formulate the paradox is by way of David Armstrong's argument:

> The Regularity theorist's problem is to justify an inference from, say, observed emeralds to unobserved emeralds, while denying that there is any intermediate law. For his concept of law is that it is simply the greenness of observed emeralds plus the greenness of unobserved emeralds. How then can it help him to add the unobserved class to the observed class, and then argue from the observed class to this total class using the mathematics of probability? His problem is to get from the observed class to a completely disjoint class. No logical probability can help here. (And if it did, it would equally help with unnatural as well as natural predicates.)[376]

[376] Ibid., 58.

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,

- 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*.

- 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.[377]

If projectable and nonprojectable 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 no purely formal or syntactical constraints can be sufficient to distinguish projectable from nonprojectable 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

[377] 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.

any regularity can compress the data).[378] 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.[379]

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 inquires 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 nonstandard, in which case its Gödel-sentence is false. For this agent, knowing the 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

[378] R. Solomonoff, 'A Formal Theory of Inductive Inference parts 1 and 2', *Information and Control* 7:1 (1964), 224–54.

[379] J. Buechner, *Gödel, Putnam, and Functionalism* (Cambridge, MA: MIT Press, 2008).

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 owing 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'.[380]

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). But what the inductivist gets

[380] Ibid., 73.

is the worst of all possible worlds. 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.[381] 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.[382] Since the 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.

[381] Putnam, '"Degree of Confirmation" and Inductive Logic', 761–83.

[382] '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 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 der Mathematik* [2 vols. Berlin: Springer, 1979/1982], vol. 2, § 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.

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_2} 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 $P(\mathrm{M}(x_{n+1}) \mid h(x_0), \ldots, h(x_{n+1}))$ 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.

If we reframe Putnam's diagonal argument in terms of the familiar Church-Turing paradigm of computation as a special computer, or alternatively as an inductivist expert or scientist, we can say that this expert is supposed to be capable of guessing or making informed bets about the next digit in a sequence of 0s and 1s. For example, given a sequence $\{0,0,0,0,0\}$, a non-expert person might say that the next bit is also 0. But for an inductive learning machine or expert, the extrapolated sequence might be something like $\{0,0,0,0,0,1,1,0\}$. The inductive learning machine is concerned with a general inductive rule that yields an output about the guess or bet regarding the next bit given the finite sequence that has been observed so far. This expert rule is called the recursive predictor or *extrapolator* $\mathcal{T}$.

$$\text{Evidence} \rightarrow \boxed{\text{Extrapolator } \mathcal{T}} \rightarrow \text{Prediction}$$

Over time, the predictor sees more bits of the infinite data stream ε. In this scenario, at each stage or time n, when ε_n is obtained, the initial segment of the data stream $\{\varepsilon_1, \varepsilon_2, \ldots\varepsilon_n\}$ is available for examination. If we take H as the set of hypotheses of some interest or the set of all possible data streams which may arise, there is an actual data stream $\varepsilon \in H$ which can be extrapolated as the predictor inspects increasingly larger initial segments of ε and outputs increasing sequences of guesses or bets about the bitstring that might arise. $\mathcal{T}$ can be said to be reliably extrapolating H in the limit if, for every individual ε in H, there is a state or time n such that, for each later time m, the extrapolator's prediction is guaranteed to converge on correct prediction: $\Gamma(f_\varepsilon[n]) = f_{\varepsilon_{n+1}}$ where f_ε is a recursive zero-one valued function, $f_\varepsilon[n]$ is the initial segment of f_ε of the length n and $f_{\varepsilon_{n+1}}$ is the next value.

Putnam's diagonal argument shows that no recursive $\mathcal{T}$ can extrapolate every recursive function $f_{\varepsilon}[n]$ if it is to satisfy both CP1′ and CP2′. Furthermore, at no time or stage can the data imply the correctness of the hypothesis. In the vein of Kevin Kelly and others' elaboration of Putnam's diagonal argument,[383] let us assume that there is an effective procedure or computable function $f(e, x)$ that allows us to calculate in advance how many particular observations or bits x must be successively given to $\mathcal{T}$ in order for $\mathcal{T}$ to be able to predict the next observation x for each finite data segment e. $\mathcal{T}$ can be said to be 'recursively gullible' if there exists precisely such a computable function.[384] Or, more simply, $\mathcal{T}$ is recursively gullible when, regardless of what it has inspected so far, when fed observation x frequently, it will begin to predict that x will arise next. Using Putnam's diagonal argument, it can be proved that a recursively gullible $\mathcal{T}$ does not extrapolate H_{Rec} when H is a set of all data segments generated by a computer (i.e., H_{Rec} is a zero-one valued recursive set). In this demonstration, at each stage, we check $\mathcal{T}$'s prediction at the end of the previous segment of x. Next we pick a datum, say y such that $y \neq x$ (cf. first Green(x_i) then choosing Not-Green(x_i)). At this point, f can be used to calculate how many instances of y need to be added to the current data segment e in order to enable $\mathcal{T}$ to predict y. Once f has calculated how many, we add that many instances of ys to e so that $\mathcal{T}$ makes a mistake once it has read the last instance of y just added. $\mathcal{T}$ makes infinitely many mistakes. Yet ε is effective in so far as effectiveness has been defined recursively by way of the recursive function f. Therefore, if $\varepsilon \in H_{Rec}$ then $\mathcal{T}$ does not extrapolate H_{Rec}.[385]

Moreover, Tom Sterkenburg has painstakingly shown that even a Solomonoff optimal learning machine falls under Putnam's diagonal argument.[386] An optimal learning machine can be defined as a pool of competing learning

[383] K. Kelly, C. Juhl, et al., 'Reliability, Realism and Relativism', in P. Clark (ed.), *Reading Putnam* (London: Blackwell, 1994), 98–161.

[384] Ibid.

[385] See Putnam, '"Degree of Confirmation" and Inductive Logic', 769.

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

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).

It might be objected that Putnam's assault on the idea of a universal learning machine is not exclusive to the computational account of predictive induction, but can equally be applied to our inductive methods. That is to say, we should extend the conclusions reached by the diagonal argument to the human mind. It then follows that the quandaries that arise from the formalization of predictive induction not only undermine the idea of a universal computational learning machine, but also challenge the human mind. Consequently, the sceptical claims made against the possibility of constructing genuine learning computers seem to be prejudiced in that they limit the implications of the quandaries of induction to computers while letting the human mind off the hook, whether in the name of human exceptionalism or an implicit metaphysical concept of the human mind. If inductive inferences are indispensable tools in forming knowledge-claims, then the problematization of inductive methods and its consequences cannot be selectively used to distinguish human knowledge from a learning machine. If anything, such a problematization dissolves the distinction between the two, since both have to face the same set of challenges.

In response to such an objection, it should be pointed out that the human knowledge or human mind is distinct from a universal learning machine because human knowledge formation is not a matter of either/or. Unlike a universal learning machine, human knowledge is not based either on a purely inductive method or a purely deductive method. It is based on *both*, as inseparably connected. For us, inductive inferences are caught up in a complex web of diverse epistemic modalities, semantic complexity, contextual information, and so on. To put it differently, our inductive methods are impure in the sense that they do not operate by themselves but always in conjunction and entanglement with other methods and modes of epistemic inquiry. This of course raises the question of what the catalogue of such methods and epistemic modalities might be—a catalogue that can at once list, distinguish, and rank epistemic methods and mental faculties. This is indeed an open question that only philosophies of epistemology and mind under the aegis of cognitive sciences including logic and theoretical computer sciences can answer. The first step toward compiling such a catalogue, though, is to abandon—on a methodological level—any inflationary method or model identified as the most decisive or sufficient, in favour of the toy model approach to human rationality and mind that has been introduced in this book. Hume's challenge properly understood is not a challenge to knowledge per se, but to paradigms of knowledge built on inflationary models of epistemic inquiry, constructed on the premise of a single method deemed sufficient to carry out the task of other methods or different faculties of mind.

What distinguishes the human mind or knowledge from a universal learning machine of the kind described above is precisely ordinary (i.e., human rationality, which is reliant on both epistemic multimodality and a complex of qualitatively distinct mental faculties). Here, however, the term *ordinary* ought to be handled with care, distinguished from common sense as identified purely with the manifest image, and defined precisely and scientifically in terms of a multilevel web of inter-related methods and faculties. Most importantly, ordinary rationality should not be equated with the vague notion of informal rationality. As Kelly and others have expressed, the appeal to ordinary rationality as informal rationality is more akin to a conversation between a cognitive scientist and a naive philosopher. Imagine the scientist and the philosopher standing next to a computer. Every time the scientist makes a claim about how the computer functions, what it can possibly do, or how it might shed some light on our own rationality, the philosopher says, 'Switch this stupid thing off. I have informal rationality!'[387] If ordinary is taken to mean informal as common sense qua a purely manifest image of our rationality, then there is indeed a question as to what exactly safeguards common-sense rationality from the aforementioned quandaries. If the answer is that evolution has provided

[387] Kelly et al., 'Reliability, Realism and Relativism'.

us with epistemically reliable heuristic-inductive tools or some innate creative intuition then, as argued in the excursus on time, the collapse into epistemic naivety is inevitable.

Similarly, if informal means not formalizable, then the question would be how we can claim that ordinary human rationality is foreclosed to formalization without providing either an inflated picture of human rationality or an impoverished account of formalization. The equation of ordinary rationality with informal rationality in the latter sense is, alas, among the weakest of Putnam's arguments, and this for a number of reasons: (1) As Buechner has argued, not every model of formalization means arithmetical formalization; (2) if ordinary rationality in its entirety cannot be modelled on the classical Church-Turing paradigm of computability, this does not mean that ordinary rationality cannot be modelled computationally, for, as has been argued in the previous chapters, the Church-Turing paradigm is only a special case of a more general concept of computation; (3) In light of (1) and (2), the claim that human rationality is not formalizable requires unjustifiably strong claims with respect to either the nature of human rationality or the scope of formalization. In either case, the price to be paid for this claim to be considered justified or rational in a non-inflationary sense is too high.

Ruling out the more familiar senses of the ordinary, then, what does ordinary mean when we associate it with human rationality? The answer is that ordinary means the complete demystification of rationality as something extra-ordinary. But more importantly, it implies that rationality, as concerned with knowledge-formation and knowledge-claims, does not rely on the power of a single method or model that can be said to be sufficient to satisfy all aspects and desiderata of the rational or the epistemic *order*. That is to say, ordinary or human rationality is essentially multimodal. By multimodality I mean what Lorenzo Magnani calls the hybridity and distributedness of methods and modes of gaining traction upon the objective world, and what Yehoshua Bar-Hillel identifies as a multidimensional perspective as opposed to a one-dimensional perspective (i.e., a line of thinking that considers a single point of view to be sufficient or decisive for knowledge-formation or the theoretical assessment of hypotheses).[388] Epistemic multimodality or the multidimensional perspective, however, should not be interpreted as a pure liberal pluralism of models and methods. The implied plurality is a constrained one, in that it admits of a ranking or prioritization of methods, modes of inquiry, and mental faculties which are in complex interplay with one another.

BLUFFING YOUR WAY THROUGH SIMPLICITY

Faced with the various ramifications of the problem of induction, at this point an inductivist will invoke 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 magical remedy for 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 nonprojectable predicates). For an inductivist proponent of theory-formation and theory-comparison, simplicity is what enables us to separate good hypotheses from bad ones, or to 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 upon 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 refers not to simplicity or to 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 on the sole basis of a general and context-independent objective notion of epistemic

[388] See Magnani, *Abductive Cognition*; and Y. Bar-Hillel, '"Comments on the Degree of Confirmation" by Professor K.R. Popper', *British Journal for the Philosophy of Science* Vol. 6 (1955), 155–7.

simplicity, one of the theories (the simpler one) can be characterized as true. But when faced with two incompatible and rival theories one of which is actually false, the appeal to the principle of simplicity cannot be made indiscriminately, 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.[389]

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 prediction. It is precisely this absolute and objective notion of epistemic simplicity—understood in terms of the formal 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 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,[390] the output sequence σ of any length

[389] '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.

[390] 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

receives greater algorithmic probability if it has shorter descriptions of the input sequence ρ given to the machine in that manner. The probability that we end up in this manner feeding the machine a sequence that starts with ρ entirely depends on the length $|\rho|$.[391] 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.

To put it more formally, the Kolmogorov complexity of an infinite string $\sigma = \sigma_1, \sigma_2, \ldots, \sigma_n$ where σ_1 is 0 or 1 can be generally formulated as:

$$K(\sigma) := min\{|\rho|, \rho \in \{0,1\}^* : U(\rho) = \sigma\}$$

where $\{0,1\}^*$ is the set of all binary strings, U is the universal Turing Machine or a formal descriptive language and ρ a variable ranging over all programs such that when U is applied to them, they produce σ as the initial segment of the output.

Now the question is as follows: If we give the machine random bits, what would be the probability of machine U returning the sequence σ or more precisely, the probability that we arrive at a U-description of σ?

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.

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

Solomonoff demonstrates that if we seek to generate a prior probability distribution over $\{0,1\}^*$, this task can be accomplished by resorting to Occam's razor where higher probability is assigned to simpler or shorter strings. The so-called Solomonoff prior (M) answers the above question by finding this probability via Occam's razor and in relation to Kolmogorov complexity. The Solomonoff prior or algorithmic probability source can be formulated as:

$$M(\sigma) := \sum_{\rho \in D_{U,\sigma}} 2^{-|\rho|}$$

where D_U, σ is the minimal U-description of σ. Or, more simply, the Solomonoff prior is the sum over the set of all programs which compute σ. If $|\rho|$ is long, $2^{-|\rho|}$ will be short, and therefore contributes with a higher degree of probability to $M(v)$. The inference of the Solomonoff prior is called Solomonoff universal induction.

Solomonoff induction shows that, when seeking the computer program or description that underlies a set of observed data (*e*-statements) by way of Bayesian inference over all programs, one is guaranteed to find the correct answer if Solomonoff prior is used. In tandem with the theories of Andrey Kolmogorov and Gregory Chaitin, Solomonoff's theory of universal induction assumes that the best theory is the one that is simpler, with simplicity defined formally as compressibility. Therefore, the best theory is the one that best compresses the observation data. The question of compressibility is at its core the question of finding patterns that are effective. Solomonoff induction proposes, then, that the best method of prediction of the unobserved data is one that best compresses the observed or available data.

Solomonoff induction has a number of curious characteristics: the Solomonoff prior is incomputable; the prior is highly language-dependent (i.e., dependent on the subjective choice of the universal Turing machine which determines its definition); and it presupposes the examined hypotheses to be computable. While these characteristics pose a challenge to the practical implementation of Solomonoff induction by ordinary or even idealized humans, they can nevertheless be appreciated as useful pragmatic constraints. Take for instance, the incomputability of the Solomonoff prior. The