Explaining the Big Bang

Alastair Wilson
a.j.j.wilson@leeds.ac.uk

School of Philosophy, Religion and History of Science
University of Leeds
Woodhouse, Leeds LS2 9JT, UK

School of Philosophical, Historical and International Studies
Monash University
VIC 3800, Australia

9th April 2024

Abstract: The big bang is the cause of every event in our universe, and hence it explains all subsequent cosmic history. But can we explain the big bang itself? This chapter explores a number of different styles of explanation that might be offered. These include causal explanations of the big bang, either by a physical or a non-physical cause: here I focus especially on Roger Penrose's conformal cyclic cosmology. They also include non-causal explanations of the big bang in terms of an underlying physical or non-physical fundamental basis for the universe, or in terms of fundamental physical laws.

1. Introduction

Current cosmology – in the form of the Lambda-CDM model – tells us that the universe began with a big bang. (Lambda stands for the cosmological constant, also known as 'dark energy'; CDM stands for 'cold dark matter'.) This model displays an excellent match with observed distributions of matter and cosmic background radiation. It tells us that the big bang initiated a causal chain, a cascading process which led to space and time as we know them, and (eventually) to the creation of matter in all its curious patterns. The big bang is the causal ancestor of all subsequent events: every spacetime trajectory eventually reaches back to approach it (Figure 1). In that sense the big bang explains everything that has happened since. But can we explain the big bang itself?

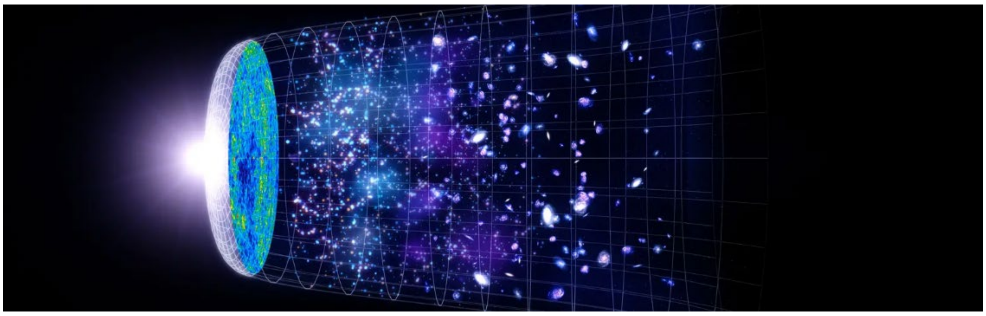

Figure 1: All spacetime trajectories approach the big bang.

Image: NASA.

This chapter is about some of the different possible kinds of explanation we can give for the big bang. I will contrast causal explanations of various kinds, based on different candidate physical laws, with non-causal explanations based on principles which are at least partly metaphysical in character. While causal explanations of the big bang require new physics – as we shall see, models of this kind have been offered by various theorists – non-causal explanations of the big bang may require more in the way of new philosophy, or at least some reconceptualizing of our existing physics. It must be noted that our current state of knowledge in early-universe cosmology remains radically incomplete. There is no theoretical consensus on physics below the Planck energy scale, where quantum-gravitational effects start to dominate. Without working in the context of some specific approach or other to quantum gravity, however, very little can be said that has general applicability. Much of what follows therefore discusses physics which is itself highly speculative. For present purposes that is no obstacle, indeed it is part of the plan; my aim here is to survey and discuss the variety of shapes that a candidate explanation of the big bang might take.

2. Kinds of Physical Explanation

Let us start by distinguishing two kinds of explanation: causal and non-causal. For the purposes of this chapter, it will help to treat this distinction as exhaustive, at the cost of potentially making the category of non-causal explanations a disunified one.

I will operate with a broad conception of causal explanation as (roughly) *law-governed explanation*. Causal explanations in physics will always then involve the application of some physical law(s): consider as canonical examples the explanations of planetary orbits in terms of a centripetal gravitational force; and explanations of resultant current in terms of applied potential difference and resistance in a circuit. The Earth's gravity causes the moon to orbit, and I can cause my torch to light up by putting a battery in it.

The category of non-causal explanation covers all and any other kinds of explanation. Canonical examples of non-causal explanations in physics are explanations of temperature and pressure of a gas in terms of the collective motion of its molecules and explanations of the stability of matter in terms of the Pauli exclusion principle. The gas has heated up because the the molecules in the gas are now moving faster; the exclusion principle explains why electrons form nested orbitals in atoms (Figure 2). In cases like these, our ability to model the higher-level theory in terms of the lower-level theory through partial theoretical reduction is typically what enables us to explain the higher-level phenomenon in terms of the lower-level phenomenon. We account for the thermodynamics of gases in terms of statistical particle mechanics, and we can account for all the key features of stable atomic structures in terms of quantum electrodynamics. (For discussion of theoretical reduction and how it relates to non-causal explanation, see [1].)

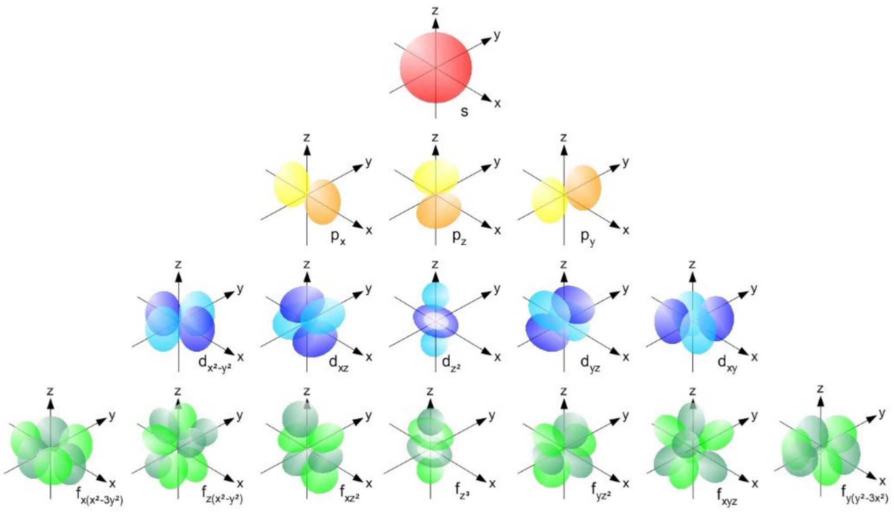

Figure 2: The exclusion principles restricts how many electrons can occupy an energy state, giving nested orbitals. Image: Wikimedia Commons.

To draw the distinction between causal and non-causal explanation more precisely, we can appeal to the distinction between dependencies which result from the operation of physical laws (causal explanations) and dependencies which result directly from the constitution of physical systems (non-causal explanations). If an explanation is driven by an explanatory principle which is a law of physics, it is a causal explanation; if it is driven by some more general metaphysical or logical linking principle, then it is a non-causal explanation. This way of drawing the distinction makes sense of the way we think about the examples above, and it generalizes in a natural way to the context of theories of quantum gravity [2].

In contemporary theories of quantum gravity, space and time are often seen as emergent – like the waves on the surface of a deep ocean. What we experience as space and time is the product of quantum processes operating at a deeper level, and we can't any longer rely on our ordinary conception of distinct events playing out against a 3+1-dimensional spacetime background. If causation were tied to events situated in ordinary space and time, then there could be no causation in quantum gravity. Nonetheless, our question 'why the big bang?' can still be coherently asked at the quantum gravity level, as we shall see.

3. Causal Explanations of the Big Bang

How might we give a causal explanation of the big bang? If we were purely focused on explaining where matter comes from, then the challenge of explaining the origins of the universe might be easier than it is. David Albert [3] memorably criticized accounts of the big bang such as that of Lawrence Krauss [4] which promise to get something from nothing, but which do so by understanding 'nothing' in terms of a spacetime 'vacuum state'. Quantum field theory tells us that even the vacuum state, supposedly corresponding to empty spacetime, is still full of physical activity. The spacetime vacuum state is seething with particles constantly being created and destroyed. As a first cause, it is very much a something rather than a nothing.

Suppose we are not satisfied with an explanation of the origin of matter which presupposes the existence of spacetime in a vacuum state. Suppose instead we ask: where did spacetime itself arise from? To answer this question, we need to look back into the Planck epoch and consider specific theories of quantum gravity. Here our familiar categories of cause and effect come under pressure, since theories of quantum gravity may describe physical processes which do not unfold in space or over time [5]. This is where we start to see the value of our approach to distinguishing causal from non-causal explanation in terms of the principles which mediate the explanation. Traditional approaches to understanding causal explanation, which link causal explanation to the determination over time of systems in space by other systems in space, struggle to handle causal chains which reach back into a physical epoch prior to space itself or time itself. Once we instead generalise our conception of causality so that dependencies count as causal if they are described by laws of nature, including pre-spatiotemporal laws, then we can trace causal chains back into the Planck epoch, resulting in an ordering of events which lacks the geometric features of spacetime but which nevertheless sustains a causal ordering. Henrique Gomes has described a scenario of this kind, called the many-instant landscape [6]. Figure 3 shows how records of the past can exist in such scenarios even in the absence of a completely ordered sequence of instants.

Fig. 3 The types of sequence of records that can arise. The first gives rise to a notion of a granular history. The second and third represent a more general version of records of records. The fourth represents the sort of records which are not consistent, as their ordering disagrees

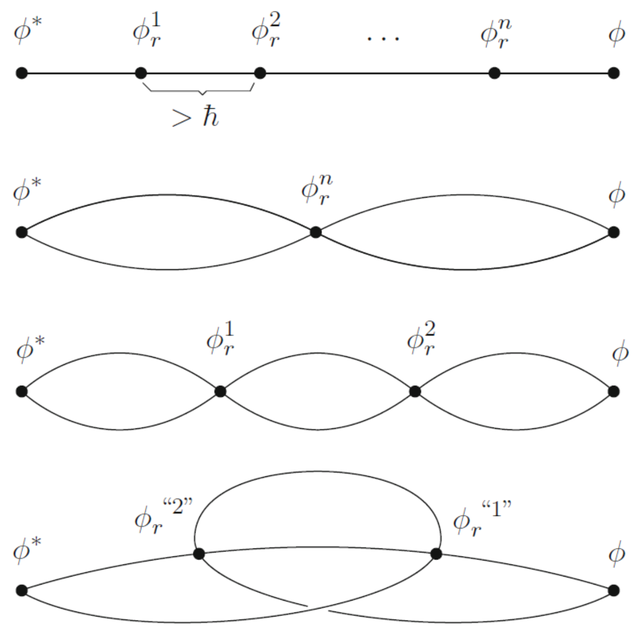

Figure 3: The many-instant landscape, in which causal ordering and temporal ordering can come apart. Image: Henrique Gomes, reproduced from [6].

However, tracing our causal chain back to the domain of quantum gravity in the end only postpones answering our question. All candidate theories of quantum gravity describe something physical that was going on in the Planck epoch, some quantum precursor of ordinary space and time – so where did that come from? Consider, then, the quantum state of the universe at the big bang, and ask what explains it. There are two options: either there is a causal explanation or there is no causal explanation. Amongst causal-explanation options, we can distinguish *creation explanations* and *recurrence explanations*. Amongst recurrence explanations, we can further distinguish *linear* recurrence explanations and *circular* recurrence explanations.

Consider first creation explanations. Here the general idea is familiar – although the big bang might be the first physical cause, it is not the first cause if it is caused by some supernatural cause – usually, a divine designer. These explanations go beyond the limits of what counts as physics, but that is no reason by itself to reject them out of hand. However, a more telling problem is that none of the explanations seem to me to have good epistemic standing by the standards usually applied within physics. In particular, they do not explain the less probable in terms of the more probable. In explaining an initial quantum state such as the one we observe in terms of a creator, we immediately raise the question of why there is a creator of the kind which would create an initial quantum state of that kind. This style of objection is developed in detail by Graham Oppy [7].

Let us set this sort of objection aside for the sake of argument, and suppose that we have independent evidence for a designer who would be inclined to create an initial quantum state like our own. It seems rather doubtful, however, that any of the current features of our universe are what you would expect from any independently-motivated designer. Quite apart from the extensive suffering in sentient creatures we know of, there is no reason to think that our physical universe is optimized for any plausible divine plan. Indeed, the universe doesn't seem on the face of it to be optimized for anything in particular. Although there is life, there isn't a maximal amount, or sophistication, or variety of life: we see a lot of empty space, many barren planets.

I will say no more about design explanations. Consider next recurrence explanations, of the linear kind. Here a prominent recent example is Penrose's conformal cyclic cosmology. Penrose was inspired by an interesting mathematical correspondence between a very hot, dense, small state of the universe (as it was at the big bang) and an extremely cold, empty, expanded state of the universe (as it will be in the far future). His radical theory to explain this correspondence is that those states become literally identical when taken to their limits: that the far-future limit of an old and empty universe is at the very same time the birth of a young and busy universe. How can the very same state be a cold empty universe from one perspective and a hot dense universe from another? The answer lies in the notion of a 'conformal rescaling', a geometrical transformation which in effect alters the size of an object but leaves its shape unchanged. Penrose showed how the cold dense state and the hot dense state could be related by a conformal rescaling, so that they match with respect to the shapes of their spacetimes (although not to their sizes). This led to his bold hypothesis: the hot dense state and the cold empty state coincide. The death of one universe, when we 'zoom out' far enough, is revealed as the seed of a new big bang. (Figure 4).

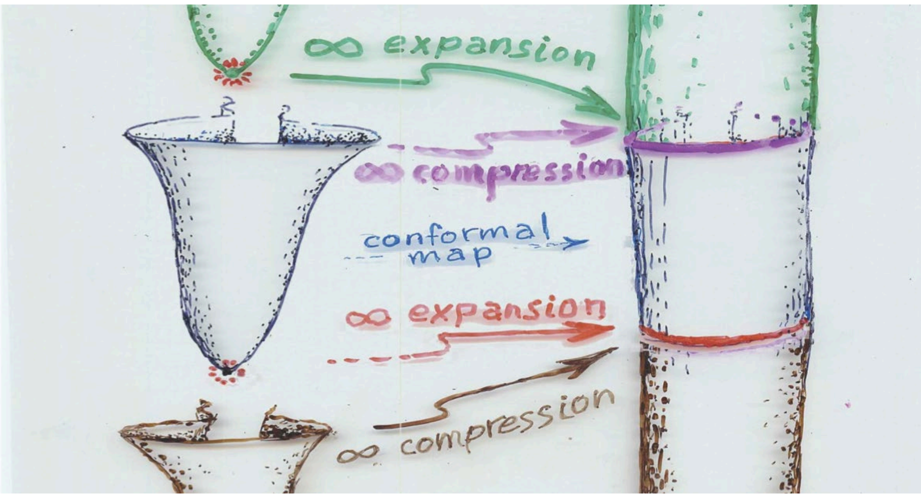

Figure 4: In conformal cyclic cosmology, the death of one spacetime is the birth of another.

Image: Roger Penrose.

In conformal cyclic cosmology, the direction of explanation goes from old and cold to young and hot: the hot dense state exists because of the cold empty state. But this 'because' is not a cause followed in time by its effect. The cold dense state and the hot dense state are in effect located on different timelines, whereas causality is usually understood as internal to a single timeline. The cold empty state continues on forever from the perspective of its own temporal geometry, but the hot dense state to which it gives rise has a new timeline all its own.

Does conformal cyclic cosmology explain why there is something rather than nothing? The big bang arises from an almost-nothing: the almost-nothing that is left over when all the matter in a universe has been consumed into black holes, when all the black holes have boiled away into photons, and when those photons become lost in a void. The whole universe thus arises from something that – viewed from another physical perspective – is as close as one can get to nothing at all. But that nothing is still a kind of something. It is still a physical universe, however empty. We therefore avoid the troublesome idea that something ever completely comes from nothing.

In Penrose's vision, there are endlessly repeating cycles, with the initial quantum state of each universe explained by some feature of the universe before. For him, each cycle involves quantum random events turning out a different way. Then our big bang would have been caused by the death of a previous universe, and we might even be able to glimpse it through faint traces in the left-over radiation seen by the Planck satellite. Penrose and his collaborators believe they may have spotted these traces already [8, 9]. But we can also consider a one-cycle form of conformal cyclic cosmology, with the beginning of that one cycle explained by some feature of its own end. Then physical reality consists in a single cycling around through the big bang to a maximally empty state in the far future – and then around again to the very same big bang.

Recurrence explanations for the big bang come in two forms, then: cyclic cosmologies made up of an endless chain of new cycles, and cyclic cosmologies made up of a single repeating cycle. What makes the difference is whether the universes are one and the same, with one and the same set of inhabitants – or whether they are distinct universes with distinct inhabitants (no matter how closely they might resemble one another).

To see how a circular recurrence explanation might play out, we could ally conformal cyclic cosmology with the many-worlds interpretation of quantum theory, and locate all the quantum parallel universes within one single cycle. No matter how small the chance of something occurring, if it has a non-zero chance then it occurs in some quantum parallel world. There are people just like you in other worlds who have won the lottery, or have been swept up into the clouds by a freak typhoon, or have spontaneously ignited, or have done all three simultaneously.

In a many-worlds setting, then, we can envisage a new twist on conformal cyclic cosmology, (although not one which Penrose himself would endorse): the big bang might be the rebirth of one single quantum multiverse, cycled through over and over. Everything possible happens – then it all happens again and again and again. This many-worlds twist simplifies the ontology of conformal cyclic cosmology. We might now be able to make do with just one single cycle, since all the possibilities physics permits could be captured within that cycle. However, the price is that we would lose any distinctive observable consequences of the previous cycles, since all the information they contained would be 'washed out' in the transition from cold and empty to hot and full. The overall case for the resulting view would therefore have to be an indirect one, involving as much philosophy as physics. But I think that is no reason not to explore it.

4. Non-Causal Explanations of the Big Bang

In this section, our focus turns to prospects for non-causal explanations for the big bang. Suppose that some causal explanation of the big bang goes on to be wildly successful. Some questions would still be left unanswered. For example, if Penrose's vision is vindicated by the future progress of cosmology, we still wouldn't have answered a deeper explanatory question. Why does the whole system of Penrose cycles exist? Thus we finally end up with the pure question of why there is something rather than nothing – a paradigm question of metaphysics.

Non-causal explanations, like causal explanations, come in multiple types. I will discuss *grounding explanations*, *plenitude explanations* and *necessitarian explanations*. These are all types of metaphysical explanation. Mathematical non-causal explanations of the universe are also possible – and have been attempted in a tradition which goes back to the Pythagoreans and beyond – but these are beyond my scope in the present chapter.

Grounding explanations are a matter of anchoring the fact that the universe exists in some other kinds of fact. These explanations tell us what layers of reality exist underneath physical reality. If physical reality is grounded, then physical facts are in general non-fundamental and the entire body of physical fact is dependent on facts of some underlying kind. For those who find the idea of a divine mind plausible, thoughts in that mind would be a natural candidate for this underlying kind of fact. But other kinds of non-physical substrata for grounding a physical universe as a whole – including its beginning in the big bang – are possible. These include mentality (as in phenomenalist proposals, and cosmopsychist proposals like that of Philip Goff [10] and facts about what is good (as in axiarchic proposals like that of John Leslie [11, 12].

Grounding explanations rely on general principles which can be thought of as generating something out of something else. In that sense they have much in common with causal explanations; see [13] for extended discussion of the analogy. However, the general principles driving grounding explanations are metaphysical rather than physical in character; they typically specify the constitution of derivative systems. Normally, this is a matter of subsystems constituting systems, but some have argued that subsystems are instead constituted by the larger systems of which they are parts: this is the thesis of *priority monism*, often defended by appeal to quantum entanglement [14, 15].

A potential mechanism which might drive a grounding explanation of the big bang might be the spacetime-from-entanglement proposal [16, 17]. This theory's description of spacetime emerging from a non-spatiotemporal underlying reality is bona fide physics, but it is nevertheless metaphysical in the sense of specifying the underlying nature of space and time. Instead of explaining how the big bang is caused, they tell us what the nature of the big bang might be: as, so to speak, a spatiotemporalizing of a pattern of entanglement. This idea is illustrated in figure 5.

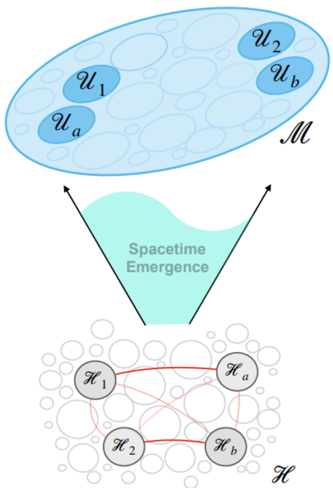

FIG. 1. Representation of how space emerges from quantum mechanics. In the *quantum* Hilbert space $\mathcal{H}$, different quantum sub-systems, $\mathcal{H}_p$, such that $\mathcal{H} = \bigotimes_p \mathcal{H}_p$, are connected by their mutual information, which is larger for darker red lines. In the emergent *classical* manifold $\mathcal{M}$, the relative distance of these sub-systems, $\mathcal{U}_p$, decays monotonically w.r.t. the amount of mutual information they share.

Figure 5: Spacetime as grounded in entanglement.

Image: Franzmann et al. [17].

Plenitude explanations are a very different kind of non-causal explanation of the universe. According to plenitude explanations, the universe exists because all universes of some relevant class containing our universe exist. For example, Lewisian modal realists ([18]; see also [19]) might explain the existence of our universe by reference to the existence of a multiverse of all logically possible structures containing it.

While plenitude explanations may (or may not; the explanatory credentials of subsumption-under-regularities are much debated in the literature on Humeanism) technically answer our question, they don't push the bump in the carpet very far away. Why does our universe exist? – because all universes exist. But why do all universes exist? In addressing this latter question we need a different type of explanation, maybe an axiarchic approach ([11]) or a design explanation ([20]). Or perhaps the existence of the relevant kind of multiverse is instead simply explanatory bedrock: a 'brute fact'.

The final style of explanation I shall consider is necessitarian explanation: the universe exists because it must. There simply are no possibilities in which a universe does not exist, so some universe necessarily exists – and it could just as easily have been this one as any other, so we ought not be surprised. This necessitarian response can be conceived as a positive form of explanation – a truth is explained by its own necessity – but it is perhaps more natural to think that what is happening here is that the explanatory demand is shown to be illegitimate. Because there is no contingency in whether the universe exists, there is no contrast between existing and non-existing possibilities to be drawn, and hence no question of why the contrast resolves one way or another. Then the initial explanatory demand dissolves.

At this point we can distinguish two approaches to necessitarian explanation: the fatalistic and the non-fatalistic approach. On the fatalistic approach, our universe is the only possible universe: things could not have failed to be as they are. On the non-fatalistic approach, our universe is one of multiple possible universes, but some universe or other from that range had to exist.

In one particularly interesting version of the non-fatalistic necessitarian approach, there are a limited range of possibilities, but they all contain a big bang just like ours. Then the best explanation for the big bang could still be in terms of its necessity – it happened because it had to – even if there are multiple open physical possibilities for how the big bang turns out. A limited necessitarian view of this kind is defended in my book *The Nature of Contingency* [21]; all the genuine possibilities for nature correspond to the different worlds in a quantum multiverse.

5. Zero-Explanations of the Big Bang

Consider finally zero-explanations, in which something is explained without there being anything which explains it. It must be said at the outset that this explanatory strategy is of doubtful coherence: perhaps there are simply no explanations of this kind. But a zero-explanation strategy promises to provide explanations of the big bang which are uniquely satisfying, since they disarm the strategy (favoured amongst small children) of simply pressing the question 'why' again to any explanation that might be provided. If something is explained without there being anything which explains it, then there is no explainer there which we can seek to explain in turn.

In the model of zero-explanation I have explored elsewhere [22], zero-explanations are mediated by general explanatory principles, in much the same way as causal explanations and grounding explanations are. If the big bang is zero-explained, then there is some principle which explains how the big bang is explained. The principle itself is not an explainer of the big bang, though – breaking the regress. Of course, we can ask why the principle itself holds. But now we are asking why something is a law of physics, rather than why the universe exists, so we have at least managed to answer all explanations concerning the existence of things.

Causal zero-explanations of the big bang would involve a law of physics that a universe – perhaps a universe with a certain quantum state – should exist. Various proposals of that kind deriving from physics have been considered by Eddy Chen [23], and some toy theories with more metaphysical dynamics have been explored by Bradford Skow [24]; an illustration of the latter is given in Figure 6.

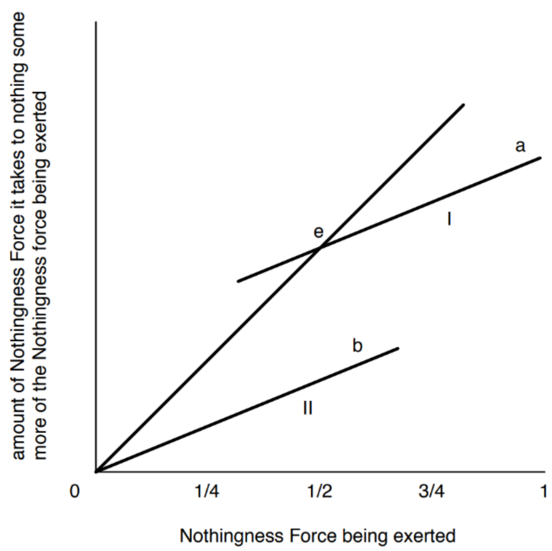

Figure 6: A metaphysical dynamics to explain why there is something rather than nothing. Image: Bradford Skow [24]

By contrast, non-causal zero-explanations would have the same general structure but would be driven by laws of metaphysics. As examples, we can consider the axiarchic principle that matters turn out for the best, or the rationalist principle that things turn out in the least arbitrary way possible.

6. Conclusion

In this chapter have discussed a perhaps-dizzying array of different styles of explanation one might offer for the big bang – while still leaving many explanatory strategies aside. I hope to have shown that there is no impossibility in principle in explaining the big bang, either in physical or metaphysical terms, and I have discussed a variety of serious theoretical proposals which do explain it in some form or other. While some of those proposals generate new explanatory questions of much the same kind, others offer the tantalising promise of leaving nothing unexplained. Time will tell which of these explanatory strategies ultimately gains traction as fundamental physics and cosmology progress.

Acknowledgments

My thanks to Roger Penrose, Katie Robertson and Miranda Rose for feedback; all remaining confusions are of course my own. This work forms part of the project A Framework for Metaphysical Explanation in Physics (FraMEPhys), which received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 757295). Funding was also provided by the Australian Research Council (grant agreement no. DP180100105).

References

[1] Robertson, K. and Wilson, A. (forthcoming). "Theoretical Relicts: Progress, Reduction and Autonomy", *The British Journal for the Philosophy of Science*.

[2] Wilson, A. (2020b). "Classifying Dependencies". *In The Foundation of Reality: Fundamentality, Space and Time*, eds. G. Darby, D. Glick & A. Marmodoro. Oxford: Oxford University Press

[3] Albert, D. (2012). "On the Origin of Everything", *The New York Times*, March 23rd.

[4] Krauss, L. (2012). *A Universe from Nothing: Why There Is Something Rather Than Nothing.* Free Press.

[5] Wilson, A. (2021). "Explanations of and in time". In *Explanation Beyond Spacetime*, eds. C. Wüthrich, B. Le Bihan, N. Huggett. Oxford: Oxford University Press.

[6] Gomes, H. (2018). "Timeless Configuration Space and the Emergence of Classical Behavior", *Foundations of Physics* 48: 668-715.

[7] Oppy, G. (2006). *Arguing about Gods*. Cambridge: Cambridge University Press.

[8] Penrose, R. (2010). *Cycles of Time: An Extraordinary New View of the Universe*. Oxford: Bodley Head.

[9] Penrose, R. (2020). "Black Holes, Cosmology, and Space-Time Singularities", Nobel Prize Lecture.

[10] Goff, P. (2023) *Why? The Purpose of the Universe*. Oxford: Oxford University Press.

[11] Leslie, J. (1970). The Theory That the World Exists Because It Should", *American Philosophical Quarterly* 7(4): 286-298.

[12] Leslie, J. (1989). *Universes*. New York: Routledge.

[13] Wilson, A. (2018). "Metaphysical Causation), *Noûs* 52(4): 723-751.

[14] Schaffer, J. (2010). "Monism: The Priority of the Whole", *Philosophical Review* 119(1): 31-76.

[15] Ismael, J. and Schaffer, J. (2020). "Quantum Holism: Nonseparability as common ground", *Synthese* 197: 4131–4160.

[16] Cao, C., Carroll, S. and Michalakis, S. (2017). "Space from Hilbert space: Recovering geometry from bulk entanglement" *Physical Review D* 95(2): 024031.

[17] Franzmann, G., Jovancic, S., and Lawson, M. (2023). "On the Relative Distance of Entangled Systems in Emergent Spacetime Scenarios", *Physical Review D* 107: 066008.

[18] Lewis, D. (1986). *On the Plurality of Worlds*. New York: Blackwell.

[19] Tegmark, M. (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. Knopf.

[20] Kraay, K. (2010). "Theism, Possible Worlds, and the Multiverse", *Philosophical Studies* 147(3): 355-368.

[21] Wilson, A. (2020a). *The Nature of Contingency: Quantum Physics as Modal Realism*.

[22] Hicks, M. and Wilson, A. (2023). "How Chance Explains", *Noûs* 57(2): 290-315.

[23] Chen, E. (forthcoming). "Strong Determinism", *Philosophers' Imprint*.

[24] Skow, B. (2010). "The Dynamics of Non-Being", *Philosophers' Imprint* 10(1): 1-14.