A Model of Human Thought: Philosophy

Tyler Neylon

216.2018

What does it mean for something to be true?

What can we do to verify that something is true?

How can we discover new truths?

This article is about a conceptual model of truth that may help answer these questions. I'm personally motivated to consider these questions because I'd like to re-evaluate the scientific method — the search for truth.

I'll argue that this line of pursuit aligns with an understanding of human thought itself. As a result, these questions also pertain to another deep interest of mine, which is writing software that we would recognize as a person.

1 A traditional view

The usual view of truth is that an idea is true exactly when it corresponds to the actual state of the world. A concept you need to bundle along with any definition of truth is a concept of what kinds of things are candidates for truth; previous philosophers have argued that suitable *sentences* or *propositions* are good candidates for truth. The word *suitable* here refers to the need to avoid tricky cases such as *"This sentence is false."* I won't dwell on the tricky cases because I think they are mostly a distraction from the meat of the issue.

The reason I'm not a fan of this definition is that it doesn't really explain much. It doesn't tell us how we can verify truthiness. We're left asking what it means to correspond to the state of the world. It seems to simply exchange one word for a few more without helping us learn anything new in the process.

Rather than arguing that this semi-tautological perspective is *wrong*, I'll simply provide another that I think is more useful. Consider it the difference between Roman numerals and our usual Arabic base-10 decimal notation. It's harder to do long division with Roman numerals, whereas the structure of a number captured by base-10 notation is *useful* — it enables us to take certain actions that might be more difficult with other notations, even though it adds no new

information. This is the kind of *usefulness* I'm arguing exists in an alternative model of truth.

2 A practical view

I won't be so bold as to redefine truth itself, but I will introduce a phrase to capture an idea that I think we all implicitly use without quite naming it. This is the idea of *effective truth*; when I say something is *effectively true*, I mean that it tends to help us achieve a goal we have in mind.

An easy example to consider would be to say that a certain book at a certain store costs $10. We could achieve our goal of buying that book by walking into that store with a $10 bill and paying for that book, ignoring tax and pretending that people still use paper money.

A more interesting example is to understand the uses and limits of our own internal model of physics. Suppose you take a cylinder of wood and roll it on a flat surface. You can reasonably expect it to move smoothly, gradually slowing down. Imagine doing the same thing with an empty glass jar — same result. In fact, most solid cylindrical objects will act the same, so we have a useful model for predicting how objects will move on their own based on this.

Now imagine someone gives you a jar half-filled with water, and you also roll it on a flat surface. If you've never done anything like this before, then you might be surprised by how uneven the motion is. Even if you've done this before, you'll find that you have much less certainty about how it will move than in the other cases. The point is that often our ideas about how the world work are useful simplifications. In this case I'd say the simplification is the idea that "most cylinders act as if they were solid and had uniform density."

You might respond that this seems like an arbitrary example that is a simplification, whereas other ideas are more clearly true or false. I'd respond by saying that, as far as we know, all of our ideas of parts of the world are models rather than exact representations. To return for a moment to the $10 book, suppose the store burns down before you get there. Would you say it still costs $10 at that store? How do you know they won't raise their prices? What if there was a computer error in their pricing system, and half of people were charged $10 and the other half were charged $12? These may feel like edge cases, but the point itself is that unclear edge cases *always* exist.

If you concede that human ideas are imperfect models that help us achieve our goals, then you can see how it's useful to speak about effective truths rather than to insist on an idealized version of truth that could only make perfect sense if it knew all the rules of physics to the smallest detail.

2.1 Relevance and social truths

I want to briefly mention two other potential essential properties of truth.

One is that the *relevance* of an idea seems to be important for that idea to be discussed as true or false. Suppose a certain particle of dust in a far-off galaxy may or may not one day collide with another certain particle of dust. The point of this example is that no one cares. Neither outcome will have any bearing on any human.

The traditional model of truth applies equally well to these particles of dust as it does to the more pertinent question of whether or not an asteroid will collide with Earth and destroy our species. I won't argue that this is a downfall of the traditional model, but rather that how humans think is closer to effective truth because we don't bother with things that don't matter. From a certain point of view we basically *can't* think about completely irrelevant ideas because we simply don't see them.

The other potential essential property I wanted to mention is *social truth,* by which I mean the sense that some ideas are accepted or not based on large-scale social behaviors. For example, what kind of clothes are currently in fashion? While some specific people have more influence on this than others, this kind of truth is ultimately up to the actions of the many. Another example of social truth are questions of widely-known fictional characters such as Santa Claus. It's generally considered true that Santa Claus lives at the North Pole, although there is no such person. In cases like this, a story has evolved over time, perhaps influenced more strongly by some individuals than others, but the story takes on meaning through its collective acceptance as a kind of truth.

Similar to relevance, the idea of social acceptance can be viewed as indirectly meaningful through the lens of effective truth. For example, if you want to understand the advertising around, or integrate your children into a world of a Santa Claus-celebrating society, then it's useful to be aware of the mythology behind Santa Claus. In general, social truths appear to be ideas that tend to build on top of, or further enable, connection-building and status-building. Knowing these truths is effective at achieving social integration.

2.2 Shades of truthiness

From here on, let's work within the view that an idea is true exactly when it's effective. An interesting result of this view is that truth is no longer a dichotomy. Instead, just as some ideas are better at helping us achieve a goal than others, there are similarly many degrees of the truth of an idea.

One example is the idea that a trip from your apartment in Manhattan to Penn Station takes 20 minutes. This is a simple idea that, in practice, may often be true. But there are many factors that may need to be taken into consideration to

make this idea more accurate. Traffic is worse during rush hour. If the subway is not running normally, you may need to find an alternative route. The availability of trains and cars decreases on weekends. If you're aware of the many moving pieces, then you're in a better position to get to Penn Station on time. It's not quite *false* that you live 20 minutes away, so much as it is *useful* to act as if that were true; and in fact even *more useful* to keep in mind modifications that allow you to plan your trip more resiliantly.

2.3 Truth itself as a useful fiction

So if we move away from the dichotomy of truth in favor of thinking in terms of what's effective, then what becomes of truth itself? Do we still need it?

I suggest that, in practice, we already use effectiveness as the building blocks for how we decide to do what we do. Perhaps the idea of truth itself has come in handy as a counterpoint to the technique of lying.

Imagine living in a world without communication. Then the need to distinguish a truth from a falsehood is far from your thoughts. You still can learn and accomplish things by experimenting and thinking. In other words, you already have a model for how the world works without need for an explicit concept of truth.

I'll give another example to illustrate the concept of a *useful fiction*. Consider the center of gravity of an object. It is not a part of reality — nor do we even pretend it's a physical object. At the same time, it's useful to think in terms of a center of gravity in order to model how things behave physically. Many useful concepts can be seen from this angle — the idea of ownership, money, responsibility.

What unites these ideas is that they help us get things done without having an obvious physical counterpoint. We've *added* a new idea to the world, something that we think in terms of, that is entirely internal, and that is useful to us. (Again I'll suggest that we are not even really capable of thinking about things that are entirely useless to us.)

When it comes to truth, there's no point and counterpoint in reality to give meaning to the term. The world is only itself. If it has no falsehood, how could it show us truth?

So we see that the idea of truth must be internal — something we need to invent. And in fact, it is less fundamental than simply getting done what we want to get done.

2.4 Isn't it true even if no one knows it?

At this point I can anticipate a philosophical argument against effective truth as a more fundamental idea than truth itself. You might ask about an event that no one was aware of — is it not true that this event still happened? In that setting, how is truth an internal idea?

But this line of questioning is misguided. My claim is not about what is false or true. Rather, the argument of this article is that it's *more useful* to model truth on effectiveness than it is to place our first principle on a correspondence with reality. In other words, I'm never disputing the truth or falsity of any particular event; I'm not disagreeing with the traditional concept of truth, rather I'm seeing it as a higher-level concept that we've invented in a world that already contained the precepts of thought beforehand.

2.5 Charles Pierce and Karl Popper

I wish I were better versed in the relevant schools of philosophy before writing this so I could explain exactly how this line of thought fits in historically. I don't think I'm an expert at the history around these ideas, but I will mention a small number of prior ideas and how they relate.

One of the most relevant set of ideas is that of *pragmatism* as championed by Charles Pierce in the early 20th century CHECK. Pierce's perspective is typically summarized as saying that an idea is true exactly when it is *useful*. I have quite deliberately gone in a subtly but critically important direction by replacing the word *useful* by the word *effective*, and I'll explain the distinction I have in mind.

A brief but compelling argument against the idea of truth as usefulness is this: It may be useful to think of my spouse as faithful, but that certainly doesn't make it true. I suspect Pierce himself would have disagreed with this in much the same way I will, though I can't be certain of that; so I'll present my reply as my own.

There are two trouble spots with this argument. The first is almost a distraction, but is important: It's the idea that we can treat belief as a choice. Do I choose what I think is true? I don't think so at all. We can choose to act *as if* one or the other thing is true, but even in acting so, we don't really change our mind unless we receive more evidence one way or the other (which evidence, admittedly, may be based on an internal revelation).

The second — and I think more important — reply is to keep in mind what something is useful *for*. If pretending your spouse is faithful is useful *for* the sake of your peace of mind, then the argument falls through as belief is not a choice; it would be like saying that being confident is useful for being confident. On the other hand, if pretending your spouse is faithful is useful for ensuring that the only children either of you parent is between the two of you, then clearly it's *not*

useful to simply "think of" your spouse as faithful. Your action of thinking this has essentially no influence over the child-centered outcome you want to achieve.

Whatever the case, you need to look at the goals in order to evaluate the effectiveness of the decisions made to reach those goals. The correspondence between traditional truth and effective truth is that both will help you achieve your goals. To be clear, I don't claim that there is any great difference between pragmatism and effective truth, but I suspect I'm providing a different point of view by allowing effective and traditional truth to live in the same world as slightly different ideas, and to build a model of thought around these. I prefer the word *effective* over *useful* simply because *effectiveness* more strongly emphasizes that there is a goal in mind, and that this goal is brought closer by ideas that are effective.

As a brief note, Karl Popper argued in favor of an idea needing to be *falsifiable* in order to qualify as a candidate for truth. I agree that unfalsifiable ideas are not effective ones — in fact, I would go even further to say that they are not even *meaningful*. Consider an unfalsifiable idea. No matter how we act on it, by definition, nothing in the world will change either way — otherwise it would be falsifiable. Because it makes no difference to the world, it can't be an effective means of achieving any goal. Similarly, if an idea has the capacity to be effective, then we can try to work toward a goal with it, and how well we do moving toward that goal provides a kind of falsifiability. Not to say that these ideas are the same, but rather that they are compatible in that both Popper and the idea of effective truth can use falsifiability as a test to throw out particularly unhelpful ideas.

3 A model of thought

The bulk of this article focuses on truth, but these thoughts are even more interesting in the framework of a model for how humans think. This is probably a line of thought worthy of its own article, but I wanted to include an overview here to add some fun context.

The framework is this: Humans have goals, and mental building blocks of actions they can take to help achieve those goals. Effective truths are these building blocks, and the way we put them together is like an algorithm built from subroutines. Put another way, if we're thinking about a plan of action, then we can see effective truths as if-then clauses, and a chain of them as a kind of logical argument (albeit non-boolean since we have shades of truth) which begins with the state of the world as it is now, and ends with the state of the world as we'd like it to be.

Communication is such a critical aspect of thought and action today that it deserves special attention. When we express an idea to another person, we're trying to do something useful for both of us by working within a framework

where we each have goals. *Statements* are meaningful in that they are effective by distinguishing among multiple possible states of the world. In other words, they answer a question, and their meaning only truly makes sense in the context of both the goal and the question they imply.

Although I've been very brief in this section, you can begin to see how these ideas might become part of a model for machine intelligence.

4 A bigger picture

Descarte, arguing from first principles, suggested that

(philosophy as humanity's relationship with the world, not about the world itself)

(David Hume, Descartes)

(aim to close with the big picture idea and some argument as to why)