In other instances of explanatory change, however, ideas come first and must await new technologies to secure their explanatory fit to the world. I call these Copernican Delays, in reference to the seventy years during which Copernicus's concept of heliocentrism was dismissed as, in Galileo's words, a mere "hypothesis" to calculate "motions."[50] It took a new technology, Galileo's "improved spyglass," to ground the hypothesis in the hard to vary and correlated details (Jupiter's moving moons and Earth's tides) that gave it an explanatory fit to the universe it reordered.

With that pattern in mind, we can add a historical perspective to the connection between quantum mechanics and information that I just posed. The initial explanations of the quantum formalisms of the early twentieth century quickly gave way to decades of implementation during which calculation silenced explanation[51]—a delay of Copernican proportions that only ended with what's called the "new quantum age"[52] at the century's close. Like Galileo tracking the Jovian moons in 1610, Deutsch in 1985 followed the flow of information into the physical world, moving computation from Alan Turing's mathematics into physics. The new technology he envisioned enacted a new explanation of reality in which, as I've just discussed, information became newly fundamental: the *lingua franca* of all of our re-makings. This grounding of information in the *physical* performed Bacon's handshake between the intellectual and visible worlds and changed our explanations of both.

INFORMATION IN THE HISTORY OF KNOWLEDGE II: THE SEQUENCE OF SURPRISE

To recap: Companion 1 makes information legible across the ages; Companion 2, as I've just started to outline it, reads information for evidence of advancement—of explanations fitting better and reaching farther. With matrices and delays now in our toolbox of change, it's time for a second reading, but one with an added twist. In this one, we'll pursue not only explanations of information but information about explanation itself. When do we feel particularly compelled to explain? Why? My goal here in connecting specific moments of explanation is to animate the "story of learning" so that we can see ourselves in it, helping to clarify the present and conjecture the future. Let's conclude by attempting both.

Happily, I'm writing this as we face a perfect test case—a moment in which, especially in regard to information, there is much to clarify. To put that another way—with more urgency and less serendipity—we could really use some companionship right now. Without it, change, especially consequential change, can seem exceptional and thus disturbing, causing us to worry the consequences

50. Galileo Galilei, "Reply to Cardinal Bellarmine (1615): Galileo's Considerations on the Copernican Opinion, Part I," in *The Essential Galileo*, ed. and transl. Maurice A. Finocchiaro (Indianapolis: Hackett, 2008), 152.

51. That silencing of explanation has been summed up in the popular phrase "Shut up and calculate." See N. David Mermin, "What's Wrong with This Pillow?," Physics Today 42, no. 4 (1989), 9.

52. See Andrew Whitaker, *The New Quantum Age: From Bell's Theorem to Quantum Computation and Teleportation* (Oxford: Oxford University Press, 2016).

rather than focus on the explanations. The change at issue is the startling rise of transformer-driven Large Language Models (LLMs). ChatGPT and its accompanying wave of generative AI have put a twofold problem of advancement on everyone's table. First comes doubt: Is this new information technology an actual advance or a dead end—an illusion of progress diverting us from real path to AI and AGI (artificial general intelligence)? Second comes fear: If it is an advance, is it an advancement for us or an extinction-level trap for humankind?

These are, of course, in important ways, technical questions, but technologies, as we've just discussed, are not self-explanatory—either of how they work or of what they're good for. And, for these information-based technologies, the explanations turn on what we think information itself is and what it can *do*. That's where the history of knowledge matters. It documents how those assumptions have changed with an eye open to the vectors of advancement. But what prompts those changes in the first place? Can we mine the history of knowledge for a recurrent sequence there as well?

In regard to our current AI moment, for example, can our history recover continuities across moments of advancement that can take the apocalyptic edge off the current debate? Instead of mystifying GPTs for doing more than expected, we could engage them—and the new problems they generate—as the latest episode in an old and ongoing interplay between explanation and emergence, or what I'll call the *sequence of surprise*. With Companion 2 at our side, we won't be so surprised by surprise. If we jump into the debate without it, we, rather than GPTs, hallucinate, mistaking their workings and their output for a threatening singularity rather than a not unfamiliar way forward for the project of knowing the world. Finally, in keeping with Bacon's plan for a history of knowledge that doesn't just index advancement but fosters it, I'll end by going out on a limb. There, I'll switch into a much more speculative mode to try on an explanation for hallucination that might just be a better fit.

NEWTONIAN MOMENTS: SURPRISE ACCOUNTING FOR ITSELF

In the late seventeenth century, Isaac Newton took a well-worn heuristic for finding the best fit—most famously imaged as Ockham's razor—and gave it an ontological twist. "Nature *is* simple," he insisted in *The Principia* in 1687.[53] Since that was the primary rule of his "Rules for Natural Philosophy," we shouldn't be surprised to hear it echoed today by the pioneers of natural language processing. In 2017, they laid the groundwork for our new era of generative AI with this sentence: "We propose a new simple network architecture."[54] Transformers, their new type of neural net, dispensed with the complex workings of their

53. Isaac Newton, *The Principia: Mathematical Principles of Natural Philosophy*, transl. I. Bernard Cohen and Anne Whitman, with Julia Budenz (Berkeley: University of California Press, 1999), 794 (emphasis added). In the Latin text, this phrase is "natura enim simplex est" (Isaac Newton, *Philosophiæ naturalis principia mathematica* [London, 1687], 402, https://www.loc.gov/resource/gdcwdl.wdl_17842/?sp=394&st=image&r=-0.242,0.005,1.853,1.236,0).

54. Ashish Vaswani et al., "Attention Is All You Need," arXiv, submitted 12 June 2017, last revised 2 August 2023, 1, https://doi.org/10.48550/arXiv.1706.03762.

predecessors—"recurrence and convolutions"—in favor of a simplified focus on "attention mechanisms."[55]

In that move, they followed the vector of Bacon's 1620 plan for the renewal of knowledge (*The Great Renewal*). Bacon turned from the convoluted logic of his Scholastic predecessors to (1) pay attention to "things as they are" and (2) scale up to as many things as possible. For Bacon, as we've seen, that meant collecting and writing as many "histories" of those things as possible; in our terms, he proposed a new database of information on everything from "clouds" to "mountebanks." For OpenAI, scaling up also meant attending to as many things as possible: the generative pre-trained transformer (GPT) was born in 2018 with a simple formula: combine "transformers and unsupervised pre-training" then add "a large amount of data."[56]

Bacon eventually got what he wanted—the knowledge revolution we call the Enlightenment. And, according to such recent headlines as "ChatGPT Heralds an Intellectual Revolution,"[57] we now have a similar result. Simplicity works—not because things stay simple, but because simplicity surprises us with both unexpected forms of complexity and new forms of simplicity. In fact, the most common thread across the divide between those who produce and study these new technologies and those who use and worry about them is *surprise*. From Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher to mathematician Stephen Wolfram and Google and OpenAI research scientists, the most compelling feature of GPTs has turned out to be not the expected gains in performance of machine language tasks but the unexpected and "unpredictable phenomenon that we refer to as *emergent abilities* of large language models."[58] Emergence is a simple phenomenon—it occurs when more is different. And, by definition, it's surprising, since it appears to come out of nowhere; something is "emergent if it is not present in smaller models but is present in larger models."[59]

By November 2022, ChatGPT team member Jason Wei had already cataloged "137 emergent abilities of large language models."[60] But the "big surprise . . . of ChatGPT," as Wolfram put it, even to those who built and scaled up this attention machine, is that a "neural net would be able to successfully produce 'human-like' text"; in fact, the "discovery" of ChatGPT is that this ability is even "possible at

55. Ibid.

56. Keval Nagda et al., "Ascent of Pre-trained State-of-the-Art Language Models," in *Advanced Computing Technologies and Applications: Proceedings of the 2nd International Conference on Advanced Computing Technologies and Applications—ICACTA 2020*, ed. Hari Vasudevan et al. (Singapore: Springer, 2020), 274.

57. Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher, "ChatGPT Heralds an Intellectual Revolution," *Wall Street Journal*, 24 February 2023, https://www.wsj.com/articles/chatgpt-heralds-an-intellectual-revolution-enlightenment-artificial-intelligence-homo-technicus-technology-cognition-morality-philosophy-774331c6.

58. Jason Wei et al., "Emergent Abilities of Large Language Models," arXiv, submitted 15 June 2022, last modified 26 October 2022, 1, https://doi.org/10.48550/arXiv.2206.07682.

59. Ibid.

60. Jason Wei, "137 Emergent Abilities of Large Language Models," Jason Wei's personal blog, 14 November 2022, https://www.jasonwei.net/blog/emergence.

all."[61] Its ability to chat—at length—appears to tell us something that we didn't know before—something, in fact, that we didn't know that we didn't know: "that language is at a fundamental level somehow simpler than it seems."[62] In a striking departure from the standard condescension arguments—"OK, this AI is better, but look at what it still can't do that we can"—Wolfram argued that "what we should conclude is that tasks—like writing essays—that we humans could do, but we didn't think computers could do, are actually in some sense computationally easier than we thought."[63] The surprise of emergence can at times come with a surprising touch of humility, especially for academics whose self-esteem is tied to the assumed complexity of their skill set.

And, at times, surprise comes with a wallop of fear—that has in turn produced more surprise. Enacting the error correction technique he popularized and that popularized him, Geoffrey Hinton startled friends and foes alike by backpropagating himself into the forefront of AI doom. His reason cements my case for invoking Newton's rule: Hinton underestimated the emergent power of simplicity. He had assumed "that software needed to become much more complex—akin to the human brain—to become significantly more capable."[64] It doesn't.

You can, like Hinton, feel shaken or you can recognize that we've entered another episode of the recurrent interplay of emergence and explanation. After surprise comes curiosity—the need to know what just happened, why, and what's next. Here's the sequence: (1) Nature is simple; (2) Scaling up simplicity sets the stage for emergence; (3) Emergence generates the need to understand our own surprise. It invites—some would say "demands"—new knowledge. For those working on these technologies, such as Wei and colleagues, "emergence motivates future research on why such abilities are acquired and whether more scaling will lead to further emergent abilities."[65] For everyone else, there's the need to explain it—either explain it away, as in Noam Chomsky's clearly irritated dismissal,[66] or, for those who don't feel the need to be unsurprised, explain what's emerged and create knowledge adequate to the consequences. "The point of being an A.I. researcher," as a Stanford PhD candidate put it to *The New York Times*, "is you should understand what's going on. We're constantly being surprised."[67]

Bacon had that same feeling—that he was living in a moment of recurrent surprise. He saw this as "good fortune," and he linked it to new and powerful "resources"; these three "mechanical things" were print, gunpowder, and the

61. Stephen Wolfram, "What Is ChatGPT Doing . . . and Why Does It Work?," Stephen Wolfram's official website, 14 February 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

62. Ibid.

63. Ibid.

64. Will Knight, "What Really Made Geoffrey Hinton into an AI Doomer," *Wired*, 8 May 2023, https://www.wired.com/story/geoffrey-hinton-ai-chatgpt-dangers/.

65. Wei et al., "Emergent Abilities of Large Language Models," 2.

66. Noam Chomsky (with Ian Roberts and Jeffrey Watumull), "Noam Chomsky: The False Promise of ChatGPT," *The New York Times*, 8 March 2023, https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

67. Rishi Bommasani, quoted in David Brooks, "The Fight for the Soul of A.I.," *The New York Times*, 23 November 2023, https://www.nytimes.com/2023/11/23/opinion/sam-altman-openai.html.

nautical compass.[68] All became agents of emergent surprise as they rescaled language, energy, and even space itself. And as with the AI researcher, Bacon thus saw the point of "learning" not as re-presenting established truth but as understanding what was going on. To catch up with surprise, he reoriented learning to advancement, and from 1605 to 1620, he laid out a plan and called for others to join him. Within decades of the moment Bacon made that call, the Royal Society and Newton answered it. An outpouring of new knowledge followed. To see Enlightenment from this perspective is to zero in on the kind of knowledge the sequence of surprise—simplicity/scalability/emergence—elicits, back then as well as now. That knowledge is what I've been calling explanatory; it is *surprise accounting for itself.*

A history of knowledge can identify certain moments of explanatory accountability as especially consequential; I'll tag them here as Newtonian Moments. When the *Principia* demonstrated that not only could math apply *to* nature but that there are mathematical regularities *in* nature, the relationship between simplicity and complexity took a surprising twist. If nature was not only simple but mathematical, then its simplicity could account for all of its complexity. Rules applied. One to all. When a single equation could describe "the falling of stones in Europe and America,"[69] Natural Philosophy had its Newtonian Moment.

That was the moment that—thanks to information's new genres of currency (newspapers, periodicals, Republic of Letters)—launched the flood of explanatory ventures so distinctly new and useful that even those who were in it called it "Newtonianism." All of these efforts sought explanatory regularities for emergent complexity expressed in the genres of simplicity: from laws and formulas to binaries (sublime versus beautiful) and concepts (survival of the fittest). Institutional effects of this new explicability followed, as its proliferation over the course of the eighteenth and nineteenth centuries required a division of labor: the branches of moral as well as natural philosophy morphed under pressure from this new explanatory imperative into the modern, specialized disciplines.

Then, in the early twentieth century, the sequence of surprise was triggered again, yielding another Newtonian Moment in which information (transformed in the moment) played a key role. The prompt this time was not Nature as mathematical, but the nature of mathematics itself. And the surprise was not something that worked but something that didn't. Kurt Gödel and Turing demonstrated that numbers did not add up to be the complete and consistent logical system that mathematicians had wanted and expected it to be. But that "no" opened the door to a "yes." Turing realized that the power of numbers must lie elsewhere—and embodied it in the simplicity and emergent reach of the Turing machine. One tape moving one box at a time yielded the enabling principle of modern computation: the jump to computational universality. One to all once again—this time, a single machine to simulate all machines.

On that basis, Turing's 1936 essay titled "On Computable Numbers, with an Application to the *Entscheidungsproblem*" is often invoked as the conceptual

68. Bacon, *The New Organon*, 100.
69. Newton, *The Principia*, 795.

starting point of modern computation and thus the beginning of our Information Age.[70] But the word "information" does *not* appear anywhere in that essay. Turing couldn't turn to that term, because it still operated from and within the matrix of currency. There, as we've seen, it signaled content of a new and useful kind—a usage Turing did employ in his letters and elsewhere. However, in that matrix, "information" had no purchase on his initial foray into computability.

For it to gain that purchase, information had to find a place in a new matrix configured in part by Turing's new approach to computability. First, and most famously, Turing thought through computation as what Bacon called a "mechanical thing"; the first paragraph of "On Computable Numbers" ends with the word "machine."[71] This edged computation—and the terms we use to describe it—from the realm of abstractions into the physical world, the same world inhabited by human computers, the physical entities that the Turing machine was designed to emulate. But that emulation was only the first step. It was proof of concept for computation's new relationship to the possible: the jump into universality[72] in which a single machine could simulate all other machines in its class.

Ten years after publishing "On Computable Numbers," after gaining experience building codebreaking devices during World War II, Turing set out to construct that machine. That meant not only assembling new devices, such as mercury tubes, but forming a new vocabulary to describe them. This was information's second chance. On the first page of his "descriptive account" of his "proposed electronic calculator," Turing used the word he didn't use before—and this time he underlined it, calling attention to the new work he wanted it to perform.[73] Not surprisingly, the context is his explaining the advantage that his mechanical computer will have over a human one: "Instead of repeatedly using human labour for taking material out of the machine and putting it back in at the appropriate moment all this will be looked after by the machine itself."[74]

There's the opening for a new use for an old term. The "material" that was moved and stored—what the machine worked on and with—was, Turing specified, "<u>information</u>."[75] But "information," to be used in the construction phase of computing, had to meet the new specifications I just discussed. The term's explanatory reach had to be extended into the physical (stored in "mechanical memory") and the possible (translated into the "digital").[76] Turing wasted no time, taking just a few sentences to do both. He wrote:

70. A. M. Turing, "On Computable Numbers, with an Application to the *Entscheidungsproblem*," *Proceedings of the London Mathematical Society*, 2nd ser., 42, no. 1 (1937), 230–65.

71. Ibid., 230.

72. Deutsch, *The Beginning of Infinity*, 125–46.

73. Alan M. Turing, "Proposed Electronic Calculator" (report E882, Proposals for Development in the Mathematics Division of an Automatic Computing Engine, National Physical Library, 1945), 2.

74. Ibid.

75. Ibid.

76. Alan Turing, "Computing Machinery and Intelligence" [1950], in *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, plus The Secrets of Enigma*, ed. B. Jack Copeland (Oxford: Clarendon Press, 2004), 443–48.

> Let us now be more specific. It is proposed to build "delay line" units consisting of mercury or water tubes about 5' long and 1'' diameter in contact with quartz crystal at each end. The velocity of sound in either mercury or water is such that the delay will be 1.024 ms. The information to be stored may be considered to be a sequence of 1024 "digits" (0 or 1), or "modulation elements" (mark or space). Those digits will be represented by a corresponding sequence of pulses. The digit 0 (or space) will be represented by the absence of a pulse at the appropriate time, the digit 1 (or mark) by its presence.[77]

Here is information striding into the matrix of possibility, where the next explanatory steps into the physicality and universality of the qubit could eventually be taken.

Information in this matrix became an emergent surprise, something that had to account for itself in new ways as it took on a new explanatory load. Turing himself, for example, increased that load by leveraging information-based universality to explore the nature of intelligence and, at the very end of his life, life itself in his work on morphogenesis. At the very same time, Claude Shannon took on the task of providing a new account of information by redefining it entropically, in James Duesterberg's words, as a "quantity of surprise": "the more easily predicted a symbol is the less information it contains."[78]

Even if, as Companion 1 argues, every age is an information age, our current one, as I've unfolded it in Companion 2, is different. It's full of surprises—whether it be the surprise of emergence or the surprise of entropy. And Turing's information machine pointed to one other—a surprise that proved foundational to AI but that still tends to confound us today: our notion of error was in error. To be more precise, in the matrix of currency, information could be both new and useful only if it was accurate—if facts were only facts. But when information enters the matrix of possibility, its connection to fact takes a back seat to its *counterfactual* character: to its capacity, as I pointed out earlier, to convey different messages. That's a problem for information's legacy genres—the genres of currency—where offering something that is counter to fact manifests as the problem of fake news, of falling into error. But in machines based on what Andrew Hodges has called the "logical structure of information,"[79] errors can be processed as yet another counterfactual possibility.

Whether those possibilities are arbitrary, illogical, or inconsistent doesn't matter for a universal machine capable of dealing with patterns of any sort.[80] We had thought, Turing wrote in 1948, as he started to spec and build his information-processing device, that "machine[s] must not make mistakes," but then he added, in his low-key, offhand way, the "but" that jump-started modern AI: "But this is

77. Turing, "Proposed Electronic Calculator," 5.

78. James Duesterberg, "Zero Signifier" (lecture, NYU Digital Theory Lab, 5 April 2024).

79. Andrew Hodges, *Alan Turing: The Enigma* (Princeton: Princeton University Press, 2014), 317.

80. "Because computers," wrote Douglas R. Hofstadter, "at base manipulate numbers, and because numbers are a universal medium for the embedding of patterns of any sort, computers can deal with arbitrary patterns, whether they are logical or illogical, consistent or inconsistent" (foreword to *Gödel's Proof*, rev. ed., by Ernest Nagel and James R. Newman, ed. Douglas R. Hofstadter [New York: New York University Press, 2001], xix).

not a requirement for intelligence."[81] In fact, he realized, mistakes were crucial to developing a key component of intelligence: you have to make them to learn from them. Mechanical computers could ingest and produce errors, but error, Turing realized, is not a barrier to what he called "intelligent machinery" but a prerequisite for advancing toward that goal.

THE HISTORY OF KNOWLEDGE AND BETTER EXPLANATIONS (THE CASE OF COUNTERFACTUAL CONJECTURE)

That brings us, of course, back to the issue of advancement now: the test-case of GPTs. In an effort to re-tune the history of knowledge to advancement, I've tried to come at that elephant from a number of angles—all of them taken to illuminate our central concern with information. Bacon's highest hope for his history of knowledge was to make us better at advancing it. To put it simply, it should take us somewhere, so I'll end, as promised, by stepping out onto a limb: Can what we now know about information's history help us formulate better explanations of the work it currently does in the world?

"Better" requires error correction, so it's best not to make errors about error. The sad fact is that, almost a century after Turing, we are still making errors about information and error—a habit that is now stalking GPTs as its next victim. Critics love to have their cake and eat it too: dismissing the technology for its error count and fearing it for the same reason. That's not surprising. Pessimists tend to be implausibly optimistic about their own predictions, and AI doomsayers are no exceptions. If ChatGPT doesn't get you now with its error-filled ways, it will get you later when it nevertheless manages to turn into an all-powerful but, of course, ethically-flawed superbrain. In both scenarios, however, the critics fundamentally misconstrue the nature and function of the apparent mistakes. As we'll see, what looks like error is actually a key to GPT's emergent abilities.

In a bit of anthropomorphizing perhaps generated by the fear that this kind of AI may actually be succeeding, critics have called out its errors as "hallucinations." But what or who is hallucinating here? Yes, ChatGPT responds to prompts with outputs that include more than facts. But what are those things? To hallucinate is to wander off, and for those complaining about generative AI, the assumption is that the technology is aimlessly wandering into error. Yes, hallucinations are errors, but that output is not simply mistaken. As we've seen, GPTs, in their turn to simplicity, surprise us with emergent abilities that invite explanation. Our moment of surprise is, in the history of knowledge, language at its Newtonian Moment. Numbers—freed by Turing and Gödel from the straitjacket of systematic completeness and consistency—had their moment in the 1930s, becoming a universally computable medium for patterns of any kind. Now words are having their turn. Ushered into the computable as tokens and vectors, they have been transformed to work in new ways, confronting us with confusing behaviors.

81. Alan Turing, "Intelligent Machinery" [1948], in Copeland, *The Essential Turing*, 411 (emphasis added).