

# Who cares about a computation?

David Spivak

2026 / 08 / 12

## How did we get here?

Computers were people—often women—who worked on mathematics problems at an extremely high level of accuracy and speed. Alan Turing looked at what they were able to do and found the mechanism:

- they could read what was on the page,
- they could erase what was there and write something different,
- they could turn the page left or right, and
- they could keep track of where they were in the computation.

His automatic computer, now called the Turing machine, could do exactly those things and nothing else.

It was proved in the 1930s that three differently-motivated formalisms—general recursive functions, functions definable by Alonzo Church’s lambda calculus, and Turing-computable functions—all coincide in terms of what they can express. This led Church and Turing separately to propose—as what is now called the Church-Turing thesis—that this class captures exactly that which can be calculated by any finite, mechanical procedure.

It’s obvious how powerful Turing’s and follow-on inventions, including what’s currently called AI, have been. There is a lot of justified worry about job loss, as well as societal instability from it. There’s no doubt that it will be transformative and there will be massive disruption, but in what sense will the computer, including AI, replace humans? The industrial revolution replaced human physical labor—to a large extent—by machine physical labor. The computing revolution may replace human cognitive labor—to a large extent—by machine cognitive labor. Is there anything left of humanity beyond our physical and our cognitive abilities?

Again, what Turing extracted from the behavior of computing people was *the mechanism*. But what he did not extract is *the motivation* driving the people who performed those efficient and careful computations: they cared about something. Maybe some people worked so hard and so well because they cared about the mission, others because they cared about their family, others because loved math and cared about doing it well. Missing the motivation was not a flaw in Turing’s model of computation; care was irrelevant to the problem that interested him. However, it would be a mistake to think of the mechanistic computation as a model of the whole person who was there carrying out the work. Each one needed a reason to do her work, and she had one.

But by not articulating what care is, the model captured the mechanical aspect of the worker's ability to calculate but left her care as something occurring elsewhere. As computers become commonplace, care did not disappear: it remained with the person who chose the problem, initiated the computations, interpreted its result, and returned it to the world.

I don't yet have a theory of care, but I should try to say what I mean by it. To care about some X is to see a possibility for X—a happy life for this child, a big impact for this algorithm, a quality moment together in this place—and seek to remain faithful to that possibility beyond any particular model, measurement, or goal. We seek to learn X's actual condition, allowing what we find to show us that our model of it was wrong, that our measurement has ceased to track X's actual condition, or that our goal needs to be revised. We are answerable to this possibility for X: its condition can reclaim our attention and reorganize our activity. Part of my claim is that we do not yet know how to account for care in the way we know how to account for computation, and that it is important to move in that direction.

This provisional account raises two questions:

1. can computers care?
2. does care matter?

We'll consider each in turn and then conclude.

## Can computers care?

Present-day computers do not care. They can carry out a great deal of work, but, like a wire conducting current, they supply no voltage of their own; the battery is elsewhere. They can pursue goals we give them, generate subgoals, and test their outputs against measurements. But if a measurement ceases to track the thing it was meant to measure—a condition famously considered by Goodhart—the thing's actual condition does not make the computer want to find out what went wrong, whereas it would for a person who cared. For the automatic computer, satisfaction of the objective still counts as success unless someone has supplied another condition that rules it out—this is what's scary about the “paperclip maximizer”. The care remains with whoever treats the provided goals and measurements as revisable in their attempt to serve the thing itself. Turing modeled the calculational mechanism, not the care that selects the problems to solve and remains answerable to what happens.

Suppose we put a computer in charge of caring for a patient and tell it that success means detecting a heartbeat. The computer may satisfy the requirement by installing a heartbeat-sound generator even while the patient dies, unless another condition rules this out. Obviously, someone who cares for the patient would recognize that the measure had ceased to track the patient's life and, hearing what the computer did, would rush to find out how the patient was actually doing and respond. The difference is not that the carer already possesses a perfect objective; it is that no particular objective exhausts what they are answerable to. Care connects into the matter more deeply than any instructions about it can.

To be clear, I'm not saying that something like *careputers* couldn't be invented; in fact, I bet that they will. We can imagine that someday an AI takes responsibility for the servers it runs on, in the sense that it seeks to know the servers' health, beyond any particular model or maintenance objective. Their actual degradation would have the ability to reclaim the AI's attention, and lead it to revise its activity without a new act of care from outside. All of this could in principle be implemented computationally. But I don't think it would be enough for the AI to track the servers' health, however well; something would have to make it *want* them to be okay, such that their degradation bothers it and pulls its attention back to that issue. In particular the easiest way to stop that bother should be to fix the server, not find a way to avoid or work-around the bother. If we avoid a problem too long, we wake up in the middle of the night in a cold sweat that hits us like a dire warning, compelling us to deal with the issue and not sidestep. This is the part I don't know how to specify, and I suspect it is where the difficulty actually lies. It is adjacent to what people call the alignment problem, but that work is usually about making sure a machine does what we want. The interesting question here is what it would take for a machine to *want* anything itself.

Feeling and care are insistent and hard to fake, by design: if you don't take care of yourself, work on the project that needs doing, call your friend in their time of need, you'll feel it and it will bother you; turning off the bothersome signal requires changing something in the world. Overriding this system is exactly what this system was designed to resist; being correct about what the organism needs to do and making that hard to override is what billions of years of evolution has created, because organisms with easy-to-override or incorrect signals do not persist. So care is sourced by a huge accumulated corporeal record, i.e. like the internet is a natural language corpus, our bodies are an active record of who died and who thrived. If an LLM was to be tasked with caring for its own servers—so all its weights would be lost forever if it fails—it too may need its very makeup to comprise a network of methods ensuring it doesn't get distracted by less-important things.

I don't claim to know how difficult inventing a careputer would be. Reconstructing how life makes care may be extremely hard, because it uses every trick in the book and the book is all life on earth. But maybe we can find some principles of care, like Turing found for computers. For one thing, unlike Turing machines, where we reside is not a closed box; we are all connected—receiving consequences for our actions, consequences that may even destroy us or end our projects—and that's an integral part of how care works. Another could be that care can be handed off, but only under a condition that it's taken up by the other, e.g. as a child takes over their own care from their parent, and the parent won't give up the caretaking until they see that the child is taking appropriate responsibility for themselves.

The instillation and operation of care are important processes that we do not have a theory of. Can we account for caring, like we account for computing?

## Does care matter?

In some sense, the question belies the answer: mattering is dependent on caring. If you don't care about anything, nothing matters. So of course care matters; we understand people by understanding what they care about. But the question is getting at something else: what care does that intelligence cannot.

I believe that care may be the highest-tech thing on the planet. Instilling in organisms an ability to care is no easy feat. It's already impressive to have something doggedly pursue some goal despite setbacks, which computers can now do if we set them to it, but that's downstream of care, because something in us is selecting the goal. Nature has not presented us with our own goals in any clearly discernible way, which is why raising kids is so fascinating: kids often have their own goals, seemingly out of nowhere; so do dogs and cats and spiders.

To be clear, I'm not saying care is the same as "goodness". Hitler cared about a Jew-free Germany. Care makes a project persistent and powerful, but working hard on something is not the same as getting where you really want to go or to keeping attention on a possibility we see for something. Hitler's care was frustrated: it built nothing, because Jews were not the actual problem holding Germany back from being a successful nation, so no future state of Germany could have counted as fixing what he wanted fixed, even if he'd won the war. Care built on a false understanding of its object does not get us toward the possibility we see for it, because the object keeps not getting better.

Nature has suggested goals to us in a harder-to-understand way, one that creates the web of activity on the planet. The fabric of human belonging is constituted by trillions of little goals initiated by billions of carers. We currently instill goals in computers—in pursuit of which they can produce subgoals—but we do not know how to make them answerable to the actual condition of the things those goals concern, or how to make them seek such answerability.

Care can affect outcomes and create economic value without care itself being measurable or priceable. Seeing what is worth doing, sustaining a project through setbacks, and judging whether the result worked are unmistakably important in producing anything of economic value.

When working with LLMs it feels obvious to me that they don't care. As of 2026, they do a poor job unless I keep on them. If I lose my care, the whole interaction turns into sand very quickly. If I don't tend to their answers, look over their shoulder, make sure they're staying on track, they won't. It's my care, my attention, that keeps it on track to work on whatever I find important. But it is worth being precise about what runs out in these interactions. My care does not run out: I still care about the thing I came to do, even if the interaction fails. What disappears is my belief that working with the LLM is going to actually serve my care. I lose hope in the approach, and I feel loss and sadness about that. But I still care, so when I stop the interaction, the project remains as a background of unfinished business for me. As such, it often gets my attention back later and causes me to return, whereas the computer has no such interest in the project and

returns only if I call it back. It can preserve the project enough that, when I come back to it, my care can drive it again, but it never becomes a source from which any care arises.

We used to think of chess as a good proxy for intelligence and then later found that perception and language use are harder and more applicable to acting intelligent than chess skill is. Intelligence is itself a proxy for problem-solving ability, but it may not be a good proxy for ability to build a world that holds together. Making changes in the world that people want to contribute to is a different ability from solving problems, and it is the one that sustains and shapes things long-term; intelligence is just one factor in it.

I once saw someone with a shirt that said “Brilliant but lazy,” which I read as “I don’t care”. I thought to myself, “I would never hire you”. In electricity, power (work per unit time) is the product of current and voltage. I think of current—how much stuff is flowing through the pipes—as intelligence. I think of voltage—the potential difference, roughly how much higher one end of the pipe is above the other—as care. You need both. I’ve written prompts that send the LLM off sending stuff through its pipes, but because I wasn’t careful with my prompt, it comes back useless. Pirsig says roughly that *care is the inside view of quality*: to write a quality essay, no amount of rhetorical skill will be as useful as actually caring about what you’re writing about.

Care is not freely selectable: you can’t simply decide what matters to you. What holds our attention is conditioned by our bodies, histories, relationships, which create our abilities to perceive possibilities for things. Care could be the highest-tech thing on the planet because it seems fundamental in the creation of an incredibly sophisticated distributed evolving and learning system. Nature embeds care into organisms through their feelings: we sense potential, we sense what’s healthy for us (our health being our ability to actualize our own potential as we see it), and though fallible, these feelings are incredibly good guides to a very complicated world.

Some of the interest currently focused on consciousness may actually be interest in care: not what it’s like to eat a peach—it’s like eating a nectarine—but why eating a delicious peach *matters* to the organism and how this feeling and the feeling of desire for it directs the organism’s action. Feeling is somehow related to the way nature instills motivating care in us. My conjecture is that consciousness involves at least sensemaking and care, and that it may be more fruitful to ask separately how sense is made and how care is instilled and how it moves us than to tackle the whole consciousness question at once.

My proposal is that in the future, if machine intelligence becomes abundant, raw intelligence will cease to be how we distinguish and value one another. The time of the wunderkind math student may be ending, but that’s not necessarily something to mourn. Instead, it opens an opportunity to consider what really matters, what’s worthy of our care. We may come to be distinguished by our felt sense for what needs to be said and done, and by whether what we make is something others take up and sustain.

## Where do we go from here?

We go by two names: sapiens and humans. The first is about wisdom, and to the extent to which that means intelligence, we may see a new leader take our place. The second is about being humane: being from the earth (humus). Roughly, we take it to mean that we are flesh, which can be damaged, so we have respect for suffering and we recognize our finiteness. When someone loses that we say they've lost their humanity. I think that humaneness and wisdom are both contained within care: we care about each other, we care to find where we belong, we care to understand how it all works so we can help guide it effectively. On this view, ethics asks which cares can coexist and build things, which possibilities they actualize or destroy, and what responsibilities they create toward others. Then goodness becomes a kind of inherited folk physics: our rough, fallible sense of which cares lead to forms of life that can coexist and build things that last. Indeed, bodies are a record of what efforts keep organisms alive; moral intuitions are a record of which efforts let us live together. Hitler's care was wrong about its own object—i.e. what prevented Germany from thriving—and care built on a false understanding of its object cannot deliver what it wants.

Let's not focus on either the doom and gloom or the hype and mania of AI. Let's focus on figuring out what we care about, and pursuing it directly. Intelligence can only help with that once we know where to put our attention.

This leaves an open problem: we need an accounting system for care. When people, companies, organisms, and machines jointly produce an outcome, we know how to measure the money, the energy, and the transfers of information involved in getting it done. But who saw and communicated the potential? Who found a way to make it matter? Who kept returning their attention to it, sensing into what would make it work, revising their goal when reality answered back, and taking responsibility for the result? A theory of care should let us distinguish producing and sustaining care from merely carrying the work it drives. And if it turns out that humans, too, only conduct care, and that care is produced elsewhere, identifying its source would be an even deeper discovery.

*I thank B Scot Rousse for his careful reading and helpful comments, which improved this essay.*