Origin Gaps and the Eternal Sunshine of the Second-Order Pendulum

Simon DeDeo*

3 March 2017

Abstract

The rich experiences of an intentional, goal-oriented life emerge, in an unpredictable fashion, from the basic laws of physics. Here I argue that this unpredictability is no mirage: there are true gaps between life and non-life, mind and mindlessness, and even between functional societies and groups of Hobbesian individuals. These gaps, I suggest, emerge from the mathematics of self-reference, and the logical barriers to prediction that self-referring systems present. Still, a mathematical truth does not imply a physical one: the universe need not have made self-reference possible. It did, and the question then is how. In the second half of this essay, I show how a basic move in physics, known as renormalization, transforms the "forgetful" second-order equations of fundamental physics into a rich, self-referential world that makes possible the major transitions we care so much about. While the universe runs in assembly code, the coarse-grained version runs in LISP, and it is from that the world of aim and intention grows.

How happy is the blameless vestal's lot!
The world forgetting, by the world forgot.
Eternal sunshine of the spotless mind!
Each pray'r accepted, and each wish resign'd
— Alexander Pope, "Eloisa to Abelard"[1]

The world we see, and the worlds we infer from the laws of physics, seem completely distinct. At the blackboard, I infer that a thin skein of gas will coalesce into objects such as stars and galaxies. With a few more assumptions I predict the range of masses that those stars should have, beginning from an account of initial quantum fluctuations. Today, it's considered a reasonable research goal to reduce even that story, of the wrinkles in spacetime that seeded Andromeda, to the first principles of basic physics: Hawking radiation at a horizon, or the quantum statistics of a multiverse.

If, however, I try to infer the existence of the blackboard itself, and the existence of people who write on it and themselves infer, I am stuck. I find myself unable to predict the spectrum of desires and goals that evolution can produce, let alone the ones that arise, apparently spontaneously, from the depths of my own mind. The utter failure of otherwise reliable tools to generalize to this new domain is one that many scientists experience when they cross between fields. Not just scientists: as Sherry Turkle pointed out, even young children experience it, when confronted by electronic toys. There is something about the experience of life (or life's substrate, computation) that goes

*Department of Social and Decision Sciences, Carnegie Mellon University & the Santa Fe Institute. `sdedeo@andrew.cmu.edu`; `http://santafe.edu/~simon`

[1] `https://www.youtube.com/watch?v=E8xGabLlI8k`

beyond the physical mechanisms they're used to describing. A child faced with an apparently living machine looks in the battery compartment to see what powers it [1]. Whether it is felt by an adult scientist at the blackboard, or a child with a toy robot, it is at heart an experience of the gap between the purposeful world of human life and the aimless one of stars. Our tools can not make the leap.

Our tools do, of course, work if we are allowed to assume the existence of meaning-making beings to begin with. Fluid dynamics works well to describe the flow of traffic through my city, while variants on the Ising model allows me to predict the racial segregation I see as I pass through it, and further generalizations get us off on the right foot for thinking about how my messily-wired brain might learn and remember and experience at all.

Yet no matter how well we do once meaning-making beings are taken as a given, we stumble when we are asked to predict their very being at all. It is *this* gap, the inability to leap from one side to the other, that begs explanation, and I refer to it as the Origin Gap because it is familiar to those working in the "origin" fields: the origin of society, the origin of consciousness and meaning, the origin of life. It is the gap that gives those fields a very different flavor from their parallels in the sciences of their mature subjects. Origin of society looks very different from social science and anthropology; origin of consciousness looks very different from psychology; origin of life looks very different from biology.

The gap, I claim, is understandable, even (one might say) say predictable. In this essay, I'll first show that the existence of the gap is the consequence of a basic pair of facts in the theory of computation. Second, that particular aspects of the laws of physics make it very likely that in the evolution of the universe, such gap will naturally appear. Taken together, these facts explain how "mindless" laws lead to the emergence of new realms of intentional behavior. At the heart of this essay's explanation of the gap will be that the kind of intention, aim, and meaning we really care about also has the capacity to refer to itself.

1 The Mathematics of the Gap

From the mathematical point of view, the origin gap begins with the fact that

1. it is easy to describe everything.

2. it is much harder to describe one thing.

This has a counterintuitive feel to it. We began, both as individuals and as a species, by describing particular things (that big mountain, this frozen river, that tall woman, this cold morning). It therefore feels as if this task must be easier than the more elaborate habits of generalization, abstraction, the tools of set theory, category theory...

Yet when we make this leap, we forget how many millions of years of evolution went into teaching us how to produce these descriptions. What it means to be one thing rather than two, the identification of useful boundaries or persistent patterns, what it means for an argument to be valid: each is a question subject to endless debate. We see this in the history of philosophy, but a more contemporary example comes from the history of Artificial Intelligence (AI). AI gave a name to the feeling that rules of description could never be exhaustively specified. They called it the "frame problem", and advanced societies across the globe dumped literally billions of dollars into solving it. Until, that is, they discovered that the quickest way to solve the problem of describing something was to avoid specifying the rules at all.

Rather than define in computer code a beautiful sunset, or a valid argument, researchers now build learning machines that watch and copy human response. Don't describe a cat to a computer—have it learn what a cat is from the pictures we take to celebrate them on the internet. In this way, the code can rely on the accumulated wisdom of evolution. Which is only natural, since (of course) a computer is build by evolved creatures to serve their needs.

In as much as our lives are dominated by artificial intelligence we have, for now, given up on describing things. But it turns out that to describe *everything*, by contrast, is simple. It only needs to occur to you to do something so trivial as to try. Borges did so in his short story *The Library of Babel*, where he imagined a series of interconnected hexagonal rooms, walled by shelves and stacked with books, and each book containing the letters of the alphabet, spaces, commas, and periods in different orders.

How much is contained in everything! Of course, in Borges' library there are an overwhelming number of nonsensical books, cats typing on keyboards, but also (again, of course) the complete works of Shakespeare, as well as every variation on those works, and every possible edition with typographical errors, and (as Borges might have gone on) the plays that Shakespeare might have written were he really Francis Bacon, or Elizabeth the First, or an alien from Mars, as well as all the incorrect extrapolations of those conjectures, and so on to the limit of one's imagination, and (then) beyond.

Imagine that we have instant access to the text of any book. It's simple to find all the books that include the word "Shakespeare": just send your robot out to search book by book and return the ones that contain that string of letters. Of course, it'll also recover nonsense books, books full of jumbled letters that happen, once in awhile, to spell the name: "...casa,cWas,,,qwh g Shakespeare acqq CO..."

Here's a harder problem: how to locate the books on Shakespeare that make sense? Give instructions to the robot to gather them together. Or, imagine the layout of the Borges library as a wireframe image on your computer screen, and the rules of shelving to hand. Outline, or click with a mouse, the shelves to pull.

Under some very basic assumptions (which we'll address below), the strange thing about this more complicated query that it can imagined, but not actually made. Like the idea of squaring the circle, of producing using straight-edge and compass a square whose area is equal to a given circle, it seems that it should be possible. And yet it is not: the shape your mouse carves out, although imaginable in each fragment—"this book, not that"—is an impossible shape, a shape impossible to define and therefore to draw. In its infinitely detailed structure, it is at each scale completely unrelated to the scale above.

The existence of such shapes (or the non-existence of the rules of their construction) seems counter-intuitive at first, because it is the nature of human beings to ask for things that are possible. "Bring me all the sugar in the kitchen"; "Find me all the students in the engineering department". We are not used to asking questions that have no answer.

Yet for it to happen all we need is that any description of what it means to be a sensical book on Shakespeare requires more than just pattern-matching (*e.g.*, the presence or absence of the word "Shakespeare"). Impossible questions emerge when they become about pattern-processing, pattern manipulation, pattern computation. Something sophisticated enough, in particular to allow us to have something operate on a description of itself.

This is, of course, exactly what takes place. If we read a book on Shakespeare we do more than count words and match them to lists. We think about those words, the combinations they fall in, and what one combination means for another. We reason about a passage, follow its arguments and conjecture counterarguments. And when we give ourselves, or a machine, that power, we become fundamentally limited in the questions we can ask and answer about what we, or it, is going to do. It becomes impossible, even, to draw outlines around the behaviors we do, or do not, expect. In contrast to the condensation of gas into stars, we can not derive, ahead of time, the space of books that scholars will write about Shakespeare.

This is why the origin problems are hard. The things whose origins most intrigue us are also the points at which systems gain new powers of self-reference. And these moments lead to new categories, new phenomena, that we can literally not predict ahead of time. Once we have an example, we can ask questions about it, do science on it, just as we can take any particular volume from the Borgesian library and read it. But to begin with the space of all possible things that can happen, and then to draw the outlines of what we expect to see on the basis of a self-referential process, is something else altogether.

I'm hardly the first to draw attention to the importance of self-reference for the problems of life. Sara Walker and Paul Davies have pointed to the self-referential features at the heart of the origin of life problem [2]. Stuart Kauffman puts self-reference at the heart of both biological and social evolution, and in places conjectures explicitly G odelian arguments [3]. My own work, and that of my collaborators, on social behavior suggests that social feedback, the most primitive form of self-reference and something we see in the birds just as much as the primates [4], is at the origin of major transitions in political order [5].

The gap between physics and the meaningful experiences we associate with life thus turns out to have an unexpectedly mathematical feel. The emergence of meaningful experiences is associated with the emerge of new forms of self-reference, but questions about the basic properties of self-referential systems are (on pain of logical inconsistency) impossible to answer in the complete and general fashion we expect from derivations in the physical sciences.

Asked to sketch out the consequences of a new self-referential phenomenon—say, an organic polymer than can refer to, modify, and reproduce itself—we stumble, because the very question is unanswerable. Given a particular example (the replication machinery of the bacteria *E. coli*) we can do a great deal of science. But to delineate all the life this makes possible is equivalent to picking books of Borges' library.

2 The Physics of the Gap

It is one thing to ascribe the gap to the emergence of new systems of self-reference. But why should self-reference come into being at all? At the heart of self-reference is the existence of a memory device, and something that can navigate it in a "sufficiently sophisticated" fashion. Smith and Száthmary's famous 1997 piece [6], on the major transitions in the biological record, recognize this implicitly, placing the discovery of new information processing and recording mechanisms at the center of each transition. Social scientists [7], scholars of "deep history" [8], and cognitive scientists [9] each draw attention to new institutions, like cities, or new cultural practices, such as writing or social hierarchy, or even new abilities from physiology itself, such as genes for speech and syntactic processing, that enhance the ways in which we can remember, or transform what we remember. Major leaps occur when something previously forgettable, or lost to noise, finds a means to be recorded, translated into a referential form, and processed and combined into new forms. When social debts become stories told around a campfire—or transform into money and markets [10]—we see not just an augmentation of life as it was known, but the unpredictable creation of entirely new forms of being.

Each of these moments is a shift in the nature of the world, and a clear topic of scientific study. Whether we study the details of its emergence, or the patterns it displays that generalize beyond its historical context, any one of them is the task of a lifetime. But what makes memory, and self-reference, possible at all?

Strangely enough, it's not baked in to the fundamental laws of physics, a fact that was driven home to me early in my career, at the University of Chicago, when I worked with Dimitrios Psaltis, and Alan Cooney, physicists at the University of Arizona. We were puzzling over a strange class of models in fundamental physics. Despite their mathematical coherence, they were, at heart, unstable: any universe that obeyed their laws would sooner or later explode, everywhere and instantaneously, into fountains of energy with no apparent end, as if slipping off the top of a hill that had no bottom.

What made them unstable was how they handled time. In the physics you encounter in high-school it's crucial that Newton's laws of motion talk about the relationship between force and acceleration: $F = ma$, force is mass times acceleration, or perhaps more easily, $a = F/m$, the acceleration you experience is the force applied to you, divided by your mass. Acceleration is connected to the passage of time; it's how fast your velocity is changing, or, more formally, the "second derivative of position with time". Newton's laws then connect forces you might experience to a phenomenon we call gravity: objects create a gravitational field, and at each point that field subjects objects to a certain amount of force. Other laws talk about other sources of force: electrical, or magnetic, for example. All connect back to acceleration, the change of velocity with time.

This is all awesome and highly addictive to talk about if you have a certain bent of mind, but one of the basic facts about these laws is that you never see anything with more than two derivatives in the fundamental equations. When you write them down, you only ever talk about (1) a basic set of quantities, say, position, gravitational field, etc.; (2) how these quantities change with time; and (3) sometimes, how these changes in time change with time. If you have a theory where higher derivatives enter in, where you talk about changes in changes in changes, then the theory becomes unstable in some really uncomfortable ways, leading to things like spontaneous infinite accelerations which you never observe (or really could imagine observing) and that would really ruin your day if you did.[2]

This sounded just fine to me, until I remembered something from my high school physics teacher. The change in acceleration, the *third* time derivative of position, has its own name, amusingly enough, called "jerk". Jerk is what you experience when an elevator starts up. When it's moving at a constant velocity, you feel nothing. When it's accelerating, you feel heavier (if you're going up), or lighter (if you're going down). But when it switches from not accelerating to accelerating, or vice versa, you experience a sudden change in your weight. You're experiencing the elevator jerking you up, or the pit of your stomach dropping out when it descends.

The fact that I experience jerk is very strange. Am I not a creature that lives in the physical world? Am I not forced to obey the laws of physics? And don't I know, from a bit of mathematics, that the laws of physics only deal with quantities with two time derivatives or fewer, or risk being violently unstable if they don't? But if all that's true, how can jerk, a third-order quantity, play any causal role in my life, such as causing me to say "oof", or making me feel queasy, when the elevator moves? How can my psychological laws obey equations that are ruled out as physical laws?

I remember a spooky feeling when I put this argument together, and for a brief moment wondering if this proved the existence of a separate set of psychological laws beyond or parallel to physics. The answer turns out to be a bit simpler, if no less intriguing. The instabilities that emerge for theories with higher-order derivatives are real, and barriers to them being basic laws of the universe are real as well. But there's nothing that prevents them holding for a while, in limited ways, so that the instabilities don't have time to emerge.

[2] We were working on how to "cure" these instabilities in certain limited regimes; you can find our answers (and further references) in a series of papers we wrote together [11, 12].

And that's the reason I can feel the jerk. I have a brain that senses acceleration. It's possible for that sense to rely directly on fundamental laws (it doesn't, actually, but it could). But in order to report the sensation of jerk to my higher-order reason, my brain has to go beyond fundamental physics. It has to use memory to store one sensation at one time and compare it, through some wetware neural comparison device, to a sensation at a later time. Similarly, I can measure the acceleration that my car undergoes by hanging a pendulum from the ceiling and seeing where it points. But to measure jerk, I have to videotape the pendulum, and compare its location at two different times. There's no "jerk pendulum" I can build that relies directly on the basic laws of physics that apply everywhere and for all time. The fundamental laws are forgetful, the "blameless vestals" of the Alexander Pope quotation that begins this essay.

It's strange to think that a visceral and immediate feeling, like the drop you feel in the pit of your stomach when the elevator descends, is an experience filtered through a skein of memories. These memories present what is actually a processed and interpreted feature of the world as if it were a brute physical fact. Yet it so turns out that some things, like "force", are truly fundamental constituents of our universe, while others, like "jerk", are derived and emergent.

Jerk gets into the physical world through memory, but it's hardly the most impressive feat of memory we do. A man descending a New York City skyscraper is in the presence of far greater feats of memory and processing than just what travels down his vagus nerve. Yet jerk also gives us a clue to how those far more sophisticated memories might have gotten going. The experience of jerk is an atavism of a far more primal event, one that began well before there were brains to feel it.

This is because, while (to the best of our knowledge) higher order "memory terms" like jerk are forbidden from playing a role in fundamental laws, they do emerge in an unexpected fashion. We rarely perceive the world at its finest grain, in all of its fluctuating detail, at the assembly code level, you might say as a computer programmer. We see, instead, averages: not everything that happens in a single patch, but a coarse-grained summary of it, an average, as if the lens was smeared with vaseline. To give a full account of the role of that averaging, or coarse-graining, in the physical sciences would take us very far afield, but also (many now believe) into some of the best mysteries we have to hand, including an explanation of the second law of thermodynamics and the decoherence, or collapse, of the quantum-mechanical wavefunction.

Here we care about coarse-graining because, by averaging together nearby points, it introduces the possibility of inducing physical laws that (in contrast to their forgetful fundamental cousins) do have memory. When we smooth out the world, when we average out some of the small-scale bumps and fluctuations, we produce a new description of it. The laws that govern those coarse-grained descriptions, in contrast to the ones that applied at the shorter distances and for the finer details, can have memories, can include higher derivatives. They may, in certain cases, be unstable, but this is no longer an existential threat: it just means that, occasionally, the coarse-grained description will fail. The fine-grained details will emerge with a vengeance, ruining the predictive power of the theory. You'll be reminded of the limits of your knowledge, but the universe will not catch fire.

The technical term that physicists use for this is renormalization. Physicists use it for all sorts of problems, and call the theories that emerge for coarse-grained systems "effective theories". My colleagues and I have thought about them for a long time, as both a fact of life for deriving one scale from another (social behavior, say, from individual cognition), and a metaphor to help explain why biology differs so much from biochemistry, and why averaging-out might not just be a good idea, but might make new forms of society possible [13, 14, 4, 5].

If you're a computer scientist, you might say that while the universe runs in assembly code, the coarse-grained version runs in LISP. Here, coarse graining gives the possibility of memory and—

with some interesting dynamics for how those memories inter-relate—the self-reference that makes certain features of the future logically unpredictable based on what came before. The memories we have now, biochemical, electronic, on pen and paper and in the cloud, are far more complex than the ones than appear in a physicist's coarse-graining prescription.

You get a great deal from the averaging-out a cell wall allows you to do, another boost from the ways in which neurons average out the data from your eye, and another from how a story you tell summarizes the history of your tribe. No essay can derive the biochemical story, or the cognitive one, or the social one. Here we point to a crucial moment where they all begin: not in the perception of detail, but in its selective destruction and lossy compression. The arguments of this essay suggest that averaging-out may have been the first source of memory, and thus self-reference, in the history of the universe. Perhaps that happened first in biochemistry; perhaps it had an even earlier start.

3 Learning from the gaps

The leaps the universe has made, from non-life to life, chemical reaction to mind, individualism to society, aimless to aim-ful depend in a basic way on how new features of the world—physical features, biological features, social features—become available for feedback and self-reference. If this essay is correct, then it is those self-referential features that, in creating predictive gaps, attract our curiosity. And it is those features that, at the same time, make the problems so hard. You might say we're constantly nerd-swiped by the origin gaps [15].

Though I've focused on the the primordial scene, the origin of memory, and located it in the coarse-graining of fundamental theories, I've also suggested that this coarse-graining process might be something worth attending to at later stages as well. This suggests an intriguing possibility: that there are more stages yet to come, new accelerations and ways for us to reflect upon ourselves, and (in doing so) to create new forms of life. It's natural, at this cultural moment, to look to the world of artificial intelligence and to ask what our machines will do for—or to—us. As we create machines with inconceivably greater powers to reflect, we may be setting in motion a process that will leave behind, for future millennia, a new origin problem to solve.

References

[1] Sherry Turkle. *Life on the Screen.* Simon and Schuster, New York, NY, USA, 2011.

[2] Sara Imari Walker and Paul CW Davies. The algorithmic origins of life. *Journal of the Royal Society Interface*, 10(79):20120869, 2013.

[3] Stuart A Kauffman. *Humanity in a creative universe.* Oxford University Press, Oxford, United Kingdom, 2016.

[4] Elizabeth A Hobson and Simon DeDeo. Social feedback and the emergence of rank in animal society. *PLoS Computational Biology*, 11(9):e1004411, 2015.

[5] Simon DeDeo. Major transitions in political order. In S.I. Walker, P.C.W. Davies, and G.F.R. Ellis, editors, *From Matter to Life: Information and Causality.* Cambridge University Press, Cambridge, United Kingdom, 2017. Available at `https://arxiv.org/abs/1512.03419`.

[6] John Maynard Smith and Eörs Száthmary. *The Major Transitions in Evolution.* Oxford University Press, 1997.

[7] F. Fukuyama. *The Origins of Political Order: From Prehuman Times to the French Revolution.* Farrar, Straus and Giroux, 2011.

[8] Daniel Lord Smail. *On deep history and the brain.* University of California Press, 2007.

[9] Merlin Donald. An evolutionary approach to culture. In Robert N Bellah and Hans Joas, editors, *The Axial Age and its consequences.* Harvard University Press, 2012.

[10] David Graeber. *Debt: the first 5,000 years.* Melville House, New York, NY, USA, 2014. Updated and expanded.

[11] Simon DeDeo and Dimitrios Psaltis. Stable, accelerating universes in modified-gravity theories. *Physical Review D*, 78(6):064013, 2008.

[12] Alan Cooney, Simon DeDeo, and Dimitrios Psaltis. Gravity with perturbative constraints: Dark energy without new degrees of freedom. *Physical Review D*, 79(4):044033, 2009.

[13] Simon DeDeo. Effective theories for circuits and automata. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3):037106, 2011.

[14] Jessica C Flack. Multiple time-scales and the developmental dynamics of social systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1597):1802–1810, 2012.

[15] Randall Munroe. Nerd-swiping. Available at `http://xkcd.com/356`.