{...}

Ghost in the shell (1971)[1]

There is supposed to be an argument from Gödel's theorem to show that the mind can't be a machine, but I've never understood it. Of course I have never thought that was my fault.[2]

Penrose, for one, made a book out of it; and though I didn't believe him either it was amusing that whole issues of the journals[3] were re-purposed to try to refute him.

At any rate both sides of the argument are bullshit. It doesn't matter whether minds are machines or not. Even machines aren't machines. This can be seen in at least two ways:

---

[1] The basic argument (here reconstructed) has not changed a great deal since it first occurred to me, though it has obviously been revised and amended to reflect the march of mathematical progress.

[2] There are (at least) a couple of good reasons to be skeptical. — First, Gödel himself always thought this was a consequence of his incompleteness theorem, and was said to have been working on a formal proof of the proposition; which, however, he never finished, and didn't publish. That seems suspicious. — Second, the idea that a human agent could find itself trapped in repetitive cycles of mechanical behavior is supposed to be prima facie absurd; nonetheless it is the fundamental thesis of psychoanalysis; and indeed the Freudian method looks a lot like teaching a Turing machine its Gödel sentence. (Compare Thomas Mann: "No man remains what he was once he has recognized himself.")

[3] The book [published in 1989] was *The Emperor's New Mind*. For expressions of outrage cf., e.g., *Behavioral and Brain Sciences* Vol. 13 #4 (1990), pp. 643-705, and Vol. 16 #3 (1993), pp. 611-622.

— First, "machine" in the sense of artificial intelligence never really means "Turing machine" anyway; rather one augmented by a (true)[4] random number generator — a source of randomness for nondeterministic algorithms; neural networks, for instance, fall under this description, as do Metropolis and genetic algorithms.

— Second, even deterministic machines aren't deterministic. — That is, though you have the picture (sharpened by formal models) of a system with a delimited[5] set of states, whose behavior is determined by a function which computes the next state (in a discrete series) from the current state, and assume that knowledge of the next-step function entails knowledge of all its iterates — that "deterministic" means "completely predictable", in other words — this is a kind of optical illusion. It doesn't really work that way.

{...}

It's more amusing to explain this anecdotally.

I saw Skinner lecture once, in Berkeley in the early Seventies. This was shortly after the publication of one of his numerous paeans to

---

[4] I.e., one referring to what theory terms "an Oracle", some external source of input like a Geiger counter recording radioactive decays. — Purely computational (pseudo) random number generators fake it, by producing sequences which are determined by such complicated rules that they "look" random (a literature has been expended trying to define the implicit oxymoron), i.e. take a long time to repeat, but on the other hand can be rapidly computed. It is truly amazing how often the naive employment of these mechanisms leads to mortifying blunders. Nearly every serious programming project I have undertaken has been almost immediately been sidetracked by an attempt to write a better random number generator than the one that has just fucked me in the nose.

[5] This is tricky: machine theory allows not simply for the case of a finite state set, but also for a finite "internal" state set augmented by potentially-infinite auxiliary storage, the tape of a Turing machine or the stack of a PDSA, e.g., which can only be accessed finitely, e.g. one item at a time. — In practice, of course, all machines are really finite, and immense ingenuity is expended to overcome limitations of time and space.

Mind Control:[6] he spoke in a large lecture hall, to a full house packed with an extremely hostile crowd, and though he couldn't win them over, he did at least earn their respect. — There is a certain naive pig-headed charm some nerds possess, and he had it in great measure. If nothing else, I admired his balls.

We were all jammed in like sardines, and I was sitting in the aisle a few feet downhill from my girlfriend, so as it turned out I couldn't talk to her until afterward and it wasn't obvious we were attached. Instead I found myself embedded among a covey of attractive female undergraduates. One of them was lecturing her friends on the nature and context of the debate we were participating in, and every time she hesitated because she didn't quite know how to continue, I finished her sentence for her. — This provided me with the standard anecdote I used in later years to describe what Berkeley was like, in the Golden Age: this was the first, last, and only time a girl wanted to go home with me because I knew Beckett wrote *Endgame*.

At any rate I was fascinated by Skinner's insistence on the predictability of human behavior; there was an echo of that Freudian certitude that had always seemed so maddening, but his explanatory apparatus was cleaner, much more austere. So what was wrong with it?

Part of it, obviously, when I read over theoretical behaviorism later[7] to find the basis for his claims, was that the most consistent version of his approach made it a point of dogma not simply that one *should* not but that one *could* not assign internal states to the organism; since simple thought experiments showed that removing the brain from the skull would produce a noticeable difference in behavior, at least among people who hadn't voted for Nixon, that was obviously wrong. — Part

---

[6] Probably *Beyond Freedom and Dignity* [1971].

[7] Not that I wasn't familiar with it already from, e.g., Russell's synopses in *The Analysis of Mind* [1921], but it was instructive to read the modern literature and observe how little theory had progressed since Watson and Pavlov.

of it was that the kinds of laboratory experiments to which behaviorists confined themselves made essentially meaningless measurements of a kind which could not, for instance, tell you anything about the functioning of even the simplest digital computer.[8]

But the main thing — what was instantly suspect — was his claim that behaviorist methods would suffice to explain even the "behavior" of mathematicians. For this seemed, after all, to be a bizarre assertion: were we seriously to think that from considerations of elementary physics — presumably by solving some system of differential equations — not that Skinner ever wrote any down, of course, but an explanatory framework based on the measurement of quantities expressed in real numbers — i.e. founded on physics envy — would inevitably (as any real physicist could instantly see) lead to such a theory — that we could tell whether a mathematician was going to be able to prove a theorem? How was one *complicated* mathematical problem — indeed all of them at once — supposed to reduce to another which seemed so much simpler?[9] — And *why* it seemed bizarre wasn't difficult to figure out. For though if we asked the mathematician to prove, say, some statement in the predicate calculus it might seem unlikely on

---

[8] A technical refinement of the point, which Chomsky used to great polemical effect, was that though for the simplest class of finite-state automata internal states can in principle be defined away as equivalence classes of mappings from inputs to outputs, this [a] relies on the examination of infinite sets, and [b] the conditioned-reflex prescription applied to a finite training set of stimuli and responses only works for this simplest class, and cannot determine the behavior of machines that recognize more complex grammars. Since such machines already existed and even then were generating our utility bills, this seemed a fairly crushing objection.

[9] Actually it isn't impossible that a relatively simple differential equation, or system of them, could be universal in the sense of Turing; the solution of Hilbert's tenth problem showed something analogous for Diophantine problems, i.e. that there is an equation of the fourth degree in 14 variables that is universal: see Martin Davis, "Hilbert's Tenth Problem Is Unsolvable," *American Mathematical Monthly*, March 1973, 233-269.— One could conjecture, in other words, the existence of a universal *analog* computer. — But a simulation that modeled a universal Turing machine with a differential equation wouldn't be any *simpler*. The inherent difficulty of the problem is irreducible. So the picture you have of having found a solution is a kind of optical illusion. — "All I have to do is solve this equation, and..." — but how? In practice you have only replaced one intractable computation by another of equivalent difficulty.

intuitive/romantic grounds that we'd be able to describe the necessary "creative leap", really it isn't necessary to appeal to this at all: one could simply ask the mathematician to attempt mechanically to construct a proof using some method like semantic tableaux; and then observe that whether this procedure terminates on arbitrary input is, in general, undecidable. — I.e. you needn't appeal to a magical black-box mechanism at all; even if you know the mechanism, even if the box is transparent, it makes no difference. — So the grand reductive gesture of pretending the box *has* no internal degrees of freedom is doubly pointless.

{...}

Put another way, one need not challenge Skinner with the problem of predicting whether, say, Gauss sitting at his worktable will be able to come up with a proof of, say, Goldbach's conjecture;[10] one can simply ask Skinner to tell us whether Gauss in performing the arithmetical check will find a counterexample to Goldbach's conjecture in finite time; and if so, *when*. Because this means that the behaviorist must then in effect be able to tell us in advance whether Goldbach's conjecture is true. (And decide this by solving some magic differential equation, or system of them.)[11] — True, we can, if we are faster, stay ahead of Gauss in the computation. But this is not an *effective procedure*; we

---

[10] Communicated in a letter to Euler in the 18th century, the statement (based then on very flimsy empirical evidence, based now on dismayingly extensive tests) that every even number greater than 2 is the sum of two primes. A proof now does appear to be closer, but the feeling has generally been that if there really are "natural" elementary statements about the integers that are true but not provable, they would look like this. (Gödel himself referred to this possibility explicitly; see the notes to *Gödel 1972a* in his *Collected Works*, Volume 2.)

[11] Given the hypothetical universal system one might solve the equations "by computer", i.e. numerically, but then we simply have one machine emulating another of equivalent complexity; nothing is *reduced*, in other words.

can't guarantee an answer to the question exists in advance.[12] — We can't say how the computation will come out. — And therefore, in the most significant sense, we cannot *predict* what Gauss is going to do, *even if he is emulating a machine*.[13]

{...}

There are various equivalents[14] that illustrate the case equally well, but the canonical question is the halting problem for Turing machines: suppose we give Gauss the description of a Turing machine, and an input tape — all this is finite — and then ask Skinner to tell us his prescription for deciding, in the general case, when/whether Gauss will finish computing the answer, and what the result will be. — To explain his behavior, i.e. — But he can't, because this is known to be impossible. — Conceivably Skinner might object that the proof of unsolvability assumes the validity of Church's thesis, an essentially metaphysical hypothesis[15] which he rejects — another myth which will dissolve in the acid bath of his scientific rationality; but then he's saying that he has some method of computation (an oracle, e.g.) that is more power-

---

[12] I.e. though I may not know before I perform the computation that 16117667 times 16283543 is 262452723654181, I do know that there is an answer, and if I follow the rules for multiplication I will find it within a certain number of steps which can be bounded in advance. Not all computations come with such guarantees.

[13] I take it for granted that a human (like Gauss) can emulate any Turing machine; since after all the idea of the Turing machine is that it formalizes the abilities of a human calculator. — It is assumed, in other words, that the objection that Gauss might not have enough time or scratch paper is frivolous and irrelevant to the principle at issue. (This has nothing to do with his *behavior*.)

[14] The word problem for semigroups, e.g., which asks whether there's a general method for deciding whether two strings of symbols are equivalent under a given finite set of equational transformations, or the general Diophantine problem (Hilbert's Tenth), whether an mechanical procedure exists to determine whether a polynomial equation in a finite number of variables with integer coefficients has integer solutions.

[15] Fred Thompson was the first guy I heard call it that. He was certainly right.

ful than a Turing machine. — At which point we tell him to put up or shut up. And the rest is silence.

{...}

You can summarize the lesson of this gedankenexperiment as follows: since prediction is simply computation,[16] machines in general are not predictable; since people can emulate arbitrary machines,[17] the behavior of people is not predictable.

So behaviorism isn't completely useless; its refutation teaches us something valuable.

{...}

This doesn't explain why a mob of hippies showed up to howl for Skinner's head, of course. That had to do with the supposed conflict between the freedom of the will and determinism. But I think the real issue there is related, essentially psychological, the anxiety that you feel about the possibility not that your actions are "determined" in some complex and unknowable fashion, but that they can be *predicted*.

We all remember Dostoevsky's lengthy rant[18] in *Notes from the Underground*, the famous Crystal Palace passage about the conflict between the freedom of the will and mathematical certainty:

> ... then, you say, science itself will teach man ... that he never has really had any caprice or will of his own, and that he himself is

---

16 This seems self-evident, but in the same way all propositions do that insinuate metaphysical hypotheses. (Here again Church's thesis.)

17 By definition: when Turing refers to a "computer" in his original paper, he means a human following rules with pencil and paper; electronic computers did not yet exist.

18 It would be anachronism to call it that, but this is a classic example of what is now called a flame.

> something of the nature of a piano-key or the stop of an organ, and that there are, besides, things called the laws of nature; so that everything he does is not done by his willing it, but is done of itself, by the laws of nature. Consequently we have only to discover these laws of nature, and man will no longer have to answer for his actions and life will become exceedingly easy for him. All human actions will then, of course, be tabulated according to these laws, mathematically, like tables of logarithms up to 108,000, and entered in an index; or, better still, there would be published certain edifying works of the nature of encyclopaedic lexicons, in which everything will be so clearly calculated and explained that there will be no more incidents or adventures in the world.[19]

Or more succinctly:

> Good heavens, gentlemen, what sort of free will is left when we come to tabulation and arithmetic, when it will all be a case of twice two make four? Twice two makes four without my will. As if free will meant that!

But though the existentialist antihero of the *Notes* thus insists perversely on behaving irrationally to express his defiance of soulless rationalism, he needn't have bothered. Arithmetic itself is perverse enough.

That is, though it is already difficult enough to understand the traditional problem — your *will* is still free even if what you *want* is determined, [20]and so what — the point is really that determinism appears to entail predictability, and prediction allows control: if people are ma-

---

[19] This is the Constance Garnett translation.

[20] I cheerfully admit that emotional responses are often predictable, at least for most people much of the time; else they would be more difficult to manipulate. But here again the claims of psychology are exaggerated.

chines, then seemingly they can be *used* as machines; *that* is the terror of mechanism.

You have the oppressive sense that some puppet master like Skinner can look over your shoulder (with his "table of logarithms") and nod smugly at everything you do, because he has foreseen it all in advance; and since he knows what you will do when he pushes your buttons, *he can make you do whatever he likes*. — And Skinner of course endorses this interpretation at every turn, this is the plan for his Utopia. — That it might be determined[21] in advance but not known or even knowable — well, there is something that never occurred to the determinists; omniscient though they were supposed to be. In fact it doesn't seem to have occurred to anybody.

{...}

The anxiety is not unknown among physicists. There is a strong resemblance, e.g., between Eddington's argument (made nearly as soon as the uncertainty principle was invented)[22] that the indeterminacy of quantum mechanics permitted the freedom of the will, and Penrose's rather weird assertion (1989) that "microtubules" within the cell could turn the brain into some kind of quantum computer beyond the reach of Turing.[23] In both cases it is clearly the predictability of the mechanical that disturbs them. — Your will cannot be free if someone can know what you will do. — More than that, an artist or a musician or a mathematician cannot be truly creative, since whatever they produce is simply the result of a mechanical process. One could simply write a

---

[21] To return to the model of the system of differential equations, there are in general existence theorems that tell you they *have* solutions which are determined uniquely by their initial conditions. This doesn't mean you can say what the solutions are. (Or — the butterfly effect — that they are stable under infinitesimal perturbations, which is a necessary condition for computer simulation.)

[22] Cf. *The Nature of the Physical World* [1927].

[23] Pure science fiction, so far as anyone can tell.

Shakespeare emulation program and output *Hamlet*, without the intervention of the fifty million monkeys with typewriters. — This is a slightly more interesting problem, but the predictability issue is again key: one might in principle be able to program a (pseudo)machine to write something like *Hamlet*, but it would never turn out the same way twice, and given time and sufficiently many rewrites would turn into something else entirely. — Whether that would satisfy Penrose I don't know. But my credentials as an unreconstructed Romantic are unquestioned, and it satisfies me.

{...}

There is also an amusing functional equivalence between Skinner's implicit[24] assertion that he could predict the answer to any mathematical question from the laws governing the organism (the differential equations, or whatever) and Plato's insistence that all mathematical knowledge is something the soul obtained in a previous life/is engraved upon the Forms; it is accordingly suggestive that they envisioned similar Utopias. (And that they bore a suspicious resemblance to the Crystal Palace.) — Who were our behaviorist overlords going to be, but the new Guardians? — Moreover there are parallels with the apparent aims of the classical school of artificial intelligence, as exemplified by Minsky: if the brain was just a machine running a determined program, then those select few who could read the source code could make mere humans (aka "the lusers")[25] do whatever they wanted; traditional hacker culture was also based on fantasies of control, the domination of the programmers over the programmed.

{...}

---

[24] You have to say "implicit" because it is obvious he did not understood what was coming out of his mouth. Certainly he never understood Chomsky's critique.

[25] Traditionally the MIT school divided people who interacted with computers into two classes, programmers and users; the former were the master race, the latter, serfs and peons. — It is not an accident that, as Big Tech continues to conquer the world, more and more of it reverts to feudalism.

Another fantasy of determinism, indulged by the imaginative, is that one ought to be able to predict the course of history in advance. — This is not, precisely, the usual motivation of the self-styled grand theoreticians of history, who seem not to have advanced beyond Linnaean notions of classification — Spengler, e.g., goes on at great length in his philosophical preamble about Goethe, morphology, the incapacity of trivial concepts of causality to grasp the architecture of Destiny, etc.[26] — the game of hypothesis and prediction never caught on among the German idealists, obviously — but it is a fairly common speculation in science fiction. Isaac Asimov's *Foundation* novels are probably the most famous examples, and have been quite influential[27] in that respect: he imagines the decline and fall of a galactic empire on the pattern of Gibbon's Rome, and a dedicated cabal of monks, privy to detailed advance knowledge of the pattern history must follow, working to preserve civilization through the ensuing Dark Age, whose duration they will thus be able to minimize.[28]

The superficially convincing argument for the possibility of such prescience is the analogy with statistical mechanics: you don't need to know how each individual gas molecule is moving to calculate the pressure on a cylinder. — The argument probably fails on appeal to

---

[26] In the translation of Charles Francis Atkinson: "The means whereby to identify dead forms is Mathematical Law. The means whereby to understand living forms is Analogy." — "... there can be no question of taking spiritual-political events ... at their face value, and arranging them on a scheme of 'causes' or 'effects' ... ." — "That there is, besides a necessity of cause and effect — which I may call the logic of space — another necessity, an organic necessity in life, that of Destiny — the logic of time — is a fact of the deepest inward certainty... ." — "Mathematics and the principle of Causality lead to a naturalistic Chronology and the idea of Destiny to a historical ordering of the phenomenal world." — And so on. Of course all this is nonsense.

[27] Sometimes in unobvious ways: the Nobel laureate Paul Krugman, for example, is a science fiction fan, and has often remarked that Asimov's vision of a social science that could make rigorous predictions inspired him to study economics.

[28] This idea of a monastic order preserving knowledge through a Dark Age is another favorite theme of science fiction; see for instance Walter Miller's *A Canticle for Leibowitz*.

the butterfly effect, since there are many examples e.g. of critical battles won or lost by accidents of timing, and (pace Tolstoy) great men (and women) do seem to appear fortuitously and decisively alter the course of events — this is a more complex dynamical problem than that posed by a gas, after all — still, though one can't predict the weather exactly, one *can* predict climate change; so one might guess that on a longer time scale the rolls of the human dice may even out.

Nonetheless something similar to Skinner/Gauss does apply: the future of industrial civilization as we have it right now, for example, depends at bottom on facts of physics and astronomy as yet unknown — whether room temperature superconductors exist, whether fusion reactors can ever be practical, what results may come from mining the asteroids, whether irreversible ecological collapse is really at hand — whether an undetected asteroid is going to run into the Earth and reprise the extinction of the dinosaurs — and you can't tell how human history will turn out without knowing the answers to these external questions. — As was the past so determined: the history of the modern world follows in large part from the contingent fact that when Columbus sailed west, there was an extra continent to bump into. — So the one kind of omniscience presupposes the other. Even economics, which involves measurable quantities and superficially seems more easily predictable, depends at bottom on the ways that we can extract free energy from our environment, and thus on unpredictable boundary conditions and undiscovered facts of mathematics and physics (and chemistry and biology and geography and ...) which cannot be known without — well, without being known.[29] How could an economist in 1950 have predicted that nuclear power based on fission reactors would turn out to be more trouble than it was worth, or foreseen the laser, the transistor, the photovoltaic cell, the microchip, or Moore's Law? — Von Neumann saw none of that coming, and he was as omniscient as anyone could have been at that time — for instance,

---

[29] Here I'm sure Heidegger would insert some rhapsody on the knowable knowingness of being-known, but — thank the gods who have not yet fled — I lack his gift for tautological obfuscation.

he famously stated that four computers like his[30] primitive MANIAC[31] would suffice for all the computational needs of the world.[32]

{...}

A slightly weaker statement, whose relationship to undecidability is still not completely understood, is that a computation may not be impossible but nonetheless may be prohibitively difficult. This might seem like a frivolous objection were it not the case that relatively simple problems can be shown to be unsolvable within existing space and time.[33]

---

[30] Actually constructed by Nicholas Metropolis at Los Alamos following Von Neumann's IAS design, but why quibble. — Authorship of the acronym, which was meant to stamp out this reprehensible practice in its infancy and failed miserably, has been ascribed to both.

[31] Less powerful than a pocket calculator of the Seventies, and many orders of magnitude less powerful than the contemporary iPhone; which exceeds in computational power the fastest supercomputers of even the Eighties.

[32] As a final note, Lockheed Martin is supposed to be pitching a tool called the World-Wide Integrated Crisis Early Warning System (Google at your own risk), originally a project funded by DARPA, which is supposed to have had some success anticipating national and international crises. Apparently among other things it predicts the collapse of the Russian government within a couple of years; surely a consummation devoutly to be wished. — The historian Peter Turchin, on the other hand, on the basis of mathematical analysis of a large data set measuring a variety of historical trends, finds parallels between previous periods of crisis and the current situation of the United States, and predicts the disintegration of civil society within the decade. — And he does indeed begin *War and Peace and War* [2006] by invoking the example of Asimov's hero Hari Seldon.

[33] David Ruelle ("Is Our Mathematics Natural?" *Bulletin of the AMS*, Vol. 19, Number 1, July 1988) mentions a suggestion of Pierre Cartier that the axioms of set theory might be inconsistent but a proof of this would be so long that it couldn't be performed in the physical universe.

One class of examples would include the travelling salesman problem, which scales exponentially in the number of cities;[34] a greater degree of difficulty may be found in problems like computing Ramsey numbers, or evaluating the Ackermann function, which is defined as follows:

$$A(x,0) = 0$$
$$A\left(0,y\right) = 2y$$
$$A(x,1) = 2$$

else

$$A\left(x,y\right) = A(x-1,A\left(x,y-1\right))$$

Then

$$A(1,n) = 2^n$$
$$A(n,1) = 2$$
$$A(n,2) = 4$$
$$A(2,3) = 16$$
$$A(2,4) = 65536$$

in general

$$A(2,n) = A(1,(A(2,n-1)) = 2^{A(2,n-1)}$$
$$A(3,1) = 2$$
$$A(3,2) = 4$$
$$A(3,3) = 65536$$

and

$$A(3,4) = \ldots$$

---

[34] Given a planar map and the positions of n cities upon it, to construct a route of minimum length that visits each city exactly once; for n around 120 the number of possibilities that must be examined exceeds the number of cells of dimension the Planck length in the visible universe.

i.e., this is a recursion that will not terminate before the stars go out, and the answer couldn't be written down[35] if you used all the volumes in Borges' Library of Babel.

{...}

Regarding the Goldbach conjecture, subsequent developments have only confirmed Gödel's intuition. Certainly it is possible that there is some simple and elegant proof of this proposition, but it seems more likely there is not; and then there are curious questions about how complicated a proof, even if one does exist, might have to be. The proof of the celebrated four-color theorem,[36] for example, another result with an extremely simple statement[37] which defied demonstration for several generations, turned out not to involve (at least has not thus far) the elegant manipulation of powerful abstractions developed from mathematical theories of great scope and formal beauty — as did, for instance, the proof of the famous Weil conjectures (Deligne 1973), the proof of Mordell's conjecture (Faltings 1983), and the celebrated proof of the Taniyama conjecture (Wiles 1993/4), which entailed the last theorem of Fermat, for three centuries the most famous unsolved problem in the subject — but rather the enumeration and systematic

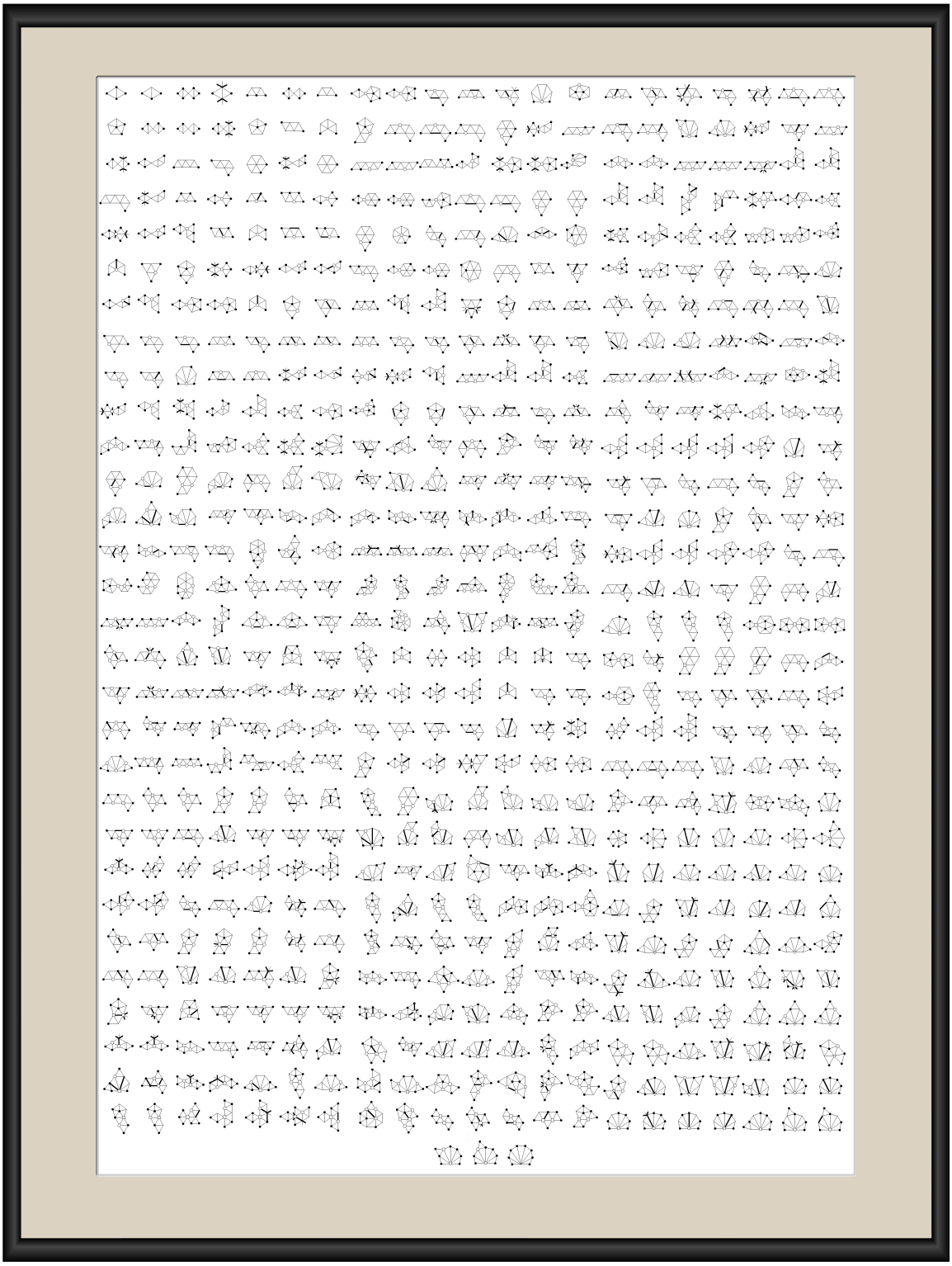
---

[35] In decimal notation, at least. Of course in effect we have already specified the number with a small finite number of symbols.

[36] Appel and Haken (K. Appel and W. Haken, "Every planar map is four colorable, Part I: discharging," *Illinois Journal of Mathematics*, **21** (1977) 429-490; K. Appel, W. Haken, and J. Koch, "Every planar map is four colorable, Part II: reducibility," *Illinois Journal of Mathematics*, **21** (1977) 491-567) considered more than 1900 configurations and more than 300 so-called discharging rules; the proof was so complicated that no one could simplify or even check it for twenty years. Finally Robertson, Sanders, Seymour, and Thomas (N. Robertson, D. Sanders, P.D. Seymour, and R. Thomas, "The four-colour theorem," *Journal of Combinatorial Theory*, Series B **70** (1997), 2-44) reduced its complexity to 633 configurations and 32 discharging rules — a simplification which allowed a complete proof to be written out and verified by computer. — An executive summary is provided by B. Bollobás, *Modern Graph Theory*. Berlin: Springer-Verlag, 1998; pp. 159-161.

[37] Specifically: that any map in the plane can be colored with no more than four colors in such a way that no two contiguous regions have the same color.

elimination of over a thousand separate cases, handled mechanically by a computer program and not, at least not immediately, understood directly by any human mathematician. This engendered a rather painful debate, and raised ugly questions: is there any guarantee a cleaner proof exists? are many unsolved propositions with simple statements destined to have similar resolutions? and so on. — Mathematics is supposed to be an elegant duel with light-sabers, not some kind of rude barbarian combat in which the victor clubs his opponent to death.



The 633 configurations of Robertson, Sanders, Seymour, and Thomas.

{...}

One might contrast the solution of the game of checkers, obtained by researchers at the University of Alberta; they examined 500,000,000,000,000,000,000 different configurations to show that there is a strategy for the game that does no worse than draw.[38] But there is nothing particularly shocking about this, because the rules of checkers are the product of a kind of caprice, and games of strategy in general have unbounded logical complexity;[39] in fact it's almost surprising it was this easy. — One would expect chess and Go to be solvable in similar fashion, though it is difficult to imagine that a computer could finish enumerating the cases before the heat death of the universe.

(It is instructive, incidentally, to consider the case of a human playing a machine at chess; the moves of the latter are completely determined by a set of algorithms; the moves of the former are not, and it is obvious no behaviorist ever considered the question of how they could be reduced to a finite set of conditioned reflexes[40] — this despite the fact that programming a computer to play chess was one of the first problems that occurred to the pioneers of artificial intelligence.)[41]

---

[38] Jonathan Schaeffer, Neil Burch, Yngvi Björnsson, Akihiro Kishimoto, Martin Müller, Robert Lake, Paul Lu, Steve Sutphen. "Checkers is solved." *Science* 14 September 2007: Vol. 317, Issue 5844, pp. 1518-1522. The program (Chinook) can be played online.

[39] As Ulam was fond of pointing out, questions about games of strategy nest quantifiers to arbitrary depth — the problem of chess, e.g., can be stated as whether for all opening moves by white there exists a move by black such that for all moves by white there exists a move by black such that, etc. — whereas in normal mathematics few definitions (Ulam's pet example was that of an almost periodic function) nest them more than four or five deep.

[40] Could operant conditioning be employed to teach a rat to play chess? — No? (Why not?) — What about tic-tac-toe? — If one rat can't be conditioned to play chess, can a roomful of them? Enquiring minds want to know.

[41] Turing himself wrote one of the first such programs; apparently to revenge himself upon his colleagues at Bletchley Park, who pissed him off by beating him so consistently.

{...}

But mathematics is supposed to be *necessary* truth. When a simple question has an enormously complicated answer[42] it looks like truth by accident.

In the case of Goldbach's conjecture results have been obtained which show that a related proposition holds for all numbers greater than an enormous lower bound; though thus far this lies far beyond the range of possible computation, it is conceivable that some combination of faster computers and improved lower bounds could make it possible to construct a complete proof by pasting together an analytic result (true for even numbers greater than some enormous N) and brute force enumeration of the rest of the cases (verified by explicit computation for even numbers less than or equal to N). If this were the case, it would present us with an example of a number-theoretic theorem about the integers, what we would like to think of as quintessential necessary truth, which would nonetheless have the appearance of being true only by accident. —Wittgenstein would have loved this, but no one else.[43]

(Obviously it is also disturbing that a proof based on a computer program depends on a proof that the program is correct; these in practice are practically impossible to provide, and, handwaving arguments about the probability of error being vanishingly small not-

---

[42] The usual situation goes the other way around — a complicated problem has a simple solution: the problem of the thirteen pennies, for example. [Not sure I can explain that without a diagram, and how are diagrams included in footnotes? Hmmmm.....]

[43] *Remarks on the Foundations of Mathematics*, III.42: "It might perhaps be said that the synthetic character of the propositions of mathematics appears most obviously in the unpredictable occurrence of the prime numbers. ... The distribution of primes would be an ideal example of what could be called synthetic a priori, for one can say that it is at any rate not discoverable by an analysis of the concept of a prime number." (Translated by G.E.M. Anscombe. Cambridge, M.I.T. Press, 1967.) — This sounds surprisingly Kantian, but there is something intuitively correct about it.

withstanding, it isn't immediately obvious that we haven't been presented with an infinite regress.)

{...}

Appended note:

The march of mathematical progress has now brought this scenario to fruition: the weak Goldbach conjecture, which states that that every odd number greater than 5 is the sum of three odd primes, had been proven by the refinement of analytical techniques due to Hardy, Littlewood, and Vinogradov, among others, to be true for all numbers greater than a bound C; a series of attempts to lower C had [2002] reduced it to about 10^1350, still far beyond the reach of computer verification. Recently, however, Helfgott[44] has lowered C to 10^27 and since computational efforts[45] have extended numerical verification nearly to 10^31, the proof-theoretic chimera has now been stitched together. — The question remains whether further refinements of these techniques can gradually reduce C to some value more satisfying to intuition: 10 certainly would work, but 100? 1000? 1000000? — Where to draw the line? — In the meantime, though the weak Goldbach conjecture is now known to be true, it falls into a kind of uncanny valley[46] between the analytic and the synthetic.

{...}

---

[44] H.A. Helfgott, "The ternary Goldbach conjecture is true"; arXiv:1312.7748v2, 17 January. 2014.

[45] These too rely on (partial) empirical verification of another open question, the Riemann hypothesis, for which enough zeroes have been computed to bound the gap between successive primes sufficiently well up to 10^27 that an odd prime can be subtracted from the triple to yield an even number less than the limit to which the even Goldbach conjecture has been verified, of the order of 10^18. Not to take anything away from Helfgott's remarkable achievement, this argument is a ridiculous kludge.

[46] A term used in computer graphics to designate the disturbing gap between the obviously phony and the photorealistic. Thus synthesized faces possess an unsettling quality.

A similar simple proposition about the primes no one has any idea how to prove is the twin prime conjecture: that there are an infinite number of pairs (p, p+2) which are both primes.[47] — About this Cohen after expressing skepticism regarding the ability of axiomatic frameworks to capture the properties of the mathematical objects they describe asks "Is it not very likely that, simply as a random set of numbers, the primes do satisfy the hypothesis, but there is no logical law that implies this?"[48]

In fact the natural metamathematical conjecture is that almost all[49] conjectures that appear to be true on probabilistic grounds are true but unprovable; i.e. that these two senses of "true, probably" and "probably true" are equivalent.[50]

The Goldbach example suggests that there may be many conjectures with relatively simple statements whose probability of truth is unity (since they can be shown to be true on the complement of a finite set) but which then are true or false globally by a kind of contingency, in that the proof can only be filled in by case by case enumeration. Indeed this situation may be typical.

{...}

To state one moral, then: philosophical intuition is not completely worthless, but like any other kind of intuition it is based on a kind of

---

[47] See (xix).2003.7.8, "Minor triumphs".

[48] Paul J. Cohen, "Skolem and pessimism about proof in mathematics", *Phil. Trans. R. Soc. A* (2005) **363**, 2407-2418. (12 September 2005.)

[49] "Almost all" has the technical definition "except on a set of measure zero" and doesn't really mean anything unless such a measure can be defined. Here it can.

[50] For reasons that may be obvious this occurred to me while meditating gloomily on a lecture about the abc conjecture.

experience; and it should not, therefore, be a surprise that its conclusions evolve when that experience broadens.

The analytic/synthetic distinction was introduced by Kant; was almost immediately questioned by Gauss, who had already understood the possibility of a non-Euclidean geometry; and then revised after radical extensions of the idea of entailment to include inferences like "*7 + 5 = 12*" and "a straight line is the shortest distance between two points", even though (as Kant pointed out) neither falls under the traditional definition of a conclusion being included in the premises.

Now, it becomes clear, it may be less a black and white distinction than a grayscale continuum, resolving under closer examination into an arbitrarily ramified hierarchies of the kind with which we have lately become familiar in complexity theory. — The more extensive our experience of what constitutes proof, the more baroque may our intuition of necessity become.

{...}

Fundamental misconceptions about mathematics and the nature of prediction notwithstanding, there was a larger fallacy involved in behaviorism: it was based upon an artificially limited, indeed an essentially invalid idea of what constituted science.

You could see it in the polemics Skinner's partisans wrote against Chomsky — here was a real theory of language at last, or at least a piece of one, and it was attacked as unscientific because it was (in Eddington's phrase) physics and not stamp collecting; because they not only did not recognize theory when they saw it, they did not understand its necessity — because they had trapped themselves in the most limited possible conception of empiricism, almost a throwback to Bacon, one in which scientific endeavor consisted entirely in the blind accumulation of disconnected "facts"; the reduction of the philosophy of nature to making statements in an observation language — which, of course, they didn't even see was ill-defined.

I suppose this was natural. Psychology had spun its wheels from Hume to William James trying to found itself in introspection. A radical break seemed to be called for, what more comprehensive revolt against subjectivism than to deny the existence of the subjective entirely, and in so doing why not banish all "metaphysical" statements altogether? this was the spirit of the age, after all.

{...}

There was, in other words, a desperate anxiety among psychologists that what they were doing was not "science". And quite understandably they sought to make what they were doing "scientific" by imitating what they saw their intellectual elders doing: performing experiments in laboratories and making measurements that produced copious amounts of numerical "data" — publishing "results" in "papers" in "journals", filling them with graphs, tables, charts, and statistical analyses — *going through the motions* — hoping that, by performing the same ritual abasements as (real) biologists, chemists, and experimental physicists, psychologists could acquire their mojo. — There is a name for this, and it is not "scientific thinking".

{...}

Freud gives as examples of what Frazer called imitative or homeopathic magic the following:

> Rain is produced magically by imitating it or the clouds and storms which give rise to it, by 'playing at rain', one might almost say. In Japan, for instance, 'a party of Ainos will scatter water by means of sieves, while others will take a porringer, fit it up with sails and oars as if it were a boat, and then push or draw it about
>
> the village and gardens'. In the same way, the fertility of the

> earth is magically promoted by a dramatic representation of human intercourse...” and summarizes the principle as follows: “If I wish it to rain, I have only to do something that looks like rain or is reminiscent of rain.[51]

Later Feynman described the practice as follows:

> In the South Seas there is a cargo cult of people. During the war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they’ve arranged to make things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas — he’s the controller — and they wait for the airplanes to land. They’re doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn’t work. No airplanes land. So I call these things cargo cult science, because they follow all the apparent precepts and forms of scientific investigation, but they’re missing something essential, because the planes don’t land.[52]

He did not, however, recognize that the cargo-cult phenomenon extends beyond pseudoscience into what is supposed to be “science” itself. — Behaviorism had a theatrical run of a couple of generations. But the planes never landed.

{...}

So that is one way of putting it: the cargo cult imitated the “behavior” of the operators perfectly, but didn’t look inside the radios to see what made them work. There must be a moral there.

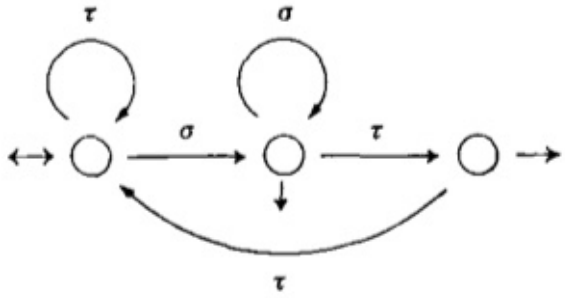
---

[51] Sigmund Freud, *Totem and Taboo*, transl. James Strachey, London: Routledge Classics, 2001. Chapter 3, “Animism, Magic, and the Omnipotence of Thoughts”.

[52] *Surely You’re Joking, Mr. Feynman!* New York: W.W. Norton, 1985.

Another way of putting it is that no one ever said what "behavior" was. There was some vague appeal to observable physical states of the organism, but "observable" and "state" and "organism" weren't defined, and the "state" per se wasn't what was referenced in any case, rather some notion of "action", presumably definable in terms of a (short, finite) temporal sequence of states — though this wasn't defined either, of course. So for all anyone could tell "behavior" might include the response of the subject to cold in the form of goosebumps, or to ultraviolet light in the form of sunburn — note that the response in this case varies dramatically from one subject to another, and that no kind of input-output table relating insolation to degree of burning will say anything about the chemistry of melanin, the real causative factor — or the humidity of expelled breath, or height and weight, or for that matter what the subject said in response to the question "What are you thinking about?" — Behavior could have been *anything*, until it was defined. In fact simply by declaring it to have meant the microstructure of brain activity, to which real scientists more sensibly have turned their attention, the program could now be pronounced a success.

Again: a rigorous definition of "behavior" would entail definitions of "stimulus" and "response", and those in turn would require an enumeration of possible inputs and outputs. — Implicitly, as Chomsky pointed out, the mathematical model behind the smoke and mirrors here is that of a finite-state automaton, which takes as inputs strings of symbols selected from a finite alphabet; each one inducing a state transition for which, in turn, a string of symbols from a finite output alphabet is produced; this may be pictured, e.g., as a state transition graph with edges labelled by inputs, for instance:[53]



It would then be easy to define away the internal states of the machine as equivalence classes of maps from inputs to outputs, and a kind of behaviorist program can be said to have succeeded.

But where do the input and output alphabets *come from*? Some kind of language is presupposed to specify just what "observable behavior" is, and in practice there is that familiar philosophical bait-and-switch, the appeal to self-evidence, and a host of unexamined assumptions are insinuated by inclusion and omission. And so we have ring bell/salivate, shout/wince, electric shocks and bits of cheese, and not, say, observations of the form "I rebuked him, and observed that he took offense at the harshness of my manner of expression" — though why not, no one will ever bother to tell you. — "Measure something" — but why measure *this* and not *that*? — And the answer, of course, is that what is and is not relevant has been decided by an implicit appeal to an unstated theory, something beyond the reach of scrutiny. — Elsewhere this is styled "metaphysics".

Moreover in practice you have only a small subset of the input-output mapping, and have to guess the rest — another version of the problem of induction — and the most natural theoretical device employed to model it is, guess what, an internal state space whose transformations are induced by inputs — in the language of the electrical engineering lab, the wiring of the black box; for a given set of pairs {(input, output)} there will be an ensemble of possible wirings, some kind of maximum-entropy probability distribution imposed upon it, and)ideally) an optimal set of yes/no experiments that will, in the limit, identify the correct internal configuration and thus determine the mapping.

But obviously this is too complicated for a psychologist to appreciate. It may be better to let them keep playing with the knobs on their empty boxes.

{...}

At any rate the simplest objection is still the most powerful: if the brain is a relatively trivial mechanism programmed by the conditioned reflex, then it shouldn't be difficult to reverse-engineer it, and build a model of one.[54] — So: Mechanical Turk; put up or shut up. — Of course this has turned out to be harder than it looked. And though admittedly the training procedure for neural networks bears a family resemblance to the process of conditioning, it is precisely that which to date has rendered it so incredibly inefficient.[55]

{...}

By way of general conclusion: though self-reproducing[56] living organisms are composed of cells, biochemical factories which contain a large but finite number[57] of kinds of molecular machines which function according to the laws of physics and chemistry — and one can, in principle, write down equations of motion for the dynamical system this ensemble represents — even in simplified form these would involve millions of variables, there is no sensible way in which one can suppose they could be solved, and in any case they are, strictly speaking, quantum-mechanical in nature and thus indeterministic; not that intrinsic thermal jiggle does not render the classical problem stochastic anyway. In consequence even when some kind of recognizably mechanical procedure is being implemented, in the operation of an enzyme, e.g., or the reproduction of a strand of DNA, nothing ever works the same way twice; not even the fabrication of the machines

---

[54] Though of course: just because you can build it doesn't mean you can predict what it will do.

[55] Neural networks are trained on sample sets which number in millions, billions, or even trillions. Human infancy does not last a thousand years. Therefore, etc.

[56] Viruses are simpler, but must hijack the machinery of cellular organisms to reproduce.

[57] Counting genes, I would guess between ten and a hundred thousand. This is probably low.

themselves. — Moreover this is not some kind of regrettable design flaw which would be eliminated in a more perfect world — as designed by Plato/Skinner/Minsky/... — this is precisely what made life possible in the first place. (It is also what renders biological design so robust.)

So in the sense that disturbs us — that mechanism is something which does the same thing the same way every time that it functions — biological machinery is not machinery at all. — Indeed to think that it is, or even that it ought to be, is simply insane. — Life is the product of evolution, and evolution consists precisely in making up rules in order to break them.

Perhaps we should call this the paradox of vitalism, then: that despite being wrong about everything it ends up winning most of the arguments anyway.

So even though there is a philosophical moral to be found here, as usual it looks like a joke.