A tool for helping science find the optimal path toward the truth: falsification trees

Introduction

How do we know whether a given scientific theory has been proven false? This essay breaks down the logic of falsification and explains why theory falsification is always inconclusive: our evidence can never *definitively* falsify our theories. This is illustrated first with a trivial example, then a historical example (the Copernican revolution), and then two contemporary examples (the foundations of quantum mechanics and the neuroscience of consciousness).

It is argued that the inconclusiveness of theory falsification creates multiple pathways for science: ways science could have evolved differently. This raises the question: how do we know we are taking the most objective pathway, the one that will reveal nature's secrets in the most efficient way?

This paper argues for a novel proposal to help the scientific community take the optimal route. The proposal combines the familiar idea of an *adversarial collaboration* with a new idea that I call a *falsification tree*. In an adversarial collaboration, advocates of competing theories try to falsify each other's theories. The idea of a falsification tree emerges from an analysis of the logic of falsification. Essentially, falsification trees help researchers search many scientific pathways at once to help locate the optimal path.

The logic of falsification

How do we know whether a given scientific theory has been proven false? The answer may seem simple: when it makes predictions that contradict what we see. But this answer is too simplistic. For it suggests that when one concludes that some observable evidence has falsified some theory T, one has reasoned as follows:

(1) If T is true, then E is false.
(2) E is true.
(3) Therefore, T is false.

Here E is a sentence describing one's observable evidence. To simplify, consider a trivial example, in which seeing a blue leaf falsifies the theory that all leaves are green. Let T be "All leaves are green" and let E be "some leaves are blue". We then have:

(1) If "all leaves are green" is true, then "some leaves are blue" is false.
(2) "Some leaves are blue" is true.
(3) Therefore, "all leaves are green" is false.

But I may have observed a green leaf that had merely been spray painted blue! This counterexample may seem too trivial, but it is enough to show that the above argument does not adequately capture the logic of falsification. In fact, it illustrates an important lesson about the nature of the scientific method emphasized by influential 20th century philosophers of science, especially Popper, Quine, and Duhem.

A theory (like "all leaves are green") does not entail that you will *experience* all leaves as being green. Rather, it only entails that *all else being equal*, you will experience all leaves as being green. So, to have genuinely falsified theory T, all else had to have been equal, for example, no sneaky spray painting! The "all else being equal" clause captures all those assumptions you need to make, if you are to conclusively conclude that some leaves are blue given your "blue leaf" experience.

But *how many* such assumptions does one need to make here? The set of assumptions seems open-ended. After all, I could have been a hermit, with no knowledge of anything resembling spray paint. To deduce that theory T is false given some observable evidence, it seems I need to make assumptions that I cannot even formulate.

Following the likes of Popper, Quine, and Duhem, we can make all this explicit as follows. When one reasons that some evidence E has falsified some theory T, one reasons more like this (where A is the conjunction of the assumptions):

(1) If T and A is true, then E is false.
(2) E is true.
(3) Therefore, T and A is false.
(4) Therefore, *either* T is false *or* A is false.

Here we use E to better capture your evidence in experiential terms. To illustrate with the trivial example:

(1) If "all leaves are green" and A is true, then "I experienced some blue leaves" is false.
(2) "I experienced some blue leaves" is true.
(3) Therefore, "all leaves are green" and A is false.
(4) Therefore, *either* "all leaves are green" is false *or* A is false.

Here A will include "no spray painted leaves!" Premise (1) therefore implies that if all leaves are green and no one is spray painting them blue (etc.) then I will not experience blue leaves. Note that A may include an infinite number of assumptions, many of which may be impossible to anticipate. Thus, the conclusion in (4) is logically equivalent to:

Either T is false or A_1 is false or A_2 is false or … or A_N is false.

Here A has been broken down into the potentially infinite set $\{A_1, A_2, \ldots A_N\}$. The key point is that theory *falsification* really involves theory *selection*. In the face of apparently

falsifying evidence, I must select whether I want to abandon my theory or maintain my theory and abandon one or more auxiliary assumptions.

To ensure that we make the right selection, it might seem like we just need to experimentally test the assumptions. But there are two problems with this. First, there may be many such assumptions that we cannot yet formulate, so this can never be conclusive. Secondly, experimentally testing an assumption just raises the same issue all over again: to test the assumption we need to infer predictions about what we should experience, but that inference will rest on many further assumptions. Theory selection is therefore inevitably inconclusive. For the most part, the scientific community makes these selections collectively, thereby selecting a trajectory in the space of possible trajectories science could have taken.

This raises the question, how do we know we are taking an objective pathway, one that could ultimately lead us to the most fundamental truths about reality? And if we are taking one among many possible objective pathways, how do we know we are taking the most efficient path? For more clarity, let's consider a historical case study.

Historical case study: the Copernican revolution

The logic of falsification outlined above can help us understand why some theories in the history of science were so tenacious. For example, the Aristotelian "geocentric" theory, which placed Earth at the center of the universe, dominated Western thought for nearly two millennia, from Aristotle (384 –322 BC) to the time of Copernicus (1473 - 1543) who provided a viable alternative, and Galileo (1564 –1642) who championed that alternative.

According to Aristotelian's geocentrism, the sun, the moon, and all the other planets orbit Earth, and their orbits are circular. However, an apparent counterexample was quickly discovered: retrograde motion, where a planet (like Mars) appears to reverse direction, and then reverse direction again. We then have:

(1) If Aristotle's geocentrism is true, then planets do not exhibit retrograde motion.
(2) Some planets exhibit retrograde motion.
(3) Therefore, Aristotle's geocentrism is false.

At this early stage in the history of science, there were two possible routes that science could have taken. Route one: take the above argument seriously and reject Aristotelian geocentrism and put all efforts into developing a viable alternative. Route two: maintain Aristotelean geocentrism and put all efforts into identifying and rejecting an auxiliary assumption.

Unfortunately, science took the second route. Ptolemy of Alexandria (circa AD 150) argued that premise (1) is only true if you assume that there are no epicycles. In other

words, Aristotelean geocentrism is consistent with our experience of retrograde motion, provided that planets perform epicycles as they traverse the circular orbits proposed by Aristotle. Astonishingly, this worldview was the dominant view for nearly two millennia. The research program involved constructing ever more complicated epicycles to fit the data. Models of the solar system became extremely complicated, as illustrated by the Astronomy article in the first edition of Encyclopedia Britannica (1771). Here, science took the wrong path, one that steered us away from the truth.

The Copernican “heliocentric” theory placed the Sun at the center of our solar system and did not require the universe to have a center at all. This paradigm shift was a hard-fought battle. Scientists who were critical of Copernicus argued as follows:

(1) If heliocentrism is true, we should fly off the Earth and see stellar parallax.
(2) We neither fly off Earth nor see stellar parallax.
(3) So, heliocentrism is false.

For many, this argument was a compelling reason to consider heliocentrism falsified. The first concern was that if Earth was hurtling around the Sun at great speed, then it would be like sitting on a roller coaster without being strapped in. The second concern was that given heliocentrism, there should be a shift of position of nearby stars against the background of distant stars (stellar parallax), which was not being seen by the telescopes of the time.

However, this argument takes a familiar form. The predictions in (1) don’t strictly follow from heliocentrism without auxiliary assumptions. The first prediction assumes there is no attractive force like gravity, holding us to Earth. The second prediction assumes telescopes then were powerful enough to see stellar parallax. Both assumptions eventually turned out to be false.

It is often said that key defenders of heliocentrism, like Kepler (1571 - 1630), were motivated not by conclusive scientific evidence, which was not available at the time, but by a mystical belief in the mathematical simplicity of the universe. One may worry that theory selection, even when correct, as it was here, may frequently be based on subjective, nonrational considerations. Considerations such as where one’s intuitions lie, or what theory sounds most interesting, beautiful, or simple, or what theory is in one’s personal interests to further, and so on.

This was the deep concern of another great 20th century philosopher of science, Thomas Kuhn (1922 - 1996). Kuhn proposed a controversial “social constructivist” view of science. The ultimate choice in theory selection, between falsifying one’s theory or instead falsifying auxiliary assumptions, inevitably depends on non-objective social considerations, like intersubjective values. As a result, Kuhn questioned the very idea of objective scientific *progress*.

The problem with a social constructivist view of science, is that it doesn't clearly distinguish science from pseudoscience and conspiracy theory. Indeed, one could imagine a modern-day geocentrist astronomer, who is given modern-day evidence, yet insists on geocentrism, holding that NASA photos are fakes. This is reminiscent of the contemporary flat-Earther, who always seeks to identify and reject some assumption we are making when we assert that flat-Earth is falsified by our evidence.

Kuhn bit the bullet here. As he once said, "myths can be produced by the same sorts of methods and held for the same sorts of reasons that now lead to scientific knowledge." The worry is that the flat-Earther and the modern scientist, while they are to some extent constrained by their observations, nonetheless choose to believe what best fits their subjective values, since their values have the last say on whether to abandon one's theory T or to instead abandon some auxiliary assumption A.

In the case of the flat-Earther, what one values the most is their theory (and being right about it). This is clearly not a route to truth. How, then, can we do better?

A way forward: adversarial collaborations and falsification trees

Our historical case study concerned a science of very tangible entities. Nowadays, we can even travel to some of them. And we can create highly accurate images of ones we cannot yet travel to. So, while it is nowadays still possible to consistently maintain geocentrism by rejecting the right auxiliary assumptions, it would be manifestly conspiratorial.

But other areas of science—especially those aiming to answer some of our most fundamental questions—do not have such a luxury, as they deal with the less tangible. Physicists cannot, for example, take a photo of an electron's wave function. Electron orbital images are theory-laden computer-generated visualizations. Nor can neuroscientists directly observe how billions of interacting neurons generate conscious experience. Such scientific debates are not so easily settled.

Here is a tool to help settle such debates. The tool is best used in the context of an adversarial collaboration. Adversarial collaboration is collaboration between advocates of competing theories, in an environment of mutual respect and friendliness. In one form of an adversarial collaboration, there are two collaborators, one is an advocate of theory T, the other is the critic. The T-critic is trying to falsify T, the T-advocate is trying to maintain T by explicating and criticizing auxiliary assumptions. Logically, this forms a tree-like structure. In Figure 1, I have drawn the first four levels of such a structure. I call it a *falsification tree* because of how it builds on the logic of falsification presented earlier.

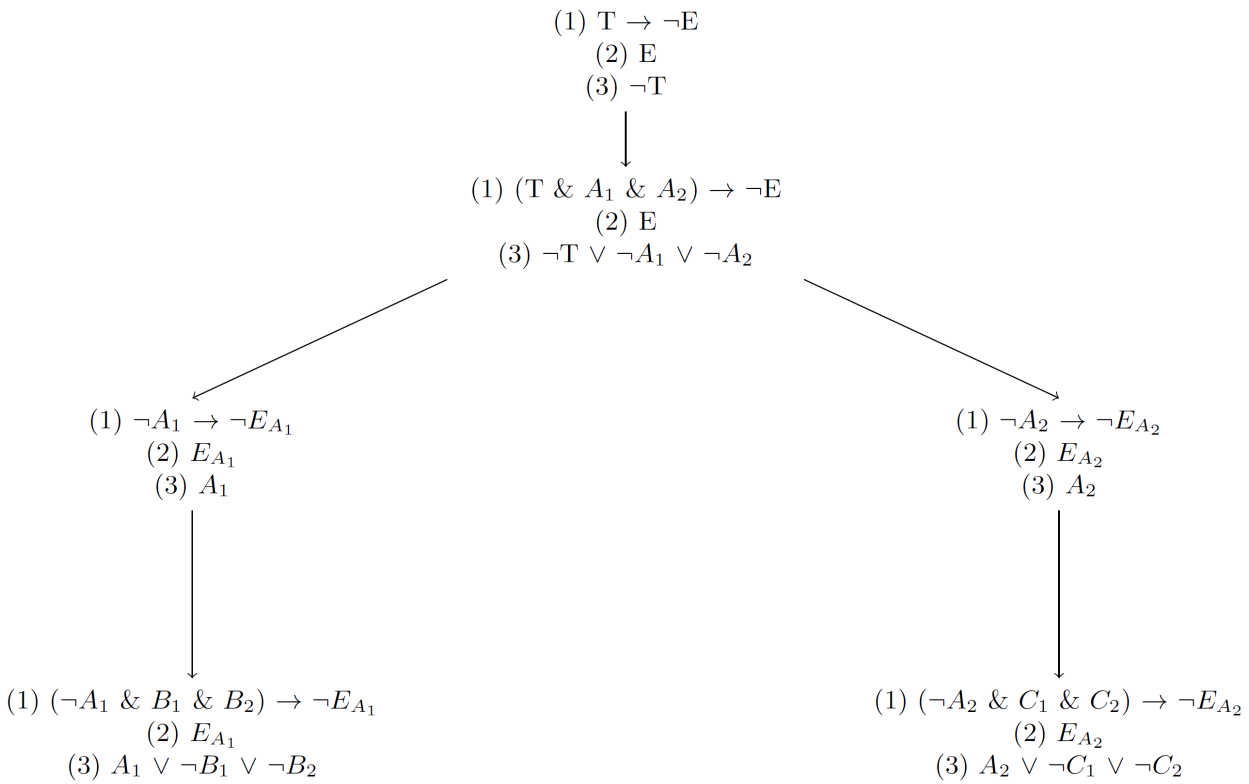

Figure 1: Falsification Tree. *Top level*: T-critic objects to T because of E. *Second level*: T-advocate identifies assumptions (A_1 and A_2) that can be rejected so that one can maintain T. *Third level*: T-critic identifies evidence (E_{A1} and E_{A2}) to argue *for* those assumptions. *Fourth level*: T-advocate identifies assumptions (B_1, B_2, C_1, C_2), that can be rejected instead. The tree may continue to branch indefinitely, but in practice branches should reach "dead-ends", where researchers can agree that defense of T no longer holds up, allowing researchers to explore other branches. A node may introduce many (more than two) assumptions and so create many branches.

The best way to understand the tree in Figure 1 is to substitute examples. In the trivial case, where T is "All leaves are green" and E is "I experienced blue leaves", A_1 might be "no leaves have been spray-painted". In that case, E_{A1} might say, "the area where the leaves were found show no trace of paint". B_1 might then say, "the spray painters covered their tracks," etc.

In the historical case, where T is Aristotle's geocentrism and E is "some planets exhibit retrograde motion", A_1 might be "planets to do not undergo epicycles". In that case, E_{A1} might refer to the changes in the apparent magnitudes of the planets, which appeared to vary in brightness as they moved through the sky, in a way that the epicycle theory could not explain. This was an anomaly with epicycle theory known prior to Copernicus. B1 might then be a modification of epicycle theory, that retains the spirit of the theory,

but accommodates the above-mentioned changes. Had this branch been more thoroughly explored earlier on, Aristotelian geocentrism may have been less influential.

The exact shape of the falsification tree may vary depending on the context of the adversarial collaboration. If two *teams* are collaborating, then the second level of the tree may identify many assumptions, thereby creating many branches at level three, which could be explored by smaller groups from each team. If only two people are collaborating, then the tree may be much thinner, with the collaborators deciding on which assumptions are more important to focus attention on. It is important to make the results of an adversarial collaboration public. If the tree is made public, then other researchers can explore existing branches and create more. The hope is that by exploring as many paths as possible we can more quickly and efficiently cut off those branches that lead to dead ends. The route through epicycles illustrates a route we took for far too long, before finally seeing it was a dead end. Rigorous adversarial collaborations that explore the branches of falsification trees are an antidote.

Contemporary case studies: consciousness science and quantum foundations

I conclude with some contemporary case studies: consciousness science and the foundations of quantum physics. These are areas of science where I see falsification trees being most useful. Both areas are plagued by significant fragmentation among researchers, where theories become isolated from critics, and are not properly critically analyzed.

In the science of consciousness. There are various contemporary neuroscientific theories of consciousness, like the integrated information theory, the global workspace theory, predictive coding theory, attention-schema theory, entropic brain theory, quantum theories of consciousness, etc. (To add another layer of complexity, there are the various competing philosophical theories of consciousness, like physicalism, dualism, panpsychism, idealism, illusionism, etc.) It is now well documented that most studies, conducted by advocates of these various theories, are not presented as testing predictions of a theory, but are instead attempts to interpret results in light of one's theory. Seldom do studies test predictions of more than one theory, trying to pit them against each other. The concern is that researchers can get so attached to their theory that they use it as a lens through which to see their data, much like Ptolemy did with Aristotelian geocentrism (and we know how that went).

In the foundations of quantum physics, consider the variation in how physicists interpret the groundbreaking experiments using entangled quantum states that were the subject of the 2022 Nobel physics prize. A common interpretation is that these experiments demonstrate that nature is nonlocal: there is what Einstein called "spooky action at a distance". But this interpretation rests on many assumptions, which others have rejected. It assumes there is no retrocausality. This is denied by retrocausal

interpretation of quantum mechanics. It assumes there is only one world. This is denied by the many worlds interpretation. It assumes that there is an objective reality in the first place that physics describes. This is denied by the various anti-realist interpretations. Realist interpretations, according to which nature is nonlocal, include pilot-wave theory and dynamical collapse theories. The fragmentation is bewildering. But it can be more efficiently reined in by adversarial collaborations that explore falsification trees.

These are complex areas of science and I have little space to delve into them. I will therefore conclude with some very brief illustrations of what can happen when we explore these debates in terms of falsification trees. Consider the many worlds interpretation (MWI) of quantum mechanics (understood in terms of the so-called Oxford-Everettian approach). Far from being unfalsifiable, the most common objection to this approach is that it has already been falsified by those experiments that collectively confirm the Born probability rule. Thus, critics have argued that if the MWI is true then the Born rule is false. MWI-advocates have responded that this claim rests on assumptions, for example, a frequentist theory of probability. Some MWI-critics seek to defend frequentism, thereby exploring one branch of this falsification tree. In other debates other assumptions have been the focus, meaning other branches have been explored. But in so many cases, branches get left hanging and are never resolved into dead-ends, while other branches never get adequately explored. Consequently, the debate over whether (this version of) the MWI is falsified by experiments confirming the Born rule seems to be at a stalemate. I believe a fully spelled-out falsification tree, constructed in the context of an adversarial collaboration, could quickly and efficiently move this debate forward.

For the consciousness case, I will illustrate with the integrated information theory (IIT). One interesting prediction of IIT is that it seems to place strong constraints on brain states underlying introspectively indistinguishable experiences: they must possess the same amount of integrated information. But it has been argued that experiences involving visual illusions (where the brain “fills in” features not encoded in the retina) can be indistinguishable from non-illusory experiences. Creating such non-illusory experiences simply involves arranging visual stimuli in a clever enough way. The problem for IIT is that brain states underlying such illusory experiences appear to possess far more integrated information than brain states underlying non-illusory experiences. This constitutes only the first node of a falsification tree for IIT. What we need to see more of in consciousness science, is branches of the falsification tree created and explored, not just for this proposed falsification, but for any that get through peer review. As mentioned above, this does not happen enough in consciousness science. But I believe that if we want to move these sciences forward—forward on the right trajectory—adversarial collaborations that rigorously lay out falsification trees is key.

In conclusion, I have argued that the inevitable inconclusiveness of theory falsification opens many pathways for science, some which may lead to the truth, others which may not. I have then proposed a way forward: the use of falsification trees in the context of adversarial collaborations. My hope is that this will accelerate science down a route that leads us to nature's most fundamental truths.