A Plea for Inexplicability

David Weinberger[*]

Independent scholar

Submitted: July 26, 2022 – Accepted: February 6, 2022 – Published: March 15, 2023

Abstract

Explanations have long played a crucial role in the west's metaphysics of control by providing levers by which we exercise control, but also by grounding our assumption that we are the legitimate and rightful masters of the world. Within this metaphysics, inexplicability looks like a failure. But machine learning may be teaching us a different lesson: Explicability is not a property of the universe, but inexplicability is. In short, the world is the ultimate black box. Accepting this may lessen our one-sided commitment to the general rules, principles, and laws that make western-style mastery seem possible. We may instead be entering a time when we are willing to value the particular at least as much as the general, and can embrace the hiddenness of the universe that grounds all that shows itself to us.

Keywords: AI; interpretability; explicability; philosophy; chaos.

[*] ✉ david@weinberger.org

A case for inexplicability — even an odd one like this — does not necessarily oppose the remarkable work being done worldwide to make machine learning (ML) more and more interpretable[1], any more than being in favor of preserving the wilds means one must oppose eco-friendly housing developments.

In fact, that stretched analogy may be more apropos than it at first seems, for there is a connection between the wilds and the inexplicable. Both challenge our western cultural commitment to mastery and control by outstripping our grip. That commitment is a metaphysics that frames our thoughts and channels our intentions, with centuries of disastrous results for the people under its thumb and for the planet itself.

But now machine learning is defining a new age that presents us with an opportunity to break that grip — even while extending our control — by exposing the deep, human and ultimately inevitable roots of inexplicability.

Or, perhaps inexplicability is not a root. Perhaps it is the ground on which we walk.

1 To Explain and Control

The western metaphysics of control views the world as existing as a resource for us to use, abuse, and use up. The "us" has long meant certain privileged classes, including elderly white men like me.

Explanations play a crucial role in this metaphysics for two reasons.

First, explanations are the proximal means by which we exercise control: Knowing how something works give us a leg up on being able to manipulate it.

Second, the fact that we can explain the world (to one degree or another), is an integral part of our claim to being the legitimate and rightful masters of the world. That's what makes the will to control not just a power grab, but the responsible act of rational creatures. Or so we have told ourselves.

Breaking the grip of this metaphysics starts with being very explicit in understanding that explanations are tools. As such, they are hugely important. Still, explicability is not a property of the universe any more than being edible is.

In short, the universe doesn't owe us an explanation. And if it gave us one, like Wittgenstein's lion that can speak, we wouldn't understand what it says.

And here's the point that brings me to make a plea for the inexplicable: Within the metaphysics of control, inexplicability looks like a failure. That's an easy but dangerously false conclusion people may draw from the urgency of the current drive for explicability.

Inexplicability is *not* a failure. It *is*, however, a property of the universe.

To break the grip of the metaphysics of control, we need — while continuing to pursue interpretability — to acknowledge and even honor inexplicability.

There is a truth to inexplicability.

2 Ignoring the Inexplicable

We can predict within minutes when the New Horizons probe will fly past Pluto four billion miles away, but we can't predict or fully explain why just those people were in a crosswalk with us this morning, or why the restaurant at the end of the day was out of eggplant parmigiana. Yet

[1] For an overview and taxonomy see Linardatos et al. (2021).

https://doi.org/10.6092/issn.1971-8853/15296

the metaphysics of control exerts such power over us that even though just about every moment of our lived lives refutes it, we still hold on to it as a framing idea.

We call "accidents" the stream of inexplicable events that stuff of our lives — little things that are fully determinate but that have causal chains too knotted for us to unravel. No big deal. We call something an accident usually to write it off, to give up on trying to pin down why it happened, to give up on assigning responsibility[2], to cease wondering about it or its meaning.

On the positive side, accidents can also be a source of wonder. Why did that particular leaf, brown and red but still with veins of faint green, fall to that exact spot on the path of our morning's walk? We can be overwhelmed by the precise and unlikely beauty of the accidental.

But more often, we live surrounded by accidents that we just ignore because they are accidents. Thus, we live in the Kingdom of Accidents at the same time as we hold explicitly to the paradigm of mastery, of control. Our experience and our understanding are in contradiction. We have developed vastly complicated ways of squaring that circle, never well.

We are a peculiar species.

3 Inexplicability Is Not an Accident

The most distinctive aspect of machine learning is that it programs itself based on the data it's given. That means that left on its own, ML will tend toward bias because data almost always reflects societal biases. Likewise, left on its own ML will tend toward inexplicability because what it learns from data will not be constrained by what human beings can understand.[3]

But ML need not be left on its own. Everyone agrees that we need tools to discover, analyze, evaluate, prevent, fix, and remediate the biases to which models are susceptible. Interpretability can be a powerful weapon against bias as well as against other ways ML can go wrong. As advances are made, it will become yet more powerful. But as models become more complex it may not be enough. And there are other ways to address the problems ML is heir to. For example, counterfactuals can verify certain types of bias without knowing what introduced the bias (Mittelstadt et al., 2018). Hyperparameters can be tweaked to produce fairer outcomes without necessarily knowing how they're doing it.

But for now in any case, some useful models have at least some elements that are inexplicable because of the nature of the accidental: Everything in the Kingdom of Accidents is causally determinate,[4] even if the causes can be too multiple, connected, and interdependent for us to ferret them out. That rarely bothers us because for the vast majority of accidents we have no incentive to engage in the social act of explaining.

But machine learning models can do surprisingly well in that kingdom because they can wring probabilistic significance from sets of tiny relationships. Clockwork laws of mechanics can explain planetary movements. Computer programs can help you figure out whether your company will boost revenues more if you use shoddier materials or increase your marketing spend. Machine learning can anticipate the next book you'll buy based on every book you've ever bought, but also perhaps based on the complete range of retail goods you've bought plus how quickly you're mousing today.

[2] In American English, and perhaps elsewhere, "accident" has a specialized meaning when it involves cars hitting each other. There the question of responsibility is not so easily avoided.

[3] Or perhaps not (Doshi-Velez & Kim, 2017).

[4] Free will throws some kinks into this, but we're not about to get sidetracked by that question, thank you very much.

Machine learning's outputs can be as unpredictable as the next license plate number you'll see in the Kingdom of Accidents, but they can be unpredictable and yet useful because they find patterns hidden within the seeming randomness that is characteristic of everyday life.

4 Transformational Not-knowing

"We don't know how it works."

We hear this sentence often, in various versions and in voices at various pitches of panic. We are coming to rely on ML for a broad range of applications, many of which are important to our lives together on our shared planet, most of which offer some opportunity for pernicious biases to make life worse. Yet we don't know how it works!

This is a transformational expression. But I believe the truly transformative part of it is not the negative "We don't know how" but the positive affirmation the phrase contains: "It works!"

The "It works!" upsets the West's multi-millennial conviction that knowing how things — ultimately, the world itself — work is our human essence and destiny. We are by definition the rational animals, aren't we? (No, we're not, but let's keep going.)

The obvious implication that sometimes machine learning works even though we don't know how it works is that it works because it's uncovered something truthful about the world… and what it's uncovered is complex to the point of inexplicability. The "it works" reveals that it's the *world* that's so complex.

Not that that's always the case. Something can be inexplicable simply because we haven't discovered the intelligible explanation. But it is sometimes the case that the complexity of the model reflects the complexity of the data that reflects a complexity of the world.

Thus, the effect of the "It works!" is — or at least can be — that the *world* is shown to be the black box.

That doesn't mean we can't ever explain anything in the world. We can because explanations are tools, never complete, and always approximate. Even a fully determinate physical event such as the length of time it takes for a coin to drop from the Leaning Tower of Pisa can never be predicted or explained with complete accuracy because we'd have to include everything from the air fluctuations caused by nearby pigeons to the gravitational pull of distant stars.[5] But that's fine. An explanation only has to be accurate enough for the task that motivates it.

We should not let our facility wielding explanations as tools convince us that we can explain everything. After all, in the Kingdom of Accidents, we've learned to only ask for explanations where they are possible: For a toilet that won't flush, yes. For every sparrow that falls, no.

5 The Rise of the Particular

In the West, we've taken principles, laws, and generalizations as the higher truth, the lenses through which we can see the truth beneath the chaos of particulars. Now machine learning may be rebalancing the general and the particular.

[5] For an excellent brief history of the relationship between explanations and predictions according to the philosophy of science, beginning with Hempel and Oppenheimer in the 1940s — as well as an argument that explanations are tools we use to manage complexity — see Douglas (2009). Note that Douglas' article is solely about scientific explanation, which can by no means be presumed to be the same as the sort of explanations we want of machine learning outcomes and operations.

Of course, machine learning doesn't reject all generalizations; overfitting is the hollowest of successes. But machine learning's generalizations don't do what we have valued traditional generalizations for.

First, we liked traditional rules, principles and other forms of generalization because we could understand them.

But machine learning comes up with generalizations without caring if we understand them.

Second, we valued generalizations because of their broad explanatory power — the fact that they apply to many different cases.

But ML's generalizations may not generalize beyond the model that results from them.

Third, we liked generalizations because we could apply them to particulars. They're the levers used by the metaphysics of control.

But we can't apply machine learning's generalizations. We have to use a model to do that.

This third point isn't just a technical nit. Having to use a model to apply the generalizations means we give up our role as the knower of rules who thereby have legitimacy to rule over particulars.

In this sense, ML generalizations are a triumph of the particular. They arise from particulars, and may be as inexplicable as those particulars — as inexplicable as life in the Kingdom of Accidents. And they may be generalizations that apply only to particular situations, the way the solution to a murder mystery is specific to its details.

I'm not pretending to know exactly what this metaphysical switch would mean if it happens at all. But let me give you a worryingly vague idea of why I think it matters.

First, ML extends the human grip quite dramatically, and thus is a form of mastery and control. But because it can be inexplicable, and is always statistical in its nature, this can be mastery without the all-too-common arrogance and the often blinding confidence of knowledge.

Second, attention paid to particulars is attention paid to *differences*, not just to the similarities that let us explicitly subsume particulars into general categories subject to general rules

Attending to — valuing — differences has obvious and positive social, political, and epistemic implications… although I acknowledge that it's a long way from this to a more just and equitable world.

Third, valorizing particulars may make a difference in our everyday understanding of our everyday experience, chipping away at the metaphysics of control as our controlling paradigm.

Perhaps we'll recognize the irreducibly chaotic nature of a world so rich in particulars, even if it is governed by simple rules.

Perhaps we'll accord at least equal dignity to the unknowable truth of particulars as to the universal rules that govern them.

Perhaps more of the world will become visible to us: the details not easily summed up and thus ignored.

Perhaps, we'll let ourselves see and credit the stubborn beauty of the particular.

Perhaps we'll start to heal the gap between our everyday experience of the Kingdom of Accidents and our metaphysics that denigrates mere accidents. Maybe.

As for the principles, theories, or laws we have worked so hard to discover, I don't for an instant think we're in danger of discarding them. But perhaps when applying them we'll recognize the depth, intricacy, and indefatigable spirit of the particulars that together make approximate their every encounter with a generalization.

6 The Hidden

The Age of AI is perhaps bringing us to a unique moment in which we can flip the metaphysics of control. This could be a Copernican turning point for us.

My concern is that we might miss this opportunity if the public only hears that inexplicability is always or primarily a failure — a problem that supposedly we can, should, and will overcome if we try hard enough.

But inexplicability is not a human failure. It is the human condition.

In the metaphor that guided Martin Heidegger[6], the ground of what we uncover is the vast hiddenness on which we walk[7]. To explain something is to shine a particular light on it because of something we care about. That makes aspects of the hidden usefully visible, but casts the rest of it, and the ground that enables it, into shadow.

The light we shine when we make a particular explicit shows a thing named by a word that is shared with others seen in this light. That particular as a thing and word is connected with all others in an endless web of complexity, illuminated by our situated standpoint and what we care about. Our explanations bring into the light the strands of this web we need for our explanations' purposes.

The model of models we are inhaling in the Age of ML shares this topology: a messy web generated from an overwhelming abundance of interrelationships that would have escaped our general conceptual model. The workings of the model often remain as hidden as the earth even a centimeter beneath our feet.

In my unsupported and irresponsible opinion, we are better, true-er people and cultures if we accept what machine learning is showing us by embracing the essential unknowing at the heart of our experience, and the hidden that makes explanations possible.

References

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. *arXiv*, 2 March. https://arxiv.org/pdf/1702.08608.pdf

Douglas, H.E. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science*, *76*(4), 444–463. https://doi.org/10.1086/648111

Fell, J. P. (1979). *Heidegger and Sartre: An Essay on Being and Place*. New York, NY: Columbia University Press. https://doi.org/10.7312/fell91382

Heidegger, M. (2006). *Sein und Zeit* (19th ed.). Tübingen: Niemeyer. (Original work published 1927)

Heidegger, M. (2012). *Der Ursprung des Kunstwerkes*. Frankfurt am Main: Klostermann. (Original work published 1935-6)

[6] This is such a common theme in Heidegger — beginning with *Sein und Zeit*'s idea of truth as un-covering or dis-closure (Heidegger, 2006) — that it's hard to point to a single reference. He develops hiddenness as a central topic in *Der Ursprung des Kunstwerkes* (Heidegger, 2012), but quickly moves away from the Earth-World duality that that work expounds, possibly in part because the essay's language of opposition is too crude. His subsequent notion of the fourfold incorporates the earth into the world (Weinberger 1984).

[7] On the hidden as the ground of Being I heartily recommend Fell (1979).

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 1–45. https://doi.org/10.3390/e23010018

Mittelstadt, B., Wachter, S., Russell, C., & Sutcliffe, D. (2018). Could Counterfactuals Explain Algorithmic Decisions without Opening the Black Box? *Oxford Internet Institute*, 15 January. https://www.oii.ox.ac.uk/news-events/news/could-counterfactuals-explain-algorithmic-decisions-without-opening-the-black-box/

Weinberger, D. (1984). Earth, World and Fourfold. *Tulane Studies in Philosophy*, *32*, 103–109. https://doi.org/10.5840/tulane19843210

David Weinberger – Independent scholar
✉ david@weinberger.org; ↗ https://weinberger.org/writings/

David Weinberger, Ph.D., writes about the effect of technology on our ideas. A former philosophy professor, he entered the technology field 35 years ago and has written numerous books and articles on the topic. He has a long-time affiliation with Harvard's Berkman Klein Center, has been embedded in large and small tech companies, and edits an open access book series for MIT Press.