Notes on Truth

Tyler Neylon

This printing is v0.3 of this book.
It was compiled on 512.2022 (September 12th, 2022) in Berkeley, California; edited by Sean Dennis.

The TEX source and sources for the original images in this book can be found at `https://github.com/tylerneylon/truth`.

This book uses the font Gentium Plus.
The cover image was generated by OpenAI's DALL•E tool.

If you notice any errors in this book, please email me about them at `errata@tylerneylon.com`. Thank you in advance!

This little book explores the question: *What is truth?*

This question is made of many others. What makes some things true and others false, especially where the distinction isn't obvious? What kinds of things *can* be true or false? Can any sentence be true? Any thought? Can fictional ideas be true? How objective is truth, really?

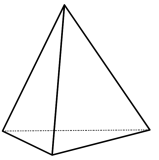
These questions are both philosophically interesting and useful in expanding the frontiers of science — science being the empirical study of what's *true* about the world.

The exploration of these questions begins with curiosity, then meanders through ideas of invention, omniscience, and the foundations of knowledge. Ten observations serve as waypoints along the path. I do not believe that perfect, final answers are possible here. But I will attempt to shed light into the shadows of uncertainty along our way, completing the trek with a specific analysis on the nature of truth.

At the heart of this investigation is the idea that truth isn't as objective and clear-cut as we might assume. It is in fact a profound concept, reflecting the nuanced nature of our world, and of our own minds.

A Map of Our Journey

Chapter		*Key Question*
Pandora's Curiosity		*What is truth?*
Definitions of Truth		*How can we describe truth?*
Inventions		*Is truth discovered?*
What Ideas Are		*What is truth made of?*
Ignorance and Omniscience		*What does truth depend on?*
The Fuzzy Edges of Thought		*How objective is truth?*
The Foundation of Human Knowledge		*How do we relate to truth?*



1

Pandora's Curiosity

I'll start by motivating our investigation of truth. Let's see how this exploration can be useful.

In theory, there's a reliable way to learn about the world called *the scientific method*. As a rough summary, the method begins with the statement of a hypothesis as well as an alternative to test against, and is completed with the collection and analysis of evidence to distinguish between the two.

In practice, there are often differences between this and the way researchers actually go about their work. In some cases, you might see a small change, such as data being collected before a hypothesis was made. In other cases, you might see entirely new approaches, outside the scientific method, for finding something out. For example, a sound mathematical proof is a different and conclusive way to argue that an idea is true. Such a logical proof is all-encompassing: The hypothesis — now a theorem — is precise and final, impervious to the need for later revision. The scientific method, by comparison, is revealed to be a process of guessing, collecting imperfect data, and doing our best to connect that noisy data with our guesses.

Beyond the awareness that the scientific method is subject to error and uncertainty, we also see that some questions don't fit cleanly within its form. For example, we may find two conflicting theories of physics which each explain certain observations, yet disagree in other ways. We must decide which theory we believe. One might argue that the imperfections in one theory are worse than the imperfections in another. One might argue that one theory is more elegant than another. These conversations aren't well captured by the traditional scientific method, yet they're common occurrences in knowledge-expanding communities, whether the

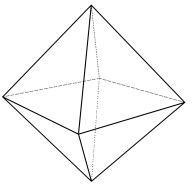
field is physics, computer science, or even, giving ourselves some wiggle room, among chefs exploring theories of gastronomical excellence.

All of this is part of another big question:

Question. *How can we learn what's true?*

This question can help us evolve how we learn. It helps us to better understand and to build on what we call our scientific method. It may even be the ultimate practical question. However, I won't try to answer this question in this book, though I'll make observations that may serve as the basis for its answer. Rather, the above query serves as a key motivator for our real focus — the nature of truth itself. We must understand truth before we can understand the discovery of truth.

To connect the dots: When you open the door toward answering *How can we learn what's true?*, you find the Pandora's box that is *What is truth?* This is the box we'll open.



2

Definitions of Truth

2.1 Correspondence Truth

Truth is a concept so fundamental to human thinking that it's hard to define in simpler terms. Perhaps the most traditional approach is an old idea called *correspondence theory*, first mentioned, as far as I'm aware, in Aristotle's *Metaphysics*:

Definition (Correspondence Truth).
An idea is true when it corresponds to reality.

I don't think this is a good definition.

Pretend we're defining the mathematical idea of a *number*, and we said "a number is an element of the set of numbers." This has a lot in common with the correspondence definition of truth. Specifically, this definition of a *number*:

- Relies on another concept (*the set of numbers*) which is more complicated than the thing being defined.
- Doesn't add much intuition about the thing being defined.
- Isn't easily testable. That is, we don't have a nice way to test if a thing is a number or not, in the context of not yet knowing about the *set of numbers*.

On a metalevel, there's something interesting about our quest for a definition of truth — we're looking for the definition of an idea that we already intuitively understand. How will we know when we've succeeded? In essence: When the definition appeals to our intuition. Specifically, for any examples we have of intuitively-understood true or false ideas, those examples should be considered correspondingly true or false by our proposed definition.

This process of finding a definition is a case of knowledge discovery which does not fit well with the traditional scientific method — although we could, with some creativity, treat intuitive examples as experiments or observations, and think of a definition as a hypothesis that we're testing.

With that in mind, let's make an empirical observation in the form of a *claim*, something that we think of as a candidate for being true or false:

Claim. *If Plato were alive today, he would like pizza more than sushi.*

In this book, I'll use the words "claim" and "idea" as synonyms for things that could be true or false. (These terms, and a few others that I use in specific ways, are listed in a short glossary at the end of this book.) Some philosophers have thought carefully about exactly what kind of thing can be true or false, but that isn't the focus of this book. I like the words I've chosen (idea/claim) because they don't strongly imply that what's being said must be correct; it's reasonable to call "1+1=3" a *claim* or *idea*, and to see that it's false.

The correspondence definition of truth fails to shed light on whether the above claim about Plato's food preferences is true or false. The fact is that Plato isn't alive today, and — for all practical purposes — we have no way to determine what kinds of modern food he would most like. What we could do is make educated guesses, and try to convince each other that one of these guesses is more likely to be correct. This method is only loosely connected with an examination of reality, such as by discussing historical evidence of ancient Greek diets.

I'm not denying that truth is a correspondence with reality. Rather, I'm saying that we might find an improved definition of truth if we take a careful look at how we decide something is true or false. This is analogous to complaining that a definition such as "a number is an element of the set of numbers" can be correct but unhelpful.

Learning from the faults in the definition of correspondence truth, we see that a good definition:

- Relies on prior concepts.
- Adds intuition.
- Is testable.

The last bullet here means that we ought to be able to easily answer the question *Does this thing meet the definition?* When a kid wants to ride a roller-coaster, it's

easy to see if they meet the definition of "tall enough" because they can stand next to a sign. If a theme park defined "tall enough" as "able to comfortably look an adult emu in the eye," then this definition would be more difficult to test, despite being intuitively clear and relying on known concepts.

2.2 Different Kinds of Truths

Since correspondence truth has not given us a great definition of truth yet, let's continue the search by looking at *how* we decide something is true. In this section I'll look at definitions of true ideas — or subsets of true ideas — classified by *methods of deciding* whether you believe them to be true. These definitions tend to be more intuitively useful and easier to test than the correspondence definition.

I'll start with a definition that's implied in the communication of any advanced mathematics:

Definition (Mathematical Truth). *A mathematical idea is true when we can provide a logically sound proof for it.*

As a brief reminder, a *sound proof* (sometimes called a *sound argument*) is a series of logical statements in which the conclusion must logically follow from the premises, and in which the premises are true in the context of the statement being proven.

The math-loving part of me would like to say that a proven mathematical idea achieves an ideal level of truthiness. But in practice, this isn't quite right. There are a number of reasons why mathematical ideas don't typically achieve "perfect truth:"

- Math relies fundamentally on axioms being true, but those axioms are not proven. They are presented as self-evident, and in practice they have subtle and non-trivial consequences. One example is a geometric axiom called Euclid's parallel postulate, which states that, given a line and a point, there is a unique line through the point parallel to the line. If we assume this is true, then we can arrive at laws of geometry on a flat surface. If we assume this is false, then we can arrive at *equally valid* laws of geometry on non-flat surfaces, such as the surface of a sphere. The point here is to show that, as much as we'd like axioms to be self-evident, they aren't always so.

- Virtually all mathematical proofs are not fully formal arguments. That is, math literature is written for human consumption, and consists of largely natural language persuasion, as opposed to a computationally-verifiable proof. In practice, we sometimes find mistakes or omissions in these proofs, and they can be quite subtle. For example, mathematicians have famously disagreed about the correctness of a particular proof by Camille Jordan about a theorem that's now known as the Jordan curve theorem. When experts

disagree about the soundness of a proof, it reveals the difficulty in deciding whether a proof is truly correct!

- Finally, even the rules of logic themselves are subject to debate. If we truly want to assume nothing, then it would be good to *know* we are using the correct rules of logic, rather than to assume them. You may feel secure that our time-tested rules of logic make sense — but there's good reason to question some proof techniques.

  I'll focus, as an example, on proof by contradiction. A proof by contradiction only makes sense if we know the axioms being used do not contradict each other. When axioms don't contain a contradiction, they're called *consistent*.

  Here's the catch: mathematicians can't prove that some key axioms are consistent — at least not until they rely on *new axioms* for such a proof. But when you add a new axiom, you no longer know that the larger set of axioms is consistent! What we'd love is a single set of axioms A where we can prove, using only the axioms of A, that our axioms are consistent. That way we can simultaneously believe all the proofs of A, as well as the proof that A is consistent. Unfortunately, this is logically impossible for modern number systems.[1]

  This means that the axioms we use for modern number systems are not fully proven to be consistent. Most mathematicians believe they're consistent, but — just as we must accept an axiom without a proof — we must likewise choose to accept the *consistency* of axioms of number systems. Tying this back to logic: Mathematicians accept proofs by contradiction, but they do so by *choosing to accept* this proof technique, rather than *knowing with logical certainty* that this technique is always valid.

Surprisingly, even when we strive for an ideal form of mathematical certainty in the truth of a statement, there's a great deal of uncertainty. There is subjectivity in our choice of axioms, in our assessment of an argument's correctness, and in our belief in rules of logic. Most mathematicians perceive a well-known proof as complete and unassailable, but the reality is that a human-written and human-read proof (vs a computationally-verified proof) is just as much a natural language argument as is a lawyer's closing statement in a courtroom. The only difference is that the audience has a higher — but still ultimately subjective — standard for what will convince them.

There is uncertainty not only for mathematical truth, but for every other kind of truth we'll consider, as I'll argue throughout the rest of this section.

[1] For the curious: I'm referring to Gödel's second incompleteness theorem as applied to Peano arithmetic.

I'll phrase this as:

Observation 1. *For every idea presented as a truth, there is a reason given to believe the idea is true. When we examine these reasons, we find uncertainty.*

I'm not saying the ideas are always incorrect. I am saying that when we think we *know* something is true, it's more honest to say that we *guess* it is true. I'll take another look at the difference between knowing and guessing in Chapter 7 below.

If math is the most pure form of *abstract* truth I can imagine, then the most pure form of *practical* truth comes from physics; I mean ideas justified by repeatable experiments:

Definition (Verifiable Truth). *A verifiable idea is true when we can test the idea, and it passes the test. Such a test is an action whose outcome can either support or refute the idea.*

As an example, consider:

Claim. *Water boils at 100°C.*

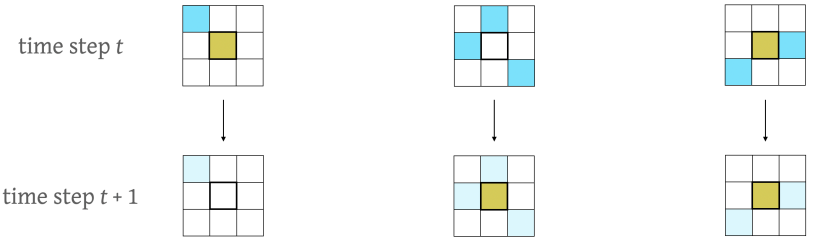
We can boil some water and measure its temperature to test this. Of course, if we were to try this experiment in a setting with high or low air pressure, we would find the boiling point to be slightly different. It turns out that there are other variables to account for (like air pressure) in repeating an experiment — variables that we may not be aware of.

In considering mathematical truth, we found many sources of uncertainty. Do all verifiable truths also contain uncertainty?

Let's try to imagine a specific verifiable claim, one where we can account for all the relevant variables. If we can find such a claim, then we could get a result that would always, without exception, be consistent with the claim. If the result is based purely in logic, then we're back to our concerns with mathematical truth, so this is only a new idea if we think about physical experiments.

Let's imagine that one day we completely understand the laws of physics. Further, let's imagine that there turn out to be quite simple rules, and that all of the physical principles we use in practice, such as Newton's laws of mechanics, are emergent properties of the simple rules. To solidify this framework, we can use an existing set of simple rules called Conway's Game of Life, named for its inventor, John Conway.

The Game of Life takes place in a grid of square cells, each of which is either dead or alive at any given point in time. The grid evolves over time. Each cell's status (alive or dead) is determined by the status of its 8 surrounding cells from the previous time step. The following illustration fully describes all of the evolution rules.



A cell is **dead** if it had 0-1 or 4-8 live neighbors.

A cell is **alive** if it had 3 live neighbors.

A cell **keeps its status** if it had 2 live neighbors— it remains alive or remains dead.

These rules apply to every cell on every time step.

In such a world, we can make a strong physics-based claim, such as:

Claim. *Living cell x will remain alive from one time step to the next when 2 or 3 neighboring cells are alive in the previous time step.*

How could such a claim have any uncertainty?

Unfortunately, it can. Even if we one day discover laws of physics that explain every experimental result throughout all of human experience, we're still making two fundamental assumptions:

- We're assuming that any laws of physics exist at all.
- We're assuming not only that laws of physics exist, but that they remain unchanged throughout all of time.

As much as we may *believe* these assumptions are true, we do not truly *know*, with certainty, that they must be true.

However much we understand the present, we know nothing about the future with certainty.

The idea of verifiable truth is so general that we could consider using it as *the* definition of truth. But there are some truths that don't clearly correspond with this definition. For example, mathematicans don't accept a theorem as true no matter how many times you test it and find it to be true — they require more than repeated experimentation. And there are other interesting ideas we call true which don't fit so cleanly under the umbrella of verifiability. Consider the claim:

Claim. *Han shot first.*

This is in reference to Han Solo's encounter with a bounty hunter named Greedo in Mos Eisley, as depicted in the 1977 film *Star Wars*. In the original version of the film, Han Solo clearly shot Greedo before Greedo had a chance to fire at Han, but in

later, edited, versions of the film, Greedo either shot at Han first or nearly simultaneously. So the truth of the claim is not obvious.

I don't think the claim that Han shot first can be verified to the degree we can verify that water boils at 100°C. It's about a fictional event in a widely-known story. For novel-based questions, we could simply ask the writer since they have a kind of authority over what "really happened" in the story they created. But in this case, the story of *Star Wars* has effectively graduated to a level of American mythology, in which case the authorship of a single living person (George Lucas, in this case) is less meaningful. The ambiguity of ownership causes some ambiguity of truth; this shows that our belief in some ideas depends on an agreed-upon creator.

Let's call the claim that Han shot first an *authored idea* — it's an idea whose truth depends on decisions by an author. This applies to ideas from fictional stories with an agreed-upon creator, but it also applies to things such as laws or standards, which are often decided by groups of specialists. For example, on the web, the default mapping between characters and numbers used to represent those characters is defined by a standard called Unicode. The Unicode Consortium is an organization that's recognized as deciding all the details of Unicode. While no person or group can decide the value of π, the Unicode Consortium can decide which emoji are included in Unicode, and therefore available on your phone. Shared awareness of decisions made about these abstract notions provides a kind of truth to ideas about emojis, or about Luke Skywalker.

Let's officially define this new kind of truth:

Definition (Authoritative Truth). *An authored idea is true when the party recognized as the authority confirms the idea.*

We could ask if Harry Potter likes cilantro. This question is not addressed in any of the writings of author J.K. Rowling. However, if she were to publicly declare the answer one way or another, it would be accepted as canonically true. Just as much as we accept the statement "Darth Vader is Luke's father," we would also accept "Harry Potter loathes cilantro."

Ideas about fictional entities may feel less real, but there's still a feeling of rightness or wrongness to them. If a kid asked you where Santa Claus lives, you would tell them he lives at the North Pole. There is a shared narrative here; if you were to say he lives at the South Pole, this would feel incorrect.

While fictional stories are not about things that happened in reality, when we talk about these stories, we're still talking about actual events. The telling of the story, the experiencing of the story, and our discussion and thoughts of the story are all real events. If the story has enough appeal, then it becomes something greater than a single telling. Stories, to humans, are a part of our experience and our learning in terms of what can happen, how people behave, and how we learn about life. Our own thoughts and actions evolve as we experience narratives, whether they are our personal experiences, those we know of our community, or those stories we hear indirectly.

For example, we generally do not learn that murder is bad by committing or witnessing murder. Instead, we learn about it as we grow. In some cases, we may hear stories of loss, and at some point, vicariously experience some pain of loss that helps us to appreciate the harm of taking a life.

To me, moral ideas never feel black and white, but so many people consider them to be such that I'll include this kind of truth:

Definition (Moral Truth). *A moral guideline is true when a society following it is better off than a society that ignores it.*

There's a lot in that definition. I've chosen to work with a variation of Immanuel Kant's categorical imperative — his style of the golden rule. It's quite a rabbit hole to consider this definition carefully, so for now I'll ask you to accept that we're not focusing on morality, but rather noticing that if I say "murder is wrong," this is an idea which we may say is true or false, and which is more clearly a *moral truth* than the other kinds of truth in this book.

There is yet another kind of truth adjacent to morality. Consider the claim:

Claim. *Forks go on the left side of your plate.*

There is a sense of agreement about this claim, and yet it is clearly not a result of a logical argument, nor of a physical experiment. Neither is it a moral directive since no one seriously suffers if a fork is placed to the right of a plate. There are other truths not far from this kind, such as the convention of wearing certain kinds of outfits to a wedding or a funeral.

It's interesting that these conventions are not understood as a single *correct* way to do things, but rather they represent a *consistency* within a group of people. In the case of utensil etiquette, the consistency provides an expression of cultural status and awareness; proper wedding attire conveys respect for the event and the people getting married. What makes a convention correct is an understanding of what most people expect. In other words, the truth of these things is based on noticing what the majority already treats as the correct decision:

Definition (Democratic Truth). *A social convention is true when the vast majority of a society consider it to be true.*

It's not always obvious how democratic truths form. Did a monarch one day see a fork on the right of her plate, felt funny about it, and decapitate her table-setting staff members? Perhaps from that day forth, forks were carefully placed on the left.

This possible fork-on-the-left origin story is an example of an idea being decided by a single individual. In some rare cases, it does seem as if a single point of history determines a convention. Consider the sandwich. The word *sandwich*, referring to food, is distinctly traceable to John Montagu, the 4^{th} Earl of Sandwich. The story, provided by plausible historical accounts, is that he liked to eat quickly

and cleanly while either gambling or working (depending on your source), and a bread-enclosed meal did the trick.

More often, it appears that conventions evolve slowly. Rather than arising from a single conscious decision, they appear to be an accumulation of natural smaller steps, with a reason behind each step. These might not always be *good* reasons, but when you examine the history of a concept, you often find more sense than you might expect. Newts were once ewts; when people saw one they would tell someone they saw "an ewt," which was so easy to confuse with "a newt" that the word changed. People once considered the Mediterranean Sea to be the sea in the middle (medi) of the land, or Earth (terra).

In the examples above, we're seeing ideas, like words, that live in multitudes among communities — such as French words in France, or mathematical lingo among mathematicians. These ideas shift and adapt as the world changes. Some ideas receive more attention while others, apparently less useful, dwindle. This brings us to another perspective on truth that's worth consideration in its own right:

Definition (Evolutionary Truth). *A communal idea is true when the persistence of that idea corresponds with the persistence of the community.*

Peculiarly, this definition doesn't directly convince us that an idea corresponds to reality. This may irk your intuition. Yet, by this definition, every good idea is evolutionarily true. For example, if one community believes in germs, and another doesn't, then over time the germ-believing community is more likely to survive, and the concept of germs with them. Ideas are the social genes of the community. They mutate and change over time, and the theory of natural selection applies to ideas just as they do to traditional genes in a species.

At the same time that good ideas tend to be evolutionary truths, there's room for other ideas to tag along for the ride. While good ideas are evolutionary truths because they promote survival, other ideas may count as evolutionary truths simply because they are good at keeping themselves alive, rather than keeping the community alive. Some beliefs of organized religion seem to fit into this category. Religion historically helped communities by encouraging them to work together, to help each other out, and to remain organized. Those are genuine benefits. Along with those benefits came ideas that do not correspond with reality, nor do they confer verifiability or utility in practical decisions. One example is the belief that Zeus exists as a god, and is the son of the titans Cronus and Rhea.

You might say that these religious beliefs are authoritative truths. But believers are likely to disagree, saying that the authority of a religion derives from it being true, and not the other way around.

Just as there's uncertainty in the previous kinds of truth, there's uncertainty here. Indeed, there are many incompatible religious beliefs, so they can't all be correct at once. In other words, it's perfectly possible for an idea to persist in a community even when that idea doesn't meet our intuition for being true.

2.3 Which definition is right?

At this point, I've presented several definitions for kinds of truths. My goal has been to follow a natural path through the jungle of reasons to believe things, using example claims as stepping stones. The definitions are not yet organized, but rather a collection that covers the categories I'm aware of. We can add a bit more organization by asking:

Question. *Is there a single all-encompassing definition for truth?*

In other words, is one of our definitions the *main* definition, with the others describing subsets of truths? Just as we can provide multiple definitions of an English word, I think there is more than one valid definition of truth. Even in mathematics, we can sometimes define a technical concept in different ways, and those ways end up being equivalent after some analysis. Accordingly, I'll provide more than one answer to which definition of truth is the one we're looking for.

I'll start by arguing in favor of this answer:

Answer A. *When we seek to describe the truths that groups of people tend to believe, we're talking about* ***evolutionary truth****.*

This answer is almost tautological in the sense that I'm aligning the context of the answer with the definition of evolutionary truth. Still, there's an interesting thought here: We're taking the vast wealth of work on biological evolution and seeing that it applies to what we accept as true. Proponents of this notion include biologists such as W.D. Hamilton and Richard Dawkins, who consider evolution to apply to ideas in social settings.

Based on this perspective, humans may deserve less credit than we tend to give ourselves for having many brilliant ideas. Just as evolution created life, flight, eyeballs and brains without individual insights, perhaps the simple mechanism of billions of people guessing and checking many possibilities deserves some credit for human innovation.

Among our definitions, evolutionary truth stands out because it describes ideas that are believed, yet it doesn't care about individuals' justifications for the ideas. This distinction is similar to the difference between language that is correct in theory, such as the grammatically correct phrase "that's neat," versus language that is used in practice, such as "it's lit fam," having roughly the same meaning. When we have an opinion about the correctness of something (such as the proper linguistic structure of "that's neat"), we have a *normative* perspective — we're thinking about an idea of the way we think things *should* be. If we are observing without judgment (noticing that "it's lit fam" is used in reality), our perspective is *descriptive* — we're looking less at how things *should* be, and more passively seeing how things simply *are*. Evolutionary truth is our descriptive definition of truth. It is more about how ideas survive than it is about what an omniscient being would agree with. With that

in mind, it's tempting to consider which ideas survive even if they're not correct by any other definition of truth.

Which definition of truth might be a good benchmark to filter out truly good ideas from ones that are persistent yet unhelpful?
Consider this final definition:

Definition (Effective Truth). *A useful idea is true when someone using it tends to achieve their goals by doing so.*

For example, suppose you believe that red bowling balls are luckier than any other color. It just so happens that your red bowling ball fits your hand better than your other one, which is orange. In this case, your belief helps you to achieve higher bowling scores, so it's effectively true. I've chosen this example to pique your sense of imperfection, but next I'll argue that all of our ideas are like this; that is, I don't think the imperfection is in the definition of effective truth, but is a necessary property of truth itself.

Consider the simple and useful formula:

Claim (Newton's second law). $F = m \cdot a$.

Unlike your belief that red bowling balls are lucky, this idea probably feels unsuperstitious and reasonable. Here's the thing: It's not true. Newton's laws only apply to non-tiny objects that aren't moving very quickly (a physicist would say Newton's laws apply when the speed is slow enough to be called non-relativistic). Even under those conditions, it turns out Newton's laws are only approximations due to relativity.

If Newton's second law and the red-is-lucky idea are both in some sense false, then how do we work with the intuition that Newton's second law feels more correct? Here's one difference: If we were to scientifically explore both, we'd find that one of them is closer to correct more often than the other. Can we say that one is true and the other is false? I don't think so; both will fail to be perfectly true in many experiments. And both will appear more true than chance in the contexts above.

What we end up with is a way to evaluate different shades of truthiness. This perspective of truth relies on the context of goals to make sense. A "very true" useful idea, given a certain goal, will achieve that goal almost every time. If a goal can almost never be achieved, then we may still care about "slightly true" ideas — ones which achieve their goal only a small percentage of the time; consider a drug that saves 2% of patients with an otherwise fatal disease.

Similar thinking also works for evolutionary truth in the sense that if an idea is very persistent, then it better fits the definition of being evolutionarily true. In both cases — for both effective truth and for evolutionary truth — there are degrees to which an idea is true or false.

Here's the theme of this book: Truth is not as straightforward as we'd like it to be. It's worth spelling out clearly how we can measure *degrees of truth* in terms of the descriptive/normative viewpoints mentioned earlier.

- [Descriptive perspective.] **Communal ideas are as true as they are persistent**. That is, an idea believed by a community of people has more or less *evolutionary truth* to it according to how well that idea tends to persist, meaning both that people agree with it, and that those people continue to survive.
- [Normative perspective.] **Useful ideas are as true as they are effective.** That is, an idea being used to achieve a goal in a certain context is more or less *effectively true* according to how often someone can better achieve the goal by using the idea.

I'll summarize this:

Observation 2. *For both effective and evolutionary truth, ideas are neither completely true nor completely false, but have degrees of truthfulness.*

The concept of effective truth is not new, although as far as I know the particular definition I'm providing here has its own nuances. Some philosophers have previously explored similar approaches to truth, such as Charles Peirce and William James; their line of thinking is called *pragmatism*.

Is effective truth the same as our human intuition for truth? Something at first feels mismatched, and I'll explain this mismatch as an objection to effective truth. We want an idea to be true when it describes how the world is; it feels like achieving a goal is secondary to being correct about the world. Here's an example idea: A cloudless daytime sky is blue. It seems there's no goal to be achieved by knowing the sky is blue, so the idea is not effective (it's not goal-based) while still being true.

Here's a response to that objection: Although the idea the-sky-is-blue does not itself specify a goal, there are many goals I can achieve with that idea. For example, I can paint a landscape picture that includes the sky. I can look up and determine if it's likely to rain or not. If I can succeed in achieving my own goals by making use of an idea, then that idea has been effectively true. Later, in observation 4, I'll argue that we cannot even think of ideas unless they are somehow associated with some goal.

Another objection to effective truth can be found in thinking that red bowling balls are lucky. If I only have a red bowling ball and an orange one, and there's something wrong with the orange one, then my belief that red bowling balls are lucky becomes effective for me. Yet it feels wrong, because it doesn't feel true that the color of a bowling ball can directly increase my bowling scores. In general, perhaps we achieve a goal by doing a good thing for a bad reason. In that case, the idea could be effective, but doesn't match our intuition about what a true idea is.

Every time we put an idea to the test, we use it in a way that will either confirm the idea or refute it. Different tests may reveal different aspects of the idea. For example, if we try to apply Newtonian mechanics to near-light-speed space travel, we'll find physical experiments where an observation directly contradicts Newton's laws — that is, Newton's laws will be wrong. When we find a new context

where an idea is wrong, we have to either label the old idea as false, or to modify the idea to better understand where it applies.

In the case of red bowling balls, when applied to the single choice of the red versus the defective orange bowling ball, the idea of red being lucky simply works. The phrase "is lucky" is vague, but in this application, it means, "I usually get a higher score when I do this." If we tried to apply that same idea to other bowling balls then we would have to either discard the idea completely or add a restiction to it — just as Newtonian mechanics needs to be qualified. In other words, if you believe the laws of Newtonian mechanics have some truth, then you must likewise accept other qualified approximations for reality as true in the same sense.

I think that effective truth corresponds to our day-to-day intuition for truth. Here's a general argument to support this:

Argument: *That effective truth matches our intuitive idea of truth.*

- For every idea, we can only understand its truth by making decisions with the idea and seeing if the decisions result in things we expected. I call this *testing the idea.*

- [**Case 1**] Suppose an idea will pass every test we ever give it. Then the idea matches our intuition for being true, and simultaneously matches the definition of effective truth.

- [**Case 2**] Suppose there is some test which the idea would fail. Then the idea fails the definition of effective truth, and also fails to be seen as perfectly true by our intuition.

- *Note:* There is no room for discoverably bad reasoning ("red is lucky"): If a line of bad reasoning is discoverable, then there is a test that can reveal the incorrectness (eg, scientifically test all red bowling balls, and we find that red is not, in fact, lucky). If the idea relies upon undiscoverable bad reasoning ("because souls are purple"), then the undiscoverable part of the idea is superfluous and has no bearing on decisions or outcomes; thus we can ignore it and focus on the parts of the idea which allow testability.

To simplify, there are only two cases, and in both cases our intuition is found to match effective truth. To unsimplify, we must face the reality that (a) many useful ideas are known to be simplifications or only partially true; and (b) that, even if an idea has passed every test so far, we cannot know if it will continue to pass all future tests. We'll return to those two caveats below.

I'll summarize the above discussion as the second answer in terms of choosing among our definitions of truth:

Answer B. *Our practical intuition for truth matches* ***effective truth****.*

I've included the word *practical* here to separate this kind of intuition from a more *mathematical* intuition that we also have for truth; I'll briefly revisit this mathematical intuition in section 7.2 below.

With answer B in mind, most of this book focuses on effective truth.

Some philosophers, such as Karl Popper, might prefer the definition given for verifiable truth over that for effective truth. I began with verifiable truth because it is a simpler notion that appeals to the tradition that ideas are either completely true or completely false. And this dichotomy shows us the distinction between verifiable and effective truth; the definition of effective truth works well with ideas that are only partially true.

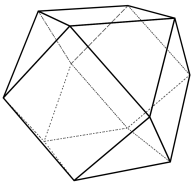
How does effective truth relate to the other definitions of truth? Evolutionary truth is different in that it allows for kinds of truth which are persistent yet not effective. The other kinds of truth are special cases of effective truth. These special cases of truth have been separated according to the kinds of goals they achieve. Authoritative truths achieve goals that standardize ideas using a central deciding authority; democratic truths achieve goals that align expectations among many people; moral truths achieve goals that improve society-level quality of life; and mathematical truths achieve goals in mathematical settings.

So far we've explored truth by considering either the goals we achieve, or the reasons we believe the ideas, depending on your perspective. But we can see truth in another light. People tend to believe in certain intuitive attributes of truth. In particular, I'll highlight two properties that I'll call P1 and P2:

Claim (Spoiler: I will disagree with this). *An idea being true is:*

- [**P1.**] **Independent of minds.** *That is, whether an idea is true has nothing to do with anybody asking the question.*
- [**P2.**] **Context-free.** *That is, whether an idea is true does not depend on how we think about or test the idea.*

Effective truth seems to be at odds with these properties: Effective ideas are tied to their believers and their goals in that the ideas we see as true depend on believers' actions and interpretations. The remainder of this book explores what I consider the most fruitful path toward understanding truth — the relationship between ideas and believers.



3

Inventions

A simple model of the world breaks all things down into *inventions* and *natural objects*. A natural object is something that existed independently in the world before it was found, such as Mars or the gravitational constant. On the other hand, an invention is something purposefully made, such as a car or a song.

For the sake of this book, I'm going to slightly expand the usual meaning of the word *invented* so that it applies to anything at all that has been purposefully created. For example, it's not clear if the traditional idea of invention would apply to birds' nests, but I'm shifting definitions so that nests count as invented since they're created by birds to serve the purpose of a home. I'm using this adjusted meaning because it handles edge cases that would otherwise be unclear, and because it's interesting to notice what's purposefully created and what's not.

3.1 Useful Fictions

With these categories defined, let's take a look at an interesting invention: an object's center of gravity. The philosopher Daniel Dennet calls the center of gravity a *useful fiction* because it's not a physical part of the world, although it is useful.

It lives in an interesting space between being real and imaginary. A well-defined solid object always has a center of gravity, and no human decides where it is. In this way, the center of gravity is somewhat independent of human minds. Yet it does not correspond to any particular molecule within that object. Just as the financial value of a dollar bill only exists as an agreement between minds, the center of gravity is an idea, and not a physically separate entity. In this way, a center of gravity is somewhat dependent on human minds to exist. I'm telling you about this dual nature because other concepts share this duality, and we'll use this lens to explore the dependence of truth on the minds of believers.

It's strange to call a part of physics, or to call a mathematical idea, a fiction. I'm using the word *fiction*, as Dennet did, in a nonstandard way. Instead of meaning that the idea is incorrect, I mean that it's created.

Useful fictions are inventions. The idea of a center of gravity is something created by people with the purpose of better understanding physics. There are many other useful fictions in our universe. To help identify them, notice that:

- A useful fiction x is a *fiction* in the sense that we can completely describe the physical state of the world without needing to describe object x.

- A useful fiction x is *useful* in the sense that we, humans, accomplish some goal with it.

The concept of ownership is a useful fiction, as are the notions of calculus and of beauty.

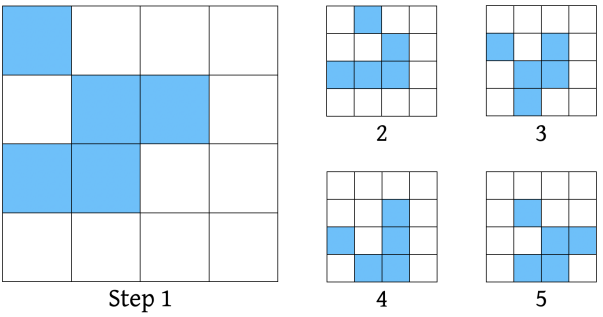
You might object by noting that, based on the points above, many natural-seeming objects could count as useful fictions. For example, limestone: In theory, we could describe the full state of the world in terms of basic particles without any need to actually define what limestone is. Is something as natural-seeming as limestone really dependent on minds to exist?

I claim that the concept of limestone — and other natural-seeming concepts — are indeed somewhat dependent on minds to exist. My formal argument is that anything meeting the two bullet points above for a useful fiction is something that is goal-motivated, and is based on a conscious decision within a mind to be named and generalized; and most human concepts have both of those properties.

I can try to establish some intuition that most human concepts are in fact dependent on minds to exist. To do so, I'll momentarily pretend that our rules of physics are based on the Game of Life; this will sidestep the fact that we do not yet know the rules of physics of our own world.

In the Game of Life, the word *glider* means a particular configuration of cells that moves diagonally across the board, assuming it does not encounter another object.

Here are all the shapes of a single glider across 5 time steps:



Notice that step 5 is in the same configuration as step 1. In this way, a glider retains its shape while moving diagonally.

The universe may contain nothing but gliders, yet we still have no need for the concept of a glider to describe everything in it. We could simply describe grid cells as being alive or dead. The glider concept is useful, created, and dependent upon minds for it to exist. Despite being analogous to simple molecules, they are still a useful fiction.

To drive home the point: If we could describe the entire universe in terms of particles smaller than atoms, then atoms themselves become useful fictions just as gliders, consisting of a few cells, are fictions within the Game of Life. Everything that exists at a larger scale than atoms, such as emus, brains, ideas in brains, or adjustable-rate mortgages — all of these things are conceptually built completely in terms of lower-level elements of the world, and are useful fictions.

3.2 The Ubiquity of Invention

It may bother your intuition about the sense of what is invented or natural to call an atom an invention, so let's explore this distinction. Consider the concept of water. Water seems quite natural; it would be strange to say it's an invention. Now consider a hypothetical *molecule X* which was painstakingly synthesized in a pharmaceutical research lab to treat a disease. Molecule X does not occur naturally on Earth, and in fact is quite difficult to create. It seems more natural to say that molecule X was invented. Why do we freely call molecule X an invention, while we want to say water is natural? It seems the difference to us is in the pre-existence of water, and in the purpose of molecule X.

Now consider a remote planet where molecule X is abundant, but water is nonexistent. Inhabitants of this planet could reach the opposite conclusion if they created water themselves. So it seems that our intuition for what is natural and what is invented depends on our personal experience, and is not objectively based on the state of the universe alone.

Our judgment of some concepts being natural is itself something we find useful, and something we have added to the world; it is a useful fiction.

Here's why I'm discussing all of this:

Observation 3.
Truth is a useful fiction. It is an invention, and not a natural part of the world.

This observation disagrees with the common intuition that truth is independent of minds (the idea that truth is independent of minds is property P1 from Chapter 2's final claim). How can I justify observation 3?

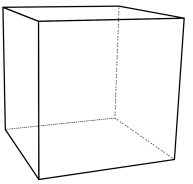
One justification is the fact that truth meets the two bullet points above describing useful fictions: Truth is not something we need to describe the world, and it's useful.

I'm going to provide more than one argument to support observation 3 because it can feel quite unintuitive. I think it's easier to see that the concept of falseness is invented; I'll use that in the first such argument:

Argument: *That truth is an invention.*

- The idea that "x is true" only makes sense as a denial of the opposite idea that "x is false." If there were no concept of a false idea, then the concept of a true idea would be meaningless.
- The idea of falseness is invented. We can describe the world without having to include the idea of falseness. By describing the state of every physical particle in the universe, we describe all brains and, implicitly, all concepts within those brains — and we've done so without needing to explicitly spell out an idea of falseness. There is a purpose to the idea of falseness, which is to understand mistakes or deception.
- Combining the above two points: Since the idea of truth is essentially paired to the idea of falseness, it must be the case that both ideas are invented.

A natural counter-argument is that perhaps falseness is invented while truth is not. I'll expand on this point in the next section, arguing that pairs of concepts which are alternatives to each other can only exist together, rather than individually.



4

What Ideas Are

In order to better understand the relationships between alternative notions (up vs down, red vs orange, true vs false), it's helpful to begin with the question:

Question. *What kind of thing can be true or false? In other words, what is an idea?*

As a reminder, I've been using the words *idea* and *claim* to talk about things that might be true. The word *idea* is adjusted in this book to exclude things like "the idea of the color green" because that isn't something that could be true or false.

Some philosophers go into a tizzy when you ask what kind of thing can be true or false. You might say a grammatically correct *sentence* can be true or false, but then you must deal with any random yet grammatically correct string of words such as "a yellow sadness confused the used passenger;" or logical ambushes like "this sentence is false." There's a serpentine rabbit hole here that I don't want to get stuck in.

Instead of trying to find a perfect answer, I'm going to speak at a high level and suggest two connections that help us better understand the word *idea* as it's used in this book. I'm not going to define exactly what an "idea" is. Rather, I'll keep the word *idea* as a placeholder for our intuitive understanding of what can be true or false, and I'll lay out some key properties of this intuitive notion.

4.1 Questions and Answers

The first connection I want to make is between ideas and answers to questions.

If an idea tells us some information, then it must distinguish between different possibilities. If I say "I'm having a nice day," then I'm distinguishing this from the possibilities of having an astounding day or a dog-tired day. If there weren't multiple possibilities, then there is no information conveyed. If I said "red is red" without more context, it's confusing because it's not clear what alternative I'm excluding.

There's a nice mathematical analogy here. It's almost as if there were a short conversation like this:

- [Mathematician A] I know the value of x is in the set of possibilities X, but exactly which possible value is x?
  (The math notation here would be written "$x \in X$; what is x?")

- [Mathematician B] Oh, it turns out that $x = 4$.

There is a question (what is x?) with a set (X) of possible answers. The information content — the idea being expressed — is an answer to this question ($x = 4$). If you say "My name is Inigo," then you're answering the implicit question "What is your name?" whose set of possibilities is the set of all names.

This mathematical analogy is not perfect. We often speak without thinking about the alternatives we're excluding. The question we're answering is typically implied, and may not be clear, even to the speaker. I claim that if you *did* think about any statement you make, then you'll typically find specific alternatives you've excluded. Another imperfection with the analogy is that some sets of possibilities are not clearly defined. If I say "this wine tastes of elderberries and the phlegm of an incontinent camel," then we must admit there seems to be a fuzziness to the set of things a wine can apparently taste like; you'd be hard-pressed to enumerate this set exactly.

Nonetheless, the analogy helps us to build our intuition. An idea has meaning only because it expresses one way things *are* from amongst a (non-mathematical) set of possible ways things *could be*. Similarly, a question is an expression of a set of possible ways things could be, along with a desire to learn which item in that set is correct. I'll summarize this connection as the answer to our most recent question about what kinds of things can be true or false:

Answer. *An idea — a thing that can be true or false — is an answer to a question.*

This is similar to a definition, but really I'm just breaking down one intuitive notion (ideas) into two other not-carefully-defined notions (answers and questions). I still find this connection useful, as we'll see below.

4.2 Goals

The second connection I'd like to make is inspired by asking:

Question. *Why do we ask questions?*

Perhaps every thought we think, including every question we ask, is motivated by a goal. But there's an objection: It seems that some thoughts we have are not goal-motivated. Consider raw perceptions that enter your mind simply because you experienced something — for example, you notice a car driving by, or perhaps you hear a friend ask a question. You may understand your friend's question without being directly motivated to ask it yourself.

With this objection in mind, I'll separate our thoughts into *conscious* ones that we deliberately choose to have, and *external* thoughts or perceptions which we do not choose. I'll make the claim that every one of your *conscious* thoughts is motivated by some goal you have.

Are there any remaining objections to this claim? If you try hard, you might come up with interesting thoughts that seem to be conscious yet may not have clear goals. As examples, consider a child at play, or someone distracted by a squirrel running across their path. In the case of play, I think that having fun is itself a goal. Play is an action some animals engage in. It may be helpful in learning about social interactions and other life skills. In terms of being distracted by a squirrel, I would say the initial moment of distraction is an external, unbidden thought — you see a squirrel. However, if you have any follow-up thoughts about the squirrel, those are conscious thoughts. They may be motivated by curiosity, which I see as a basic drive alongside our need for food or sleep. Sating your curiosity is a goal.

For any conscious thought I've considered, I've found a goal behind it.

Based on this, we can have conscious thoughts that are ideas, and others that are questions. The perspective of this book is that an idea is an answer to a question, so in either case there is a question present. And a question is a set of possibilities along with a desire to know which of these possibilities will help us move toward our goal.

Let's summarize this as:

Answer. *Every question we consciously consider is motivated by a goal.*

This includes both questions we ask directly as well as implicit questions that give meaning to a conscious idea. Such questions always exist, based on the previous section.

4.3 Impossible Ideas

What I've argued is that every idea is an answer to a question, and that every conscious question is motivated by a goal. Putting these two together, every conscious idea is connected to a goal.

This is a strong claim. One key implication is that we cannot consciously think of ideas that aren't motivated by a goal.

Note that it's possible for an idea to be associated with different goals. Perhaps the goal which first led us to an idea is different from a goal we later associate with it. For example, we may first think of a water heater as a way to make tea, and later learn how to make instant ramen with that same water heater. So my claim is not "every idea has one goal which motivates it," but rather "every conscious idea must have at least one goal to motivate it."

It's intriguing that there are ideas we cannot consciously think of — those for which we have no goals. They seem hard to believe in, so it would be interesting if we could find an example to bolster our intuition. Consider learning to speak Mandarin when you only speak English. This is difficult because Mandarin uses *tones* — patterns in the pitch of your voice — as a key part of a word's pronunciation. Two words in Mandarain can have the same consonant-and-vowel sequence, yet can be pronounced differently based entirely on their tones. English-only speakers typically have no concept of a word's tone. They've learned to understand consonant-and-vowel sounds in English, and to ignore other points of data from word sounds. If we imagine a world without tone-based language, then we can imagine that people would never think of tones at all. Tones would be an impossible concept — a concept we were incapable of consciously conceiving.

This isn't a perfect example of an idea we can't think of — because we *can* think of it; it's just an unknown concept for some people. It's not my fault I gave a partial example. It's literally impossible for me to express an idea that cannot be thought of. But such ideas must exist because we'll never have experienced all possible goals. For every goal we never have, there is the question of how to achieve that goal, and an answer to this question. We will never think of any of these ideas.

If it helps your intuition, consider an alien race that sees colors we cannot; their eyes see different wavelengths than human eyes. They will have many words for "colors" that we would not recognize as colors. These colors are another kind of concept we almost cannot think of. Again, technically, this example is partial (hence the word *almost*) because scientists can understand what I mean — and, again, any example I actually give must necessarily be partial. But if you can stretch your imagination a bit in thinking of these aliens as having entirely new dimensions of perception, and we never meet these aliens or encounter this dimension of perception, then you can begin to envision ideas that we're currently unable to experience.

4.4 Concepts

We have the rough workings of a model of human thought. I'll add one more ingredient to this model. Earlier, I said that this book adjusts the meaning of the word *idea* to exclude, say, the color green since a color can't be true or false. In

this section I'll formally introduce the word *concept* to talk about things like colors. Concepts are the things ideas are made of. Examples of concepts include colors and water heaters and letters of the alphabet. Intuitively, if ideas are mental sentences, then concepts are mental words. As with the term *idea*, I won't try to define concepts, but I will analyze our intuition around them.

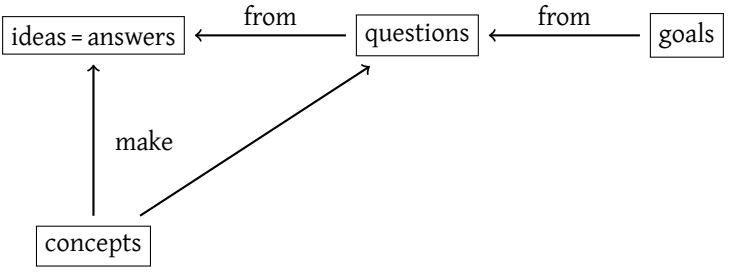
To ensure the notion of a concept is clear, let's consider another example. Suppose one day you're stung by a distinctive-looking wasp. A year later, you see a similar wasp, and you steer clear of it, remembering the sting. It's not that you have a word you can say out loud to name this wasp, but rather that you have a connection in your brain that can recognize it and understand that it can sting you. Contrast this level of awareness to a bug you may peripherally perceive, but never care about; if you saw such an innocuous bug later, you could have no associations at all with it.

It's possible that concepts originate in a brain process preceding conscious thought. I'm not going to suggest a specific origin story for concepts because we don't need to understand where concepts come from in order to begin analyzing them. I've separated conscious thoughts from external ones. Focusing on conscious thoughts (excluding, say, raw perceptions), I've suggested they are always goal-driven and deliberate. I suspect that the first time we deliberately use a concept, it will be a part of a conscious idea. If this is the case, then as soon as we consciously consider a concept, it's attached to a goal, and this attachment is unavoidable.

I'll summarize this idea, along with the previous section's ideas, as:

Observation 4. *Every conscious thought we have is goal-based. Goal-free conscious thoughts are impossible.*

Here's a diagram summarizing the mental model of ideas used in this book:



In this diagram, I've included the notion that concepts are parts of both questions and ideas. This makes sense because a question is the same thing as an idea with an unknown element in it: *What is your name?* is the same as *Your name is which* $x \in X$*?* where X is the set of possible names.

A concept can only add information to an idea if it distinguishes the idea from alternatives. There must be possible variations of the idea that are excluded. This

set of variations exists even if it's not clearly defined. If I asked why the chicken crossed the Mobius strip[1] the set of possible answers allows for creativity, and is not something we could easily enumerate.

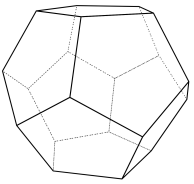
Some pairs of concepts, such as *light* and *dark*, are essentially always alternatives to each other. *Heat* has meaning only in opposition to *cold.*. The answer to some questions is expected to be *true* or *false*, such as when you ask "Is idea x correct?" In this way, *true* cannot exist alone; it is part of one concept together with *false*. This interdependence bolsters the argument in §3.2 that truth is an invention.

. . . . .

Some questions seem less useful than others because we can easily guess their answers. For example, imagine asking a random person if they're less than 110 years old. We have concepts for certain age groups such as kids, toddlers, and teens. But we don't have a name for people under 110; that mental notion isn't as useful. Concepts are useful when they center on areas of uncertainty, away from predictability. In other words, there appears to be a correlation between what we don't know (adjacent to useful questions) and the concepts related to those questions. This relationship between truth and ignorance is worth further exploration.

[1] Answer: To get to the same side.

Note that the corresponding idea, *The chicken crossed the Mobius strip to get to the same side,* seems to pertain to a new kind of truth not discussed in this book — a kind of "truth" based on comedic value, albeit rather hypothetical value in this case.



5

Ignorance and Omniscience

Ideas depend on questions, and questions assume ignorance, even if imagined. This dependence of ideas upon ignorance may seem counterintuitive, so let's try to understand how necessary ignorance is for ideas to make sense.

5.1 Ignorance

I'll use three example questions to help build up an intuition about the role of ignorance in ideas and truth. To state these questions, I'll define *pleven* numbers:

- An *even* number is any integer that is a multiple of 2;
- an *odd* number is one more than an even number; and
- a *pleven* number is any integer which is neither even nor odd.

Spoiler alert: There are no pleven numbers. But you might be able to imagine they exist, just as you might imagine there can be a triangle with three 75° angles before you learn that all triangles must have angles which sum to 180°. This strange thought experiment is purposefully awkward, and we'll explore the awkwardness behind it.

Here are the three questions:

Q1. How long will my commute be this morning?

Q2. What's the first positive pleven number?

Q3. What would the color blue be if it weren't blue?

Question **Q1** seems practical enough. At first glance, there do not appear to be any strange assumptions behind this question. Question **Q2**, on the other hand, feels weird because we're asking about something that doesn't exist. How can we answer a question of detail about a nonexistent thing? Question **Q3** seems to go one step further in this direction, being nonsensical unless we decide to think poetically or playfully, giving up on taking ourselves literally.

For a moment, imagine that you have complete and immediate awareness of all future commute times. Suddenly question **Q1** seems silly, similar to "How many minutes are in 30 minutes?" because the answer is obvious to you. Depending on how profoundly familiar you feel with commute times, question **Q1** may even feel to you like "Is blue blue?" which is so obvious as to be confusing. If you force yourself to take seriously the question "Is blue blue?" then you're in the zone of question **Q3**.

Returning to a human who doesn't know all commute times, question **Q1** is reasonable because we can imagine worlds with different commute times. This kind of imagination is akin to ignorance — to not knowing the answer. The other questions are strange because we're deeply familiar with the state of the world that gives the answers. We can't easily imagine pleven numbers existing, or blue not being blue. I included question **Q2** to show how a question's intelligibility depends on the mind considering the question. A kid just learning about even and odd numbers will take this question seriously — it makes sense to them. But it stops making sense once you realize there are no pleven numbers.

I'll summarize this line of thinking as:

Argument: *That ideas only make sense in the presence of ignorance.*

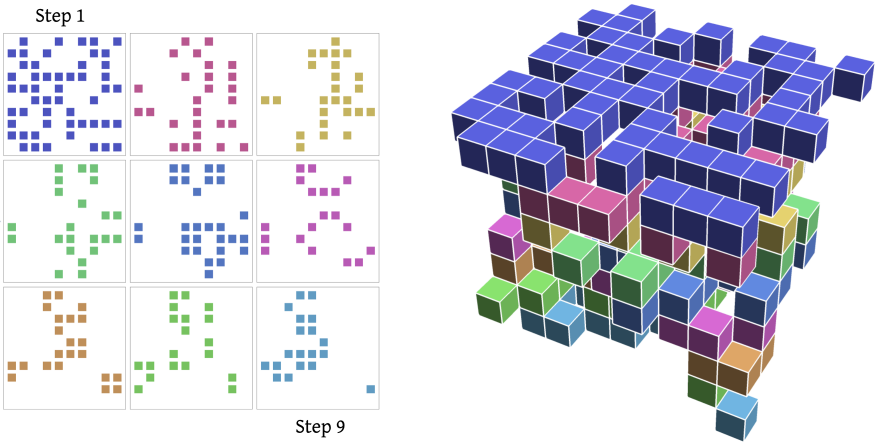
- Every idea is an answer to a question.
- A question only makes sense if we can imagine different answers are possible.
- When we're deeply familiar with information that provides the answer, the question no longer makes sense.
- Therefore, questions — and their corresponding ideas — only make sense when they're about things we can imagine not knowing.

This is another refutation of true ideas being independent of minds (from the last claim in §2.3). Ideas require ignorance — ignorance that in practice exists within, and depends upon, a mind — to be meaningful.

5.2 Omniscience

I'll describe a thought experiment that further explores how ideas relate to the limits of our awareness. The general idea is to imagine that you know everything. Pretend you're so familiar with the state of the world that questions tend to feel like "Is blue blue?" The answers seem self-evident. This is different from knowing the axioms of math and being asked if a certain theorem is true — in that case you would have enough information to derive the answer, but you'd have to think about it first. I'm aiming for knowledge without the need for additional thought to answer questions.

To solidify the thought experiment, pretend the world is a single instance of Conway's Game of Life. Any single time step is represented by a 2D grid of cells that are on or off. Visualize a 3D version of this where each slice is a single time step. In this way, all of timespace is laid out before you as a single static object. Time is no longer a mystery, but rather a dimension just like space.



In the illustration above, the same 9 steps of the Game of Life are depicted on the left and on the right. In the 3D figure on the right, step 1 is the top layer, step 2 is the layer below that, and so on. (This version of the Game of Life is implemented with wrap-around at the edges of the 10×10 grid, so that, for example, cells along the left column are considered neighbors to the corresponding cells along the right column.)

In this thought experiment you "know everything" in the sense that you know the full physical state of the world. Perhaps you don't know all possible mathematical truths, or all other possible worlds. So this is a certain kind of omniscience. Let's call it *physical omniscience*.

Now someone asks you a *physical question*, by which I mean a question about the physical state of the world. They point out a glider at a certain time step and ask how long this glider will live. You, being physically omniscient, respond: "What's a glider?" There's no reason for you to know or care about what a glider is because you can see all of history without knowing that concept. Once you learn what a glider is, then you can easily answer the question. But such a question feels arbitrary to you, in that a questioner has defined some random-seeming concept and asked a question about this concept in a situation where the answer is inevitable. It may feel as if someone defined even and odd numbers, and started to go through all the primes checking to see if each one were odd and even. It's a silly task when you know that 2 is the only even prime. But it doesn't feel as useless before you know that.

Let's summarize this idea as:

Physical questions seem nonsensically obvious to the physically omniscient.

Now let's imagine that you are both physically omniscient, and that you identify with one particular agent in the world. When you see this agent in the world you think "Hey, that's me!" What kind of goals might you have for this agent? I don't think you can have *any* goals, because you see before you the full life of your agent just as you see the indelible ink on a page. To imagine a different path is to reread The Lord of the Rings and hope that Gollum becomes an insurance claims adjuster by the end. If you're creative, you may choose to ignore what you already know, but your rational mind understands the futility:

The physically omniscient are not motivated by physical goals. (5.1)

I say *physical goals* to tease out other goals that we can imagine, such as proving a mathematical theorem. We may also imagine that you, the omniscient party, exist in some outer world, outside the confines of the world you have omniscience about (think The Matrix). You might have agency in this outside world, but that agency doesn't matter here since the point of this thought experiment is to look at how knowledge of *our* world affects ideas *about our world*, not about other worlds. That is, *physical goals* in (5.1) are restricted to actions and results observable within your inner world, the one that you know all about.

When you know the full physical state of the world, it's as if you've made a giant list of all possible worlds $\{x_1, x_2, \ldots\}$ and then learned which possible world is the real one. When you're omniscient, you know there's a certain number k that indicates the real world x_k. If the world used the Game of Life as the rules of physics, then perhaps the number k would indicate the starting conditions of the world; everything else follows. No matter what the rules of physics are, we can suppose there is some way to encode the state of the universe across all time, and the number k can be this encoding.

Building on the idea of encoding the world in a single number, I'm going to pose a strange-seeming mathematical question whose purpose is (a) to show that some math questions make us wonder what the point of the question is, and (b) to point us toward understanding the role of curiosity and abstraction in our thoughts.

Given any positive integer, I can write that number in base 2, in base 3, etc.[1] I'll create a table by writing a number's base 2 (binary) form in the top row, and then below that its base 3 form, then its base 4 form, and so on, keeping them all right-aligned. Here's a table for the number 534:

										base
1	0	0	0	0	1	0	1	1	0	*2*
				2	0	1	2	1	0	*3*
					2	0	1	1	2	*4*
						4	1	1	4	*5*
						2	2	5	0	*6*
						1	3	6	2	*7*
						1	0	2	6	*8*
							6	5	3	*9*
							5	3	4	*10*

I can make such a table for any number. I can further imagine that each number in the table is the value of a cell, analgous to the cells in the Game of Life being on or off. I can think of the top row as the first time step, and as each subsequent row as the next time step. (Notice how the number 534 effectively encodes and captures this entire world; this is "world 534" in some sense.) Now I can ask the question: *How long does the 1 in the second-from-right column survive?* That is, if I repeat the top row like this:

1 0 0 0 0 1 0 1 **1** 0

then how long does the framed cell remain as a 1 in the rows below? The answer, for 534, is that this value survives for the first 4 time steps, corresponding to bases 2–5.

As promised, I hope that you have two reactions to this question, corresponding to earlier goals (a) and (b) of this thought experiment:

1. I hope you find this question to be pointless. How is it useful? Why would we ever ask this question?

2. I hope you wonder: Out of curiosity, are there any fun mathematical patterns to questions like this?

[1] We normally write a number, like 7, in base 10 because we have 10 fingers, and thus 10 digits (0-9). If humans had only 5 fingers in total, then we'd have digits 0-4 and we'd count this way: 0, 1, 2, 3, 4, 10 (this means 5 to us), 11 (meaning 6), 12 (meaning 7), etc. Base b notation is how we'd write numbers if we had b fingers instead of 10. If this notion is entirely new to you, it's too much to learn in a footnote, and I recommend googling "introduction to number bases" to learn more.

Now consider the similarities between the following questions:

- [In reality] How long will I live?
- [In Conway's Game of Life] How long will this glider live?
- [In a number base table, as above] How long will this value survive?

These questions all have the same shape: They each specify an object in the world, and ask how long it will be around. We probably have different emotional reactions to the questions, but they feel like the same kind of question to the physically omniscient.

I've purposefully chosen a well-motivated question (How long will I live?) to illustrate that even a strong sense of motivation can evaporate in the presence of physical omniscience. The number base question (How long will the 1 survive?) is pointless because you have no control over it or interest in it. Similarly, your sense of personal agency and your physical goals both dissipate when your experience of the world is based on universal physical knowledge. Questions about survival collapse into the realm of passive observations, akin to remembering a past event that has no bearing on the future.

One purpose of the thought experiment is to help you to intuitively see the following, even when you personally identify with an element of the world:

Physical concepts seem pointless to the physically omniscient. (5.2)

The idea of *curiosity* presents a challenge to this simple perspective because virtually any question that can be asked can be motivated, in theory, by curiosity. If you're physically omniscient and curious, you might find all questions to be motivated.[2]

One thing we can do to fit curiosity into this model is to consider two kinds of worlds: the physical world, and the world of abstract thought. A historic fact is an idea about the physical world, while a mathematical theorem is an idea about the abstract world. Many ideas blend these two worlds together, such as asking if a particular theory of physics is true. In that case, we imagine a possible world based on a physical model. This model lives in the abstract world. We ask if this model is a good approximation for reality — we can do a physical experiment to check.

Whenever curiosity is the motivation for a physically omniscient person, there must be an abstract concept involved. If there weren't, there'd be nothing unknown to makes sense of the question; we'd be back to questions that feel like "is blue blue?" Once we're looking at abstract notions, we're outside the realm

[2] I suspect curiosity has an evolutionary origin that connects it with survival. For example, perhaps a curious pre-human primate would discover the utility of controlled fire, whereas a less curious species would not. This book won't try to analyze curiosity, but I think it's something we could analyze and put boundaries around. We humans find some things curiouser than others, and there are probably ways to understand this.

of physical knowledge. Hence claim (5.2) continues to make sense in the presence of curiosity with the understanding that a *phsyical concept* is one motivated by a physical goal, with no abstract component to it. As an example, the generalized concept of all gliders, and the concept of glider survival, are both abstract ideas; such abstractions allow for curiosity.

If we wanted to, we could change the thought experiment to imagine a person who knows all physical states of the world *and* all abstract ideas as well. I think that such a person would then be incapable of curiosity, as no question could be motivated. Any question would seem nonsensically obvious to such a being. Sounds like a disappointing existence.

I'll summarize the ideas of this section and the previous one as:

Observation 5.
Thoughts can only be motivated when we have ignorance about them.

5.3 Dependence on Context

Behind every question is a goal, which — for effective truth — gives us a way to test ideas. We can think of a goal, or a test for a goal, as having some inputs and some outputs. If you're adding numbers together, the inputs are the numbers being added, and the output is their sum. If you're trying to make money by trading stocks, the input is everything that influences the stock market, and the output is how much money you make or lose. To be clear, I'm considering tests to be deterministic once we know enough about the inputs. If something seems probabilistic, such as whether it will rain tomorrow morning, then it simply means we have not accounted for all the variables. It may be difficult or even impossible for a human to know all the variables, so the claim is not about how a human brain works, but rather about how the world works. Specify the inputs, and the outputs are certain.[3]

When a question has a simple relationship between inputs and outputs, then the corresponding truth — such as how to add numbers — feels easy. Complex input/output relationships make the corresponding truths feel difficult. When a problem is easy, it's less interesting.

In some special cases, such as in pure mathematics, we have awareness of all the variables that might affect an outcome. In most cases, though, we don't. When we don't know all the variables relevant to an idea, then each time we test the idea, there are elements of the test we don't control for. This makes the efficacy — the truthiness — of the idea dependent on the variables we don't control for. In other words, the efficacy of an idea can depend on the context we use it in.

[3] I'm speaking as if we know definitively that the laws of physics are completely deterministic. What if they aren't? Then we can still model the world as deterministic in the following way: For every time step forward, we have two inputs. The first input is the previous state of the world, and the other input is an incoming stream of outside data. This data can be equivalently viewed as either a source of nondeterminism, or simply as unknown information. My point is that we can speak of ideas having all relevant inputs even if the laws of physics do not allow us to know those inputs in advance.

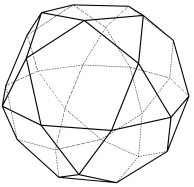
I'll summarize this as:

Observation 6. *Most ideas don't account for all inputs that could affect whether the idea works. Whether such ideas achieve their goals is context-dependent.*

Consider the idea that water boils at 100°C, which we know is not accurate unless we take into account additional factors such as air pressure. A counterargument to observation 6 goes like this: It's true that sometimes an idea like "water boils at 100°C" misses some important context — but we can always fix the idea by adding in the missing context, such as accounting for air pressure. Technically, I think this point is true, but it doesn't refute the observation because we still use simple ideas even when we understand there is unspecified context. For example, we retain the idea that water boils at 100°C without additional context because it's effectively true for practical purposes.

Philosophically, either you're forced to say that most of what you consider true is declared as fully false (since not all context is included), or you must accept that these ideas are only provisionally true. The perspective of this book is not to redefine truth, but to analyze the intuitions we have and to be honest about the role truth plays in our lives. Given the above choice, most people implicitly accept "water boils at 100°C" as provisionally true rather than as false.

The final claim of §2.3 outlined our intuitions that true ideas are independent of minds, and do not depend on context. Observations 3–6 largely refute those ideas. True ideas require minds to make sense: Without the ignorance of minds, questions do not make sense. Without the goals of minds, ideas are unmotivated. And without full contextual awareness, our tests of ideas are noisy — whether the test succeeds depends on things we have not accounted for.



6

The Fuzzy Edges of Thought

If there is a theme to this book, it's that truth isn't as objective and absolute as we usually consider it to be. Along those lines, let's look at another intuition we tend to have about truth. Consider the question:

Given two specific animals, are they members of the same species?

At first glance, this appears to be a perfectly reasonable scientific question that ought to have a clear yes or no answer. If I point to two people, the answer is yes. If I point to a dog and a cat, the answer is no.

The concept of a *species* carries with it a sense of clarity and a sense of being intrinsic to the world. This section will consider these proposed two properties of concepts:

1. [The Clarity Property] If a person fully understands the concept x, then they can always know if thing y is an x; the distinction is clear.

   Example: Is an apple a fruit? It seems that questions like this must have a definite answer, assuming the category of thing (*fruit*) is fully understood by the person answering.

2. [The Intrinsic Property] When we learn a concept, we're learning about the way the world is — we're not deciding anything about the world, but observing it.

   Example: It seems silly if I point to a water molecule and ask if we got the definition of water wrong, and maybe we should redefine *water* to exclude

this one molecule. Liquid H_2O is water, and this notion is part of nature, not something I decided.

6.1 Most Concepts Are Fuzzy

Let's take a look at the question of two animals belonging to the same species.

Traditionally, a species is defined as a group of living things which can reproduce together. But this idea doesn't always work the way we want it to. A simple challenging case occurs when a group of similar animals reproduces asexually, with a single parent resulting in two copies of itself. In this case, we cannot use the above definition, and must begin looking at features of the animal itself. But these features can be surprisingly tricky to work with. For example, chihuahuas and Great Danes are members of the same species who look mismatched, while jaguars and leopards, with similar features, are two different species.

In fact, the closer we look at the notion of a *species*, the more problems we reveal:

- [Microspecies] Scientists have found that what you informally think of as blackberry plants are actually a collection of about 400 different species that are quite similar to each other. Species within such a group are called *microspecies*. You might intuitively suppose that any two such plants are of the same species, but scientists disagree.

- [Hybridization] A mule is the offspring of a horse and a donkey; horses and donkeys are considered to be different species. Mules are thus a *hybrid*, a cross between species. Mules are typically unable to reproduce, which means that any two mules would not be considered members of the same species based on the traditional definition!

- [Ring Species] Different communities of a bird called the *greenish warbler* live in subgroups around the Tibetan Plateau. Many of these communities that live near each other can interbreed, but the most disconnected groups cannot. We have communities of similar animals — call the groups A, B, and C — where A and B can interbreed, and B and C can as well, but A and C cannot. This setup is called a *ring species*, and reveals another challenge to the reproduction-based definition of a species.

- [Evolution] Finally, we have to take into account the idea of evolution itself. Given two same-species parents and their offspring, we see the offspring as belonging to the same species. But if we extended this line of species labels throughout a complete family tree, we'd end up naming many living things as the same species, despite vast differences.

  For example, the human precursor species *homo habilis* is thought to have evolved directly into *homo erectus*. If we take the idea of a species seriously, then we're forced to believe in a family with two homo habilis parents and

the first-ever homo erectus as their offspring. Yet there would be no enormous change within this particular family. It's simply a line we draw to distinguish the two species.

The core concept of a species is not really about the ability to make offspring together. That traditional definition is useful for the many cases in which it makes sense. But it's trying to capture an intuitive notion for a species that is based more on practicality than on a single guiding scientific principle.

The notion of a species is fuzzy — it's not a mathematically elegant concept, but rather a pragmatic one with unclear edge cases.

And the fuzziness of a species is not alone. While fruits can be defined as edible plant structures associated with seeds, vegetables have no such botanically-based definition. It's true that vegetables are typically edible parts of plants, but not every edible part of a plant is a vegetable; apples aren't vegetables. Vegetables come from different parts of plants: Tomatoes (often considered vegetables) are seed-bearing fruits, carrots are roots, a leaf of spinach is ... a leaf. Some people consider mushrooms to be vegetables, in which case we can no longer say that vegetables are a subgroup of edible plants because mushrooms are fungi, not plants. Just like a species, there is no clear and elegant definition for what a vegetable really is.

What distinguishes notions like *species* or *vegetables* from more clearly-defined things like *positive integers*? I think the key is that our intuition for a species is based on looking at examples. When we learn from examples, there's no particular reason our "definition" will end up being completely clear in all cases. Indeed, I'll argue that all of our intuitively-understood, example-based concepts are fuzzy the same way that species and vegetables are.

Argument: *That example-based concepts are fuzzy.*

- Suppose that concept x is learned by example and is intuitively understood, as opposed to being understood via a clear-cut definition. For example, people learn to identify *cats* by seeing pictures of cats, or by seeing actual cats in person.

- Concepts only achieve meaning by having alternatives that they distinguish. Suppose concept y is an alternative to x. For example, x=*cat* and y=*dog*.

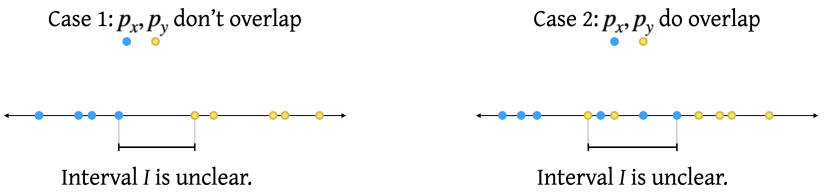
- From the defining examples for concepts x and y, we intuitively learn properties of objects that indicate whether one is an x or a y. For example, if a four-legged animal is over 2 feet tall at the shoulder, it's probably too big to be a domestic cat, but it might still be a dog.

- It's logically possible to encounter an object with properties of both x and y to such an extent that we feel confused.

  To help justify that these confusing examples can exist: Consider a single property p of an object, and suppose we can represent p as a real number. For

example, p may be an animal's height. We learned to distinguish x from y by examples, so we can let p_x be the set of p-values seen on examples of concept x (such as cat heights). Similarly, p_y is the set of p-values from examples of y (such as dog heights).

Perhaps all the p_x values are less than all the p_y values. Then there must be some non-empty interval $I = \big(\max(p_x), \min(p_y)\big)$ in between p_x and p_y. In this case, when you see a new object with a p-value in I, you can't use property p to distinguish an x from a y. If, on the other hand, the p_x and p_y values overlap, then there is still a range of p values that do not clearly indicate an object is an x or a y — specifically, the range where the example values overlap.



In short, no matter what the sets p_x and p_y look like, simply because they are finite sets of real numbers, there must be some non-empty zone of ambiguity. Objects with a value of p in the ambiguous zone cannot be distinguished based on the learning examples.[1] This argument shows that an ambiguous zone must exist for all real-valued properties.

- Because we may one day see an object that we cannot identify as an x or a y, we have found a point of confusion in our current understanding of concept x. This is what it means for a concept to be fuzzy.

- Note that it's unnecessary for these confusing objects to be seen in reality: If a concept were truly clear, then any object we could ever encounter would have a definite label associated with the concept (e.g., is-a-hot-dog, is-not-a-hot-dog). The mere possibility of encountering an indefinite (unclear) object shows that the concept has undecided edge cases.

The conclusion:

Observation 7. *Many concepts are intuitively defined by examples. Every such concept is* fuzzy *in that it's unclear if the concept applies in certain cases.*

[1] This argument is inspired by ideas from machine learning. I'm appealing to the intuition that even when we decide to find a simple decision boundary, such as a straight line in a plane, we typically do not have a *uniquely-determined* hypothesis. For example, if $p_x = \{1.3, 2.4, 3.7\}$ and $p_y = \{4.1, 5.8\}$, and I see $p = 3.9$, then I have no definitive way to tell if the object is an x or a y.

This observation refutes the clarity property of many concepts.

I'd go so far as to argue that *most* of our concepts are learned and understood from examples, rather than from a clear internalized definition. Try this thought experiment: Think of a word that you understand well, like "for" or "red;" now try to define this word without the use of any examples. If you work hard, you can probably come up with some useful definition — but notice that this process is quite different from how you use the word in your day-to-day life. You don't pause and check an instance of the object against a definition you have in mind. Rather, it seems that we have an intuition for correct usage based on experience — based on examples. Even for concepts that allow for a clear definition, if the definition is based on fuzzy concepts, that fuzziness leaks through. Consider the idea of a *veggie plate*, which we can define as a *plate of vegetables*; this definition appears clear and straightforward, but we still have to ask if we should count mushrooms as vegetables — and this fuzziness applies to veggie plates just as much as it applies to vegetables themselves.

Since most of our concepts are either learned by example, or defined in terms of earlier fuzzy concepts, most of our notions are fuzzy like the concept of a *species* is.

6.2 Social Ideas

The previous section focused on the clarity property — the claim that a fully understood concept always has an unambiguous relationship to objects. Now let's consider the *intrinsic* property — the claim that true ideas are about the world rather than based on any decisions made by a person.

We'll start with an example of an idea that is easy to accept as true: Cars in the United States drive on the right side of the road. This idea is verifiable — we can look and see which side of the street cars drive on. It's also effective — we'll live longer if we use this idea than if we ignore it. The U.S. government has written laws describing this idea, making it authoritative; and it's a practical way to drive because everyone knows to use this idea, making it democratic.

There's something interesting about this idea: It's not a fact about history; yet it is an accident of history. Americans could just as easily have decided to drive on the left side of the road. In fact, some countries such as England and Australia do drive on the left side, and they seem to be doing fine. What's important about this idea is that people *agree* to the same convention to achieve the goal of safe and organized driving.

We could say that which side of the road a country drives on is *arbitrary* in that it's not the decision itself which is good or bad, but rather it's the existence of agreement that makes it useful. When I call an idea *arbitrary* in this book, I mean that the goal of the idea (eg, road safety) is more important than the implementation of the idea (eg, which side of the road we drive on). I'm adjusting what I mean

by the word *arbitrary*, as I did earlier with the word *invented*.

Are there other conventions that could have been implemented differently? Certainly. In English we read from left to right, but we could have read from right to left. We draw the letter A a certain way, and we could have just as easily drawn it differently. Perhaps all written laws are in some sense arbitrary since they would have no value without being part of a society of people who make collective use of those laws. Even math has conventions that could have been chosen differently. We think of $\pi = 3.14159\ldots$ as a fundamental constant, and we write 2π for double this value. But there's a compelling argument that it would be more elegant to forget the value of π as a constant, and to instead define a different constant, τ, as the ratio of a circle's circumference to its radius (what we call 2π). In this alternate version of math, we would write $\tau/2$ to indicate the traditional number $3.14159\ldots$; this version of math is just as complete and correct as our own. In this sense, we can see an arbitrariness even to things that may not feel like choices, such as π.

Since these ideas only make sense when used within a group of people, I'll call them *social ideas*, and summarize the pattern we're seeing as the following observation.

Observation 8. *Most ideas are related to a social goal. Social ideas are* arbitrary*; it's more important that a group agrees on the idea than that a best idea is found to achieve the social goal.*

I'll justify this observation with a simple chain of key definitions:

Argument: *That social ideas are arbitrary.*

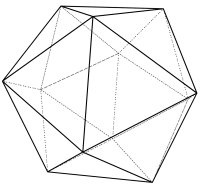
- A social *goal* is a goal which requires agreement among different minds to be achieved. For example, a group can use the idea of *currency* to achieve the goal of decentralized distribution of goods.
- A social *idea* is an idea motivated by a social goal.
- Combining the above: *How* a social idea is implemented is less important than *that* it is used, meaning that it is agreed upon within the social group. Thus social ideas always have an arbitrary element. For example, what we call a currency (dollars), and how we denote it (dollar bills), are implementation details that matter less than the use of the currency.

The above argument focuses on social ideas, but there appear to be non-social ideas that can be arbitrary as well. It seems that *how* questions tend to have an arbitrary element to them, while *what* questions may not. For example, there's only one reasonable answer when we're finding the value of $1 + 1$, or the atomic weight of a hydrogen atom. But if we ask *how* to make an optimally-delicious cup of coffee, we have choices, and often it's not clear that there exists a single best way to accomplish the goal at hand.

Observation 8 refutes the intrinsic property for social ideas — it suggests that most of our ideas involve significant elements of human choice, rather than being objective observations of the world.

Combining the ideas of fuzziness and arbitrariness, we can see how unclear edge cases in a concept (are these two things the same species?) and a desire for agreement can cause people to make arbitrary decisions (in this unclear case, we'll choose a new special case for what a species is). These decisions permeate our day-to-day lives: How to spell words, when we have holidays, whether hot dogs count as sandwiches. They also pervade the world of science. Most mathematicians consider it to be true that $a^2 + b^2 = c^2$ among sides a, b, c of a right triangle with hypotenuse c. However, even here there's a *choice*: we are assuming the world of plane geometry. Because mathematical plane geometry exists in minds, and not in the physical world, we *choose* the axioms we use.

I see choices even in the most objective, non-abstract ideas I can think of. For example, suppose we humans one day completely understand all the laws of physics on the level we understand the Game of Life. Suppose we can represent all of the physical world by thinking of a grid of cells as being either on or off at each time step. Perhaps reality will be different from this, but this suffices as a thought experiment. In this case, we still have a choice in what we consider an atom of the universe. We can think of a single cell as an atom with two possible values (*on* or *off*), or we can think of a 2×2 block of cells as an atom with 16 possible values (corresponding to the 16 possible on/off values of the 4 cells). The two perspectives are mathematically equivalent. If we apply some principle like Occam's razor to say that one perspective is superior to the other, then we are still making a choice. Even in this theoretically complete and correct view of physics, we are making decisions.



7

The Foundation of Human Knowledge

7.1 Knowledge

Let's look at the role truth plays in our model of the world, and in ourselves.

Next are some candidate properties for truth that feel intuitively correct. These include properties P1 and P2 from §2.3; property P3 is new:

P1. True ideas exist independently of people understanding them.

P2. A true idea is true in any setting.

P3. Every question has a correct answer.

This book argues against all of these intuitions:

- [Refuting P1] Ideas depend upon minds: They only make sense in the presence of both goals (observation 4) and ignorance (observation 5); many ideas are based on accidents of human history (observation 8).
- [Refuting P2] The truth of many ideas depends on how we test it (obs. 6).
- [Refuting P3] An answer to a question can only have a partial degree of truth to it (observation 2), and many questions are about cases where there is no clear answer (observation 7).

Although I disagree with P1-P3, it's still interesting to see how these notions fit into our internal model of the world.

When infants first learn a language, I don't think they have a strong notion of truth. Instead, it seems that kids start by learning useful concepts such as hunger and discomfort, who their parents are, and how they can connect with their parents. As they grow, people learn the basics of many subjects including, say, arithmetic or basic physics. If you continue to study math or physics, you eventually get to the philosophical notions of an axiomatic and logical basis for math and physics.

Even before you consider an axiomatic system of knowledge, most people have a sense that every meaningful question we can ask has a correct answer (P3 above); in the presence of axiomatic systems, this intuition is only strengthened. Although we're not born thinking this way, we gravitate toward the idea that knowledge is an infinite sudoku board, where the things we know are the numbers we've uncovered, and the rest of the board contains hidden numbers that we'll uncover with time and thought. We could informally capture this idea by saying that "truth is the foundation of human knowledge;" or, to spell it out more explicitly as a claim:

Claim ("The Truth-as-Bedrock Claim"). (Note: I'll disagree with this.)
There are a set of ideas which are true, and have always been true.
All of human knowledge consists of our discovery of the set of pre-existing true ideas.

What's new here is the focus on the human process of gaining knowledge. Earlier, when we looked at definitions of truth, we considered reasons we might believe an idea, but we didn't pay much attention to the larger picture of how new questions are answered for the first time.

In refuting the above claim, it may sound like I'm saying that truth does not exist — but that's not my goal. Rather, I'm going to argue that when we learn things, our self-education is closer to invention than to discovery. My suggestion is that our knowledge process is akin to collecting ideas we find worthwhile, and later trying to make a bundle of those ideas coherent as a group.

Before presenting a full argument against the truth-as-bedrock claim, I'll present two thought experiments that I'll call scenarios A and B. My goal is to build your intuition that humans form ideas not directly around truth, but rather around experience, and experience itself allows for misinterpretations.

. . . . .

[**Scenario A.**] Consider a simulated universe with a population of creatures. At each time step, there will be food in one of two locations. The creatures must guess where the food will be in advance. If they're correct, they survive and create offspring for the next time step. Otherwise, they starve and die.

The creatures don't know it, but the food location is always random. However, enough offspring are created so that the creatures never die out. The creatures, trying to predict the next food location, form ideas about patterns they think they see. The creatures who have survived many time steps will genuinely believe in

the correctness of the patterns they perceive. The patterns chosen will appear historically correct, and those creatures who observe a false prediction will die; the result is that survivors will see their ideas reinforced.

In this scenario, the creatures are clearly creating ideas not out of an underlying truth, but simply by matching patterns on limited data.

This scenario shows that it's possible to reasonably form ideas even if there are no underlying ideas to discover. But reality may be different — not as random as this scenario — so it's useful to consider a situation that feels closer to our world.

. . . . .

[**Scenario B.**] For this scenario, imagine that an early human group observes their first eclipse, and the next day they experience their first earthquake. They form the idea that the two events were connected. This is an example where humans make an idea within a world that could be deterministic, as opposed to the randomness of scenario A.

A natural response is to say that an idea has low confidence when it's based on a single observation. So let's suppose this same group begins to test their idea. They observe four more eclipses, and in each case there is an earthquake the next day — and they experience no other earthquakes. To their minds, their idea is strongly supported.

Given enough time, observation, and advances in understanding, they would eventually disprove their incorrect idea. But in the meantime, their belief is rationally justified. This shows that people can form false yet reasonable connections even if they live in a world that's predictable when you have enough information.

What's the philosophical difference between this group of early humans and ourselves? We base all our ideas, necessarily, on a finite number of data points. Perhaps we can have more confidence than these early humans because we collect more data points. But this difference in confidence is not the same as a difference in the fundamental process which occurs: We make mental connections and we never fully know whether those connections are, in some ideal sense, true.

With those thought experiments at our disposal, let's proceed with the argument against the truth-as-bedrock claim:

Argument: *That ideas follow learning, versus learning follows ideas.*

- If our process of building knowledge was based on uncovering a set of underlying true ideas, then (a) we would only keep ideas which worked every time we tested them; and (b) even for ideas which we later found to be incorrect, there would be an underlying, corrected version of that idea. By refuting (a) and (b), I will be refuting the claim that "Human knowledge consists of our discovery of the set of pre-existing true ideas."

- I'll start with (a). We do keep ideas that we know are not completely correct, such as Newtonian mechanics, or rules of thumb. You could counterargue

that these are useful approximations, but this is not a valid counterargument because I'm not disagreeing. Rather, I'm simply saying that we still remember and make use of ideas that we know are only approximately true. Strictly speaking, they are known to be false, yet they are part of human knowledge.

- Next I'll argue against (b), that ideas found to be wrong must have underlying true versions. To do so, I'll refer back both to scenario A (the hungry creatures who see patterns where there's randomness) and to scenario B (a rational belief that eclipses cause earthquakes). Those scenarios build our intuition that ideas are *made*, and that they may be false even if they appear reasonable. When we create such ideas, they can easily fail to have an adjusted, corrected version. They're so far from the mark, we can't even view them as approximations. This notion is strengthened when we also consider that many thoughts are fuzzy or arbitrary, further revealing a disconnect between what we believe is true versus any hoped-for fundamental truth behind those beliefs.

- [Objection 1.] One objection to this argument is that a corrected version of an unfounded idea, such as eclipses causing earthquakes, is to simply say the idea is wrong. I agree that some ideas are wrong, but saying so isn't the same as revealing a truth beneath those ideas.

- [Objection 2.] Another objection to this argument is that there seem to be some ideas that are objectively true, such as mathematical theorems. If those are true, and we prove a theorem, then it seems that we have discovered an objective truth.

  I agree that a logically proven idea appears more objective than other ideas. Yet this is not a hole in the argument because I'm not arguing about the objectivity of what we believe; I'm arguing about our process of forming beliefs. So I don't have to show that ideas are subjective. Rather, it suffices to show that our default thought process isn't based entirely on the discovery of objective truths. Some fraction of our ideas may have the special property that they can't be discredited (such as theorems). These ideas co-exist with a default process of knowledge discovery in which truth itself is a derived secondary notion, rather than a foundation.

We've seen that our thoughts are often fuzzy, arbitrary, and created practically rather than discovered organically. I'll summarize this section as:

Observation 9.
Human knowledge is built upon imperfect, fabricated concepts.
Many of our questions, being about imperfect, fabricated concepts,
don't have objectively correct answers.

This observation refutes the truth-as-bedrock claim. It is not accurate to say that human knowledge is purely founded upon a pursuit of pre-existing external truths.

7.2 Supertruth

Throughout this book, I've been distinguishing between two kinds of intuition for truth. I've argued in favor of *effective truth* as matching our intuition for what we use in practice. At the same time, I've argued that we often (incorrectly) assume that true ideas are independent of minds, independent of context, and beneath every idea is an underlying truth (properties P1-P3 above). Despite my arguments against those properties, they are common enough to warrant a closer look.

I've argued that the other kinds of truth listed in Chapter 2 are special cases of effective truth, excluding evolutionary truth. Let's focus on thoughts based around *mathematical truth*, in which ideas can only be entirely true or entirely false, and never somewhere in between. I'm going to consider this alternate version of truth — one that does obey properties P1-P3 — and I'll call this notion *supertruth*.

Some ideas appear to genuinely live in the world of supertruth. It's hard to imagine how 1+1=2 could be false to any degree without changing the rigorous and widely understood meaning of the terms. We could, for example, re-interpret + as an xor operation on strings of bits, but this is more sleight of hand than an objection to what is really meant in saying 1+1=2. Similarly, 1+1=3 is false any way you slice it; we could call it *superfalse*.

Can an idea about the world — not about abstract mathematics — be supertrue? I'll call an idea *physical* if it can be tested in the world. My question becomes: Can a physical idea be supertrue? I can imagine that we live in a world with permanent laws of physics, and that the state of the world can be completely described in some axiomatic model. If we understood these laws, then I claim that we'd be able to find a physical idea that's supertrue.

Let's see how that might work. I'll frame a supertrue physical idea as one that says something of the format "If X is true of the world, then Y is also true." A logically justified physical idea must meet these requirements to be supertrue:

- The idea exists in a world with permanent laws of physics.
- The idea uses axioms about physics that coincide with those laws of physics.
- The idea specifies all the relevant context of the antecedent, X.
- The idea only uses physical definitions that are free of unknown edge cases and are clear down to the finest possible resolution of physical structure (the ultimate analog of atoms, for example).

If our universe was the Game of Life, we might say, for example, that "A glider which never touches another living cell will live forever." I can't give a real-world example

of such a precise physical idea because I don't know our own laws of physics (not exactly), but it seems likely to me that such statements are possible for our own world.

My conclusion is that supertrue ideas about the world — as opposed to being purely mathematical — are possible, but are neither within reach, nor likely to be how we view the world anytime soon.

Suppose we do come across a supertrue idea. In practice, every test of this idea will work as expected. However, as argued in Chapter 2, we can't know with certainty that the idea is supertrue. If it's a math idea, we may make a mistake in calculation, or an error in evaluating the proof. If it's a physical idea, then we can't have definitive proof that our set of axioms are correct, even if they have never been contradicted. What I'm doing is distinguishing between an idea actually possessing supertruth, and our human ability to have complete knowledge of this supertruth. The fact that we can never be certain of supertruth strengthens the argument that our practical notion of truth is aligned with effective truth. Observation 2 stated that "For both effective and evolutionary truth, ideas are neither completely true nor completely false, but have degrees of truthfulness." Another perspective of this idea is that our *confidence* in an idea must always live somewhere in between certainty it's supertrue and certainty it's superfalse.

. . . . .

It seems that we're looking at two contradictory ideas of truth: effective truth and supertruth. And our intuition seems to support both of these notions as being "what truth is." How can we address this apparent paradox? Luckily, there's an easy resolution — realizing that our intuitions are not necessarily consistent. That is, our own intuitions are the source of the contradiction.

In case you have doubts that human intuition can be so self-contradictory, let's consider two illuminating examples. First, suppose you believe in the standard axioms of arithmetic, and then someone presents you with this statement about integers x, y, z, n:

$$\text{If } n > 2, \text{ then there are no positive values of } x, y, z \text{ with } x^n + y^n = z^n. \quad (7.1)$$

This statement is known as Fermat's Last Theorem. Suppose you're friends with a well-established mathematician who you've never known to be wrong, and she reads a proof that (7.1) is false. She tells you the proof is correct. Now you're in the state of intuitively believing in the axioms of arithmetic, and also believing (7.1) to be false. But this is a contradiction, because (7.1) is true, and that truth can be proven based on those axioms. This is an example of simultaneously believing things that logically contradict each other.

You might object by saying this is a constructed example, one that most people don't encounter in their daily lives. I've included it to show we can believe in direct logical contradictions. Other examples are less focused on pure logic. Imagine a person who's lived in New York City their entire life. Suppose they think of doves

as white birds, symbols of peace and cleanliness. With that understanding, they may honestly believe they've never seen a dove in person. However, if you see a pigeon in New York, it's almost certain that you're looking at a *Columba livia* — also known as a *rock dove*. Pigeons and doves are more or less the same animals. Like the word *vegetable*, the non-scientific terms *dove* and *pigeon* are fuzzy — but there are species, including *Columba livia*, which are considered to be both doves and pigeons at the same time. This is another example where intuition — that you see pigeons every day, that you've never seen a dove — may directly contradict itself.

I'll summarize the recent line of thinking:

Observation 10.
Our typical thoughts use intuitions that match effective truth. When we think in a more strictly logical manner, these intuitions instead match the idea of supertruth.

Most of the things we do in life are not about careful logical reasoning. If you examine all the decisions you make throughout the day, you'll find that most of them — quite possibly all of them — are practical decisions rather than precise logical conclusions. While supertruth is an idea we know about, it's a relatively obscure version of truth that we tend to use in academically-oriented settings.

Columba livia, from John Gould's *The Birds of Europe*, Volume 3.

7.3 What do we do with these ideas?

I've outlined ten observations about truth with the theme that truth is a noisy, imperfect, made-up notion that only makes sense in the context of ignorance and goals. So what, exactly, am I suggesting you learn from all this?

. . . . .

Is truth real?

It's tempting to conclude from this book that truth is not real. After all, I've argued that ideas are typically imperfect and depend on human minds to make sense. How can something be real if its existence depends on a mind?

Consider chairs. I think of chairs as real. And yet the difference between a certain configuration of atoms and *being a chair* exists solely in my mind. For example, suppose an insane alien crafted an object identical to a chair, but it made no sense as a thing for these aliens to sit on. This chairlike thing is seen by the aliens as a piece of abstract art. I'd say it's not a chair. It's not the shape of a chair that makes it a chair, but the interpretation of the beholder. My point is that everyday notions we consider to be real can perfectly well depend on minds.

While this book does posit the dependence between true ideas and minds, it doesn't refute the reality of truth itself.

. . . . .

Should we never believe an idea is fully true?

Another conclusion of this book is that we can never be completely certain that an idea is correct. This is troublesome in that most people do have full confidence in certain basic ideas; it would be inconvenient to additionally remember a confidence level associated with every idea you have. For example, consider the idea that gravity pulls things downward. There is more nuance to our common sense about gravity — we know that helium-filled balloons tend to go up and that clouds float. With nuances in mind, it feels silly to think that maybe I'll hold up an apple, let it go, and it will do something besides fall.

For all practical purposes, we can consider many ideas to simply be *true*, without qualifications. Although I can make arithmetic mistakes, I'm confident that $1 + 1 = 2$. With the caveats mentioned above, I consider the Earth-based downward pull of gravity to be a fact of life. If we were to build a probabilistic model for certainty, there are many ideas which we'd think of as being 99.9999% true. Once we get enough 9s in there, there's no value in using up mindshare with a stipulation that this idea worked the first n times, but may just as likely not work the next time.

Although we never have indisputable confidence that an idea is true, it's practical to treat high certainty as if it were complete certainty.

. . . . .

How does this book change anything?

If the reality of truth has not changed, and how we think about certainty has not changed, then what has? After all, if the observations of this book have no effect on our lives, then it would seem that they aren't effective. In a sense, this book would be declaring itself to be false if it results in no change.

One large-scale effect of this book's observations is that we change how we think about the question:

How can one learn what is true?

This book provides a basic working model of our thought process around truth. With that, we can explicitly spell out many underlying assumptions we make as we explore what might be true or false.

Let's look at an example. Consider an argument between two programmers about how to best code a particular function. In practice, there is often an underlying sense that one of the two people must be correct. With the perspective of this book, we can shift our view to see this as two people selling products that do the same thing. In one sense, one product may be superior to the other, such as if we arrive at a standadized measure of quality. But in another critical sense, we understand that there is no singularly correct answer. Rather, the decision is effectively a negotiation. When we recognize that fact, it can facilitate our feelings and communications between each other. It can give us the humility to accept that

another approach may always be better than ours (we cannot have full confidence in our own ideas), and the freedom to act in the face of uncertainty (they cannot have full confidence in theirs). This book throws a glass of cool water in the face of dichotomous thought.

In an academic sense, this book can also help us iterate on the scientific method. Because common scientific practices are somewhat based on tradition, they don't always adapt well to new ideas. One example would be the consideration of emotions in animals. It is, to this day, often considered scientifically unsound to attribute emotions to animals based on their behaviors. The argument is that animals cannot explicitly state their emotions, so we are necessarily guessing at how they feel. While this argument makes some sense, if we accept that all of science is making guesses and then seeing how they hold up to evidence, then we see that such guesses are quite reasonable.

There are many other scientific practices worth questioning. We pretend that observational studies hold zero evidence of causation, but this seems heavy-handed since causation is falsifiable through observational studies. We pretend that scientists are not emotionally attached to proving their hypotheses, and thus are reliably unbiased in their judgment, but this is unrealistic. And we pretend that p-values are a great way to check for statistical significance, when it's quite possible for this filter of significance to misfire.

At the end of the day, these two ideas — our *individual* thinking about how to decide what we believe (such as deciding between two ways to code a function), and our *academic* thinking about how to decide what we believe — are two sides of the same coin. They are each reflections of how one can learn what is true. This search for truth becomes easier when we're aware of our own central role in creating ideas, and when we keep in mind that non-dichotomous nuance is the norm rather than the exception.

. . . . .

There is another consequence of this book's perspective on truth: Often we'll think of a question as a deep mystery of the universe, when in fact we're asking something about ourselves. Indeed, if the very notion of truth is one we've created, that exists in our minds alone, then our sense of a mystical unknown is likely to encapsulate something much more mundane — and therefore more accessible — than we first expect. Consider: emotions; consciousness; intelligence; creativity; morality; a sense of life's purpose. All of these notions can feel at once profound and impossible to completely comprehend. Yet perhaps each is both less central-to-the-world, and less daunting than it appears.

To look at a single example, consider our feeling that we're alone in the universe. Many people find it mind-boggling that there may be other intelligent beings in the universe. Some folks associate the belief in possible alien life with insanity. Similarly disparaged is the idea that a machine may one day be a non-human kind of person, with its own equally valid emotions and consciousness. But if we

strip away the certainty in the notion "we are alone in the world," then the difference between human-unique intelligence and many-peoples intelligence is akin to the difference between a one-frog pond and a two-frog pond. The difference is neither so mystic nor so implausible as we may first suppose.

A part of our human psychology seems to push back against uncertainty. It feels bad to let go of the ideal of supertruth and to accept effective truth. There is a part of us that doesn't want our place in the universe to shrink by admitting our lives are based on uncertainty, and that we invent what we believe. We want to be right, and we want to be special. Yet there's beauty in *just being*, like anything else in the world. With the smallest shift in perspective — an acceptance of humility — we become less alone and more connected, having more in common with everything around us.

The world is a much, much bigger place when you see how small you really are.

. . . . .

What is truth?

Answer. *We live goal-driven lives in a world we know little about. Our goals give us questions; we guess answers to them. Some guesses effectively achieve our goals. We call these effective ideas the* truth.

Afterword

There are at least two goals an author may have when they write about philosophy. The most obvious is that they'd like to share some ideas about philosophy. I've tried to do this. Another motivation for an author is to demonstrate their own expertise about philosophy. This is not my goal, but I'm worried that it might be perceived otherwise. I want to dispel the possible implication that I believe I've written a book chock-full of original ideas. I don't.

What I've written is a combination of viewpoints from thinkers such as Charles Peirce, William James, Daniel Dennet, and Richard Dawkins. I've mentioned the folks I know of with highly relevant contributions. There are probably other thinkers who've had similar ideas before I wrote this. If I didn't mention them, I wasn't consciously aware of the connection. I don't claim originality, but I did put some thought into what's written here, so I'll call this book a combination of pre-existing ideas along with *independent* ones. If it turns out these are all considered old hat, there's still value in trying to clearly organize and present a framework of fascinating perspectives.

. . . . .

I've tried to make this book easy to read. I'm a firm believer that good philosophy can be understood by anyone who cares to understand it. Some intimidating works by, say, Immanuel Kant or Friedrich Nietzsche might suggest otherwise — and I consider this to be a detriment to their abilities as communicators, regardless of the value of the ideas expressed.

I've also tried to practice what I propose by expressing the ideas of this book not as fully true, but rather as effective stepping stones towards a more complete understanding of ourselves and the universe. I've chosen my terms carefully. I shy away from the word *proof* in favor of *argument* because the word *argument* does not imply the conclusion is ironclad. I've named my mental milestones *observations* because these are things that I see, as opposed to things I assume you agree with. An observation has an observer, a connection which serves both to be honest about my personal perspective, and to be transparent that human error is possible.

If you re-read the book, I encourage you to keep an eye out for my attempt at philosophical humility, such as in my choices of words (arguments, observations), or in my framing of ideas as perspectives, rather than inevitable conclusions.

Everything I say could be wrong — though it could also be effective.

Glossary

This book rejiggers the meaning of a few common English words or phrases. Here they are:

Arbitrary idea	*an idea for which the goal being achieved is more important than the idea itself, such as a set of glyphs (characters) as a way to write down a language*
Concept	*a piece of an idea, a mental word*
Fuzzy concept	*a concept that allows for unclear cases, such as when you see an animal halfway between a tiger and a lion — it's not clear if it's a tiger or not*
Idea	*a kind of thought that can be true or false*
Invention	*something made with a purpose*
Physical question/idea/goal	*a thought about the physical world versus about purely abstract ideas; eating a meal is a physical goal, proving a theorem is not*
Social goal	*a goal that requires the agreement of minds to be achieved, such as the use of a common currency*
Social idea	*an idea motivated by a social goal*
Supertruth	*an idea which is not fuzzy, and is true under all circumstances*
Thought	*an idea, concept or question*
Useful fiction	*another term for an invention*

Observations

Observation 1.
For every idea presented as a truth, there is a reason given to believe the idea is true. When we examine these reasons, we find uncertainty.

Observation 2.
For both effective and evolutionary truth, ideas are neither completely true nor completely false, but have degrees of truthfulness.

Observation 3.
Truth is a useful fiction. It is an invention, and not a natural part of the world.

Observation 4.
Every conscious thought we have is goal-based.
Goal-free conscious thoughts are impossible.

Observation 5.
Thoughts can only be motivated when we have ignorance about them.

Observation 6.
Most ideas don't account for all inputs that could affect whether the idea works. Whether such ideas achieve their goals is context-dependent.

Observation 7.
Many concepts are intuitively defined by examples. Every such concept is fuzzy *in that it's unclear if the concept applies in certain cases.*

Observation 8.
Most ideas are related to a social goal. Social ideas are arbitrary*; it's more important that a group agrees on the idea than that a best idea is found to achieve the social goal.*

Observation 9.
Human knowledge is built upon imperfect, fabricated concepts.
Many of our questions, being about imperfect, fabricated concepts, don't have objectively correct answers.

Observation 10.
Our typical thoughts use intuitions that match effective truth. When we think in a more strictly logical manner, these intuitions instead match the idea of supertruth.

Thanks

Thanks to my friends Richard Price, Mark Wells, Joël Franusic, and David Indermill for thought-provoking conversations that helped to inspire this little book. And thanks to Sean Dennis for editing that made it a better book.