Hector Zenil (Ed.)

Irreducibility and Computational Equivalence

10 Years After Wolfram's A New Kind of Science

Springer

Table of Contents

IV Fundamental Physics

V The Behavior of Systems and the Notion of Computation

VI Irreducibility and Computational Equivalence

Chapter 23
Free Will and *A New Kind of Science*

Selmer Bringsjord

Department of Cognitive Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI), USA
Selmer.Bringsjord@gmail.com

Abstract. What does Wolfram's new kind of science (nks_w) imply about the decidedly *non*-new topic of free will versus determinism? I answer this question herein. More specifically, I point out that Wolfram's nks_w-based position on free will is centered on the nature of physical laws, rather than formal logic; briefly rehearse the longstanding ontology of main positions on free will versus determinism in the context of physical laws; after a more detailed look at Wolfram's position, register agreement with him that in light of nks_w, the *belief* that free will is real and robust is to be expected, and is rational; but explain that nks_w provides no cogent rationalist basis for believing that we are *in fact* free. I conclude by pointing out that in light of the foregoing, if we are free, and can know that we are on the strength of what rationalists demand (e.g., proof), nks_w, while perhaps truly new, is truly incomplete. In short, I show that Wolfram, on free will, is epistemologically insightful, but metaphysically deficient.

1 Introduction

What does Wolfram's (2002) [11] new kind of science (nks_w) imply about the decidedly *non*-new topic of free will versus determinism? I answer this question herein. More specifically, I begin by briefly explaining that Wolfram's nks_w-based position on free will is centered in physics and agentless computation, not formal logic (§2); rapidly rehearse the immemorial, main positions on free will versus determinism in connection with physical law (§3); sketch out in section 4 what it takes to provide a rationalist basis for a position on free will (or on any subject, for that matter); after a more detailed look at Wolfram's position (§5), register agreement with him that in light of nks_w, the *belief* that free will is real and robust is to be expected, and is quite rational (§6); but explain that nks_w provides no cogent rationalist basis for believing that we are *in fact* free (§7). I conclude (in §8) by pointing out that in light of the foregoing, if we are free, and can know that we are on the strength of what rationalists demand (e.g., proof), nks_w, while perhaps truly new, is truly incomplete. In short, I show that Wolfram, on free will, is epistemologically right, but metaphysically deficient.

H. Zenil (Ed.): *Irreducibility and Computational Equivalence*, ECC 2, pp. 341–350.
DOI: 10.1007/978-3-642-35482-3_23

2 Wolfram on Free Will: A Physics-Based Orientation

Wolfram writes:

> Ever since antiquity it has been a great mystery how the universe can follow definite laws while we as humans still often manage to make decisions about how to act in ways that **seem** quite free of obvious laws. (Wolfram (2002) 750; bolded text here and in quotes hereafter due to me, to serve subsequent purposes)

Here Wolfram is pointing to a version of the free-will problem that involves physical laws and causation. There are other versions of the problem that are more abstract, and which steer clear of physical laws in favor of *a priori* reflection from the armchair (or its correlate in ancient Greece), and some of these were also discussed in the distant past that Wolfram points to. A seminal example is Aristotle's famous consideration of a future sea battle, given in his *De Interpretatione*, Chapter 9 (which can be found in [10]). Aristotle reflects on whether *tertium non datur* (TND) holds, and brings that issue into focus by asking whether TND holds with respect to

1. There will be a sea-battle tomorrow.
2. There will not be a sea-battle tomorrow.

Aristotle's reasoning, which we needn't assess, runs essentially as follows. If we assume for the sake of argument that 1. is true, then clearly the proposition expressed by this statement was true a week back, and a month back, and indeed 10,000 years back, *ad indefinitum*. But this is to say that it has always been the case that there will be a sea-battle tomorrow — and hence it immediately follows that all those human actions commonly associated with fighting a sea-battle (including the decision to launch an attack in the first place), commonly regarded to be up to us (= free), aren't. Exactly parallel reasoning can be carried out if the starting point is the assumption that 2. holds.

Aristotle's discussion falls under the topic of logic and the mind, definitely not, for instance, *physics* and the mind, and certainly nks_w falls into the latter domain. (Aristotle is read by some scholars as recommending rejection of TND in favor of a three-valued logic, an idea that certainly had legs: some contemporary extensional logics, e.g. the heterogeneous logic underlying Barwise and Etchemendy's [1] Hyperproof system, add to TRUE and FALSE such values as UNKNOWN.) After all, the central-to-nks_w doctrine of computational irreducibility ranges over the behavior, through time, of *physical* processes.

We turn now to consideration of the free-will problem from the standpoint not of armchair reflection and abstract logic, but physical, or natural, laws.

3 The Ontology of Free Will vs. Determinism

The classical expression of the "physics-relevant" "free will problem" is given by Chisholm's [7]. Encapsulated, the problem as portrayed by him is as follows; we shall call it 'The Dilemma.'

The Dilemma

(1) If determinism is true, then free will is an illusion; and yet on the other hand, (2) if *in*determinism is true, free will is an illusion. But since (3) either determinism or indeterminism is true, it follows that (4) free will is indeed chimerical.

To ease exposition, let's use '$\mathcal{D}$' to denote determinism, and '$\mathcal{I}$' to denote indeterminism. Indeterminism is understood to simply be the negation of determinism; that is, $\mathcal{I}$ if and only if not-$\mathcal{D}$. We thus see that (3) is an instance of a theorem in elementary deductive logic (viz., TND in either the propositional or predicate calculi; perhaps Boole never considered the sea-battle!), and is hence unassailable.

But we can be clearer. For Chisholm:[1]

$\mathcal{D}$: Every event is caused (by the conjunction of physical laws and prior and simultaneous events).

Hence, by elementary quantifier reasoning from the negation of determinism, we have:

$\mathcal{I}$: At least one event is uncaused.

There can be no doubt that the reasoning in The Dilemma is formally valid; indeed, an obvious symbolization in the propositional calculus, and a formal proof, effortlessly obtained, would quickly confirm that {(1), (2), (3)} deductively entails (4). In addition, given even garden-variety accounts of event-causation, it's not hard to see that both (1) and (2) in The Dilemma are quite plausible.[2]

Take (1) first, and understand an event e to be caused just in case prior events, combined with the relevant laws of nature, logically necessitate that e occurs. Suppose now that e is an event that many would regard to be a strong candidate for something humans freely bring about; for example, *Smith's raising his hand to signal the launching of a battle*. Suppose that this event happens at t_n, and that $\mathcal{D}$ is true. Then, given events holding before t_n, at t_{n-1} let's say, and laws of nature that are of course completely beyond the control of Smith, it's logically necessary at t_{n-1} that Smith raise his hand to vote. Since this reasoning can be iterated indefinitely, we will reach a snapshot of the universe at a time t^* eons before Smith's existence which is such that, long into the future from that timepoint, Smith absolutely must send the signal he does at t_n. This fact is

[1] And for others also seeking a rigorous statement of the free will problem; e.g., for Zimmerman's [12] and Bringsjord's [2].

[2] In a Newtonian framework, e.g., (1) and (2) are provable on axiomatizations of Newtonian mechanics. It's beyond scope for the present chapter to discuss the status of such formalizations, or formalizations of (1) and (2) in, say, quantum-mechanical frameworks. Both (1) and (2) do seem quite plausible on Wolfram's physico-computational framework.

inconsistent with the proposition that it's up to Smith as to whether he raises his hand or not, under any reasonable understanding of up-to-us-ness.

Now, the ontology of the immemorial free-will debate is derived from stances on the truth or falsity of (1), (2), $\mathcal{D}$, and $\mathcal{I}$, and runs as follows:

The Ontology of the Free-Will Debate

Incompatibilism (1): $\mathcal{D}$ is compatible with our having free will; i.e., the reasoning given above in favor of (1) is regarded to be compelling.

Compatibilism not-(1): $\mathcal{D}$ is compatible with our having free will.

Hard Determinism : $\mathcal{D}$ conjoined with incompatibilism.

Soft Determinism : $\mathcal{D}$ conjoined with compatibilism.

Not that it matters for the present essay, but Chisholm was a libertarian, as am I. A defense of libertarianism is provided in [2]. In general, it's safe to say that incompatibilism is aligned with the common-sense and ubiquitous laic notion, unabashedly affirmed herein, that the concept of up-to-us-ness is at the heart of what it means to be free. If free will consists in our ability to perform actions that are entirely up to us, then compatibilism, which must accept that free will requires only that we do what we want to do, doesn't seem to be tenable. This is so because if our desires were pre-programmed into us by some other agent, our acting in accord with our desires wouldn't be up to us, but rather up to that other agent.

4 Rationalism Encapsulated

We turn now again to Chisholm, who has provided a discrete continuum of epistemic "strength" [8]. Chisholm's spectrum of the strength of a proposition for a rational human mind is a nine-point one, and ranges from 'certainly false' at the negative end, to 'certain' at the positive end. At the halfway point are propositions said to be *counterbalanced.* There are then four positive strength factors working up from there: first *probable*, then *beyond reasonable doubt*, then *evident*, and finally the aforementioned *certain.* Certain propositions include the indubitable truths of formal logic (e.g., *modus ponens*, $0 \neq 1$, Peano Arithmetic, etc.), and presumably "Cartesian" truths such as "I exist," and "It seems to me that I'm sad." What kind of thing is evident? For the most part, the evident would be populated by those propositions we affirm on the strength of direct sense perception. For example, that there is a computer screen in front of me when I'm typing out a sentence such as the present one is evident. This proposition isn't certain: you might be hallucinating, after all; but it's — as we might say — *close* to certain. You wouldn't want to say, for example, while spying a coffee cup in front of you, in perfect health and having not ingested recently any mind-altering drugs ..., that the proposition that there's a cup in front of you is merely beyond reasonable doubt: you want to say, instead, that you are well within your epistemic "rights" in holding that it's *extremely* likely that there's a cup before you. This, again, is the category of the evident.

But moving down another Chisholmian notch in strength, we do in fact hit *beyond reasonable doubt* — which of course famously coincides roughly with what it takes in certain legal systems (e.g., that of the U.S.) to legally convict someone of murder. That is, to convict someone of this kind of crime, the evidence must make some such proposition as *Jones is guilty* beyond reasonable doubt. Finally, note that to convict on this standard, it's not sufficient to know that it's merely *probable* that Jones did it. Some proposition P being probable is the last notch before we reach *counterbalanced*, which as you've no doubt anticipated entails that a purely rational agent wouldn't bet in favor of P, and wouldn't bet against it. A perfectly rational agent who is agnostic about some proposition P would regard P to be counterbalanced.[3]

Armed with Chisholm's spectrum, we can now offer a tolerably clear encapsulation of the rationalist standard for belief in positions on free will:

> **Rationalism** The view that belief in weighty, philosophical proposition P must be supported by deductive proofs or arguments, where the inferences in this reasoning are each formally valid, and the premises are at least probable.

This doctrine can be partitioned into at least a *strong*, *moderate*, and *weak* subforms. *Strong rationalism* is the view (and as it happens, *my* view) that any human person believing some weighty, philosophical P ought to have on hand at least one outright proof of P; that is, have on hand a formally valid chain of deductive inference originating from premises that are each certain.[4] The doctrine of *moderate rationalism* holds that if Jones abides by this doctrine and believes P, then Jones must have on hand at least one formally valid argument for P whose premises $P_1, P_2, \ldots, P_n$ are each at least evident. And following suit we can say that *weak rationalism* requires only that the premises involved in deductive reasoning for the P in question are at least probable. Readers will no doubt get the driving idea from the foregoing; the story would continue on, all the way through an exceedingly fine-grained ontology of rationalism.[5]

---

[3] What about the "negative" side of Chisholm's continuum? Since neither the empiricist nor the rationalist, if abiding by their respective programs for belief fixation, would assent to propositions on the negative side of *counterbalanced*, we have no need here to explore this epistemic terrain. Interested readers can consult [8], and a recent "AI-ish" exploitation of Chisholm's framework in [5].

[4] Some readers will inevitably ask: "Is there any such thing?!" I'm well aware of the fact that even some axioms in some axiomatic set theories are controversial, and hence perhaps not certain. (Even the power-set axiom in ZFC has its detractors, e.g.) Nonetheless, whatever one can deduce in deductively valid fashion from, say, $1 = 1$, would be certain, and one would be well-advised to believe such a consequence. For instance, $1 = 1 \vee Q$, for any proposition Q, would be an acceptable disjunction for even a strong rationalist to believe.

[5] For example, we could distinguish between the strength of inferential links in the argument for P.

5 Wolfram on Free Will: A More Careful Look

It's now time to look in more detail at Wolfram's treatment of free will in *A New Kind of Science*. To do so, let's pick up right after the short quote from this book presented earlier in section 2. We read:

> [F]rom the discoveries in this book it finally now seems possible to give an explanation for [how the universe can follow definite laws while we as humans still often manage to make decisions ... in ways that **seem** quite free of obvious laws]. And the key, I believe, is the phenomenon of computational irreducibility. For what this phenomenon implies is that even though a system may follow definite underlying laws its overall behavior can still have aspects that fundamentally cannot be described by reasonable laws. For if the evolution of the system corresponds to an irreducible computation then this means that the only way to work out how the system will behave is essentially to perform this computation—with the result that there can fundamentally be no laws that allow one to work out the behavior more directly. And it is this, I believe, that is the ultimate origin of the **apparent** freedom of human will. For even though all the components of our brains presumably follow definite laws, I strongly suspect that their overall behavior corresponds to an irreducible computation whose outcome can never in effect be found by reasonable laws. (Wolfram (2002) 750; bolded text due to me, to serve subsequent purposes)

We can quickly erect a modicum of logico-computational machinery to demonstrate that Wolfram here is entirely correct.

Consider two human persons, Alice and Bob (M_b). We'll assume that M_b is a deterministic Turing machine (TM) based on the binary alphabet $\{0,1\}$ and having two one-way tapes t_1 and t_2, one read/write head operating on each.[6] Tape t_1 enables perception for M_b: a symbol appearing on t_1, and read, indicates that that symbol is perceived by M_b. The other tape, t_2, is used for "internal thinking" on the part of Bob. To further fix our context, we assume that Bob's life unfolds in discrete time steps

$$t_1, t_2, t_3, \ldots$$

into the future, in accordance with the following pattern: Bob thinks for four steps, then perceives (either 0 or 1, and the head on t_1 then moves one square to the right, and awaits the next datum from the external world) in one time step, and then four in a row for thinking, and so on *ad indefinitum*. We write 'C_i^M' to

[6] Wolfram is generally fond of depicting cellular automata rather than TMs (though he does spend appreciable time on Register machines), and indeed he gives a fascinating example of an "unpredictable" one in his principal discussion of free will: see the graphic on p. 750 (2002). But no loss of generality or insight results from restricting our attention to TMs. In addition, while I claim to have proved that human persons can't possibly be TMs, or indeed anything of the sort (e.g., see [3, 4]), for the sake of exposition and argument we here ignore such reasoning, which makes the identification of Bob with M_b palatable.

refer to a *configuration* of TM M.[7] Consider an equation schema $\mathbf{E}$ designed to yield a configuration of Bob for any timepoint given as input; that is, consider:

$$f(t_k) = C_k,$$

where the function f provides the "meat" of this equation, and is itself a Turing-computable function.[8]

We are now in position to see that Wolfram, in the quote immediately above, is right.

6 Wolfram Is Correct — Epistemologically

Suppose that we are interested in whether Alice believes Bob to have free will. Not unreasonably, we shall stipulate the following epistemic principle $\mathcal{E}$: a sufficient condition for such a belief on the part of x about TM y is that despite x's having complete knowledge about the transition rules that determine the state of y at t_k given the state of y at t_{k-1}, the "overall behavior" of y cannot be anticipated by x. More precisely, we stipulate that despite knowledge of transition rules, x does not, indeed *cannot*, predict, on the strength of an equation of the form of $\mathbf{E}$, the configuration that y will be in for some future timepoint. Next, we shall agree with Wolfram that if x is in this position of ignorance about the future states of y, then x will ascribe free will to y; that is, x will believe that y has free will.

We can now prove that Wolfram is right with respect to Bob, as long as we assume that Bob, *qua* TM, is for instance as complex as the impenetrable, unpredictable 6-state machines which have never been predictable within the confines of the Busy-Beaver Problem.[9] Needless to say, Bob's mind is unquestionably more complex that such TMs! The proof is trivial once we realize that the Wolframian setup we have established implies that the following proposition is now an easy lemma: $\neg\exists f f(t_k) = C_k^{M_b}$,

From this lemma it follows directly by *modus ponens* on $\mathcal{E}$ that Alice believes that Bob has free will. Since we have here fleshed out computational irreducibility with respect to Bob, Wolfram's reasoning is certified. Moreover, his reasoning, given the account supplied above, is without question rationalist in nature — since the inferences are deductively valid, and all premises appear to be at least evident. In particular, given the framework set out in section 4, we can declare that Wolfram has provided a case for the belief in free will that accords with the standards of moderate rationalism.

[7] The concept of a *snapshot* or *configuration* is standard in presentations of TMs. E.g., see [9].

[8] I have shown that human mentation includes information-processing more powerful than what a TM can reach [6], but I leave this aside in the present essay.

[9] Wolfram provides an elegant, succinct description of the Busy-Beaver Problem: ((2002) 889 & 1144).

7 Wolfram Is Wrong — Metaphysically

But there is a hitch here, a very serious one. In general, the hitch is that it doesn't follow from the fact that x believes some proposition that that proposition is true. Some humans still believe that Earth is flat, after all. But how does this specifically relate to the case at hand? If you look back to all the bolded parts of the quotes from Wolfram's *ANKS*, you'll see, clear as day, that Wolfram has proffered only an explanation for why humans, in general, *believe* that they have free will. For example, we earlier saw this:

> And it is this, I believe, that is the ultimate origin of the **apparent** freedom of human will. For even though all the components of our brains presumably follow definite laws, I strongly suspect that their overall behavior corresponds to an irreducible computation whose outcome can never in effect be found by reasonable laws. (Wolfram (2002) 750; again, bolded text due to me, to serve present purposes)

This is just one example from many that I've pinpointed via bolded text, but the situation, especially given the many other bolded words, should be clear as day. If someone's will is *apparently* free, it hardly follows that that will is *in fact* free. Nowhere in *ANKS* does Wolfram even intimate that he maintains that our decisions are in fact free.

But what we are ultimately concerned with is whether, *in fact*, at least some of our decisions are truly up to us. On this issue, which is the real one, Wolfram is deafeningly silent. Moreover, it would seem to be implausible that free will, or up-to-usness, is in fact in place in the universe as conceptualized under nks_w. Why?

Well, think back to Alice and Bob. But we are now concerned not with whether Alice, under reasonable physico-computational and epistemic assumptions, believes that Bob $= M_b$ has free will; rather, we are interested in whether or not Bob is *in fact* free. We have only two general factors that are relevant to this question. And neither factor is of help to Wolfram on the question before us.

To see this, consider first the first factor: the one that served to support Alice's belief that Bob is free: namely, that relevant instantiations of equation **E** for Bob's future behavior are simply unavailable. But the unavailability of equations of this form in no way rationalistically entails that Bob in fact is free. For just because we can't predict, for some future timepoint, what state Bob will be in at it, doesn't ensure that the state he is in at this timepoint is due to the free operation of his own will. You may be unable to predict what general configuration a puppet will be in to start the next act of a puppet show (because, among other reasons, you are unfamiliar with the relevant choreography), but it hardly follows from this inability that a puppet has free will.

And now what is the second factor? It's that by all accounts the Wolframian world-view seems to be inconsistent with up-to-us-ness. Notice that I don't *assert* this inconsistency; I claim only that there *seems* to be outright inconsistency. The reason for my claim can be seen by turning yet again to the Alice-Bob scenario;

specifically, to the fact that while Alice can't predict what Bob's configuration will be in at an arbitrary future timepoint, Alice *can* predict the configuration Bob will be in at t_n if she knows the configuration he's in at t_{n-1} (since, as we legislated, she knows Bob's transition rules). In short, nothing seems to be up to Bob whatsoever: the state he is in at any given moment appears to be entirely necessitated by the co-operation of transition rules (which are of course directly analogous to causation in The Dilemma) and the input to them (i.e., the configuration at the moment immediately preceding). In sum, The Dilemma could be recast within the Wolframian pan-computational framework, and would thereby lose none of its original force; in fact it would *gain* in force.

8 Conclusion

To sum up, the situation is clear: A rational person (i.e., for us herein, a rationalist), having open-mindedly studied *A New Kind of Science*, and assumed to have an understanding of the longstanding ontology of the free-will debate, including specifically The Dilemma, will not be enlightened as to what the solution to that dilemma is. However, on the bright side, Wolfram can be credited for his commendable scholarship, for the elegance of his discussion of free will, and for a compelling argument in support of the proposition that if our world is indeed computationally irreducible, human persons, when assumed to have sufficiently refined cognitive capacities for perception and reasoning, will indeed believe themselves to be free. The downside is that if we are free, and can know that we are on the strength of supporting argumentation and/or proof of the sort required by moderate rationalism, it follows that nks_w is at best incomplete, and at worst — if in fact the computationalism in nks_w rules out up-to-us-ness — incorrect.

References

[1] Barwise, J., Etchemendy, J.: Hyperproof. CSLI, Stanford (1994)

[2] Bringsjord, S.: Free Will. In: What Robots Can and Can't Be, pp. 266–327. Kluwer, Dordrecht (1992)

[3] Bringsjord, S.: What Robots Can and Can't Be. Kluwer, Dordrecht (1992)

[4] Bringsjord, S., Arkoudas, K.: The modal argument for hypercomputing minds. Theoretical Computer Science 317, 167–190 (2004)

[5] Bringsjord, S., Taylor, J., Shilliday, A., Clark, M., Arkoudas, K.: Slate: An Argument- Centered Intelligent Assistant to Human Reasoners. In: Grasso, F., Green, N., Kibble, R., Reed, C. (eds.) Proceedings of the 8th. International Workshop on Computational Models of Natural Argument, (CMNA 2008), Patras, Greece, pp. 1–10 (2008), http://kryten.mm.rpi.edu/Bringsjord_etal_Slate_cmna_crc_061708.pdf

[6] Bringsjord, S., Zenzen, M.: Superminds: People Harness Hypercomputation, and More. Kluwer Academic Publishers, Dordrecht (2003)

[7] Chisholm, R.: Freedom and Action. In: Lehrer, K. (ed.) Freedom and Determinism, pp. 11–44. Random House, New York (1964)
[8] Chisholm, R.: Theory of Knowledge, 2nd edn. Prentice-Hall, Englewood Cliffs (1977)
[9] Lewis, H., Papadimitriou, C.: Elements of the Theory of Computation. Prentice Hall (1981)
[10] McKeon, R. (ed.): The Basic Works of Aristotle. Random House, New York (1941)
[11] Wolfram, S.: A New Kind of Science. Wolfram Media, Champaign (2002)
[12] Zimmerman, M.: An Essay on Human Action. P. Lang, New York (1984)