The Internet and Other Unreliable Oracles: Gettier Problems in the Disinformation Age

In 1963, Edmund Gettier discovered a series of special problems in epistemology, challenging a traditionally accepted definition of knowledge as “justified true belief.” One year earlier, J.C.R. Licklider, a scientist from MIT, proposed a "galactic network" of interconnected computers that foreshadowed the Internet and would ultimately reshape how knowledge is spread through society. The Internet is notably agnostic about whether it spreads information or disinformation, making it a sort of unreliable oracle that may itself typify a special subset of Gettier problems.

I. Gettier problems

The Platonic definition of knowledge as "justified true belief" presents a bevy of well discussed complications. Gettier problems are one such area of trouble. Classic Gettier problems involve someone believing a chain of facts for justification where multiple flawed premises (or the ignorance of some other fact) cancel out, leaving the deduced belief both true and justified, though with a justification that seems deeply problematic on close consideration.[1]

In a classic example, Smith and Jones are applying for a particular job. Smith has strong evidence that Jones will get the job (perhaps the President assured Smith that Jones would get the job). Smith also has strong evidence that Jones has ten coins in his pocket (perhaps Jones pulled them out and counted them ten minutes ago in front of Smith). Smith concludes, reasonably, that the candidate with ten coins in their pocket will get the job. Unbeknownst to Smith, Smith now has 10 coins in his pocket (perhaps Jones, an amateur magician and pickpocket, quietly slipped the coins in Smith’s pocket for practice). Surprisingly, Smith was selected for the position (perhaps Jones practiced his pickpocketing tricks a bit too noticeably in the interview). Smith’s conclusion is correct, and it has a valid justification, though the supporting components of the justification have issues.

At the fundamental level, Gettier problems involve questions about what constitutes sufficient justification. How should we classify true beliefs that rest on a compelling and reasonable (though subtly flawed) justification?

Some have expressed skepticism[2] regarding the practical relevance of Gettier problems, though a full understanding of incremental knowledge may become increasingly necessary for the verification of claims from intelligent systems.[3]

II. Internet problems

Internet debates are popularly regarded as a source of strongly argued but flawed positions, [4], so this domain might provide relevant case studies for understanding the epistemological status of Gettier problems.

To explore this area, consider how we acquire beliefs in the first place.

III. Origins of beliefs

There are many sources of beliefs. We take on beliefs from parents and communities and learn others through education or direct observation. We conduct experiments, whether in formal settings, or informal, through simple trial and error or making advance predictions about anticipated causes and effects. We arrive at others through internal contemplation. These sources of belief can be roughly categorized into transmission, observation, or deduction.

A dominant portion of the beliefs we claim to know are actually gathered from allegedly competent and trustworthy authorities. Conversations with parents, teachers, authors, narrators, pundits, neighbors, peers, or colleagues provide us a massive series of claims, with an innumerable number of implied secondary claims. The contribution of private deductive realizations and eureka moments is undoubtedly nonzero, constituting the origin of some of the greatest discoveries of humanity, but likely constitutes a small fraction of one's web of ontological commitments.

For example, although I have not visited the moon, I do believe some Apollo-era American flags stand there,[5] and that these missions were not simply filmed on a Hollywood sound stage. I feel justified in these beliefs based on the reliability of the sources I have considered, including works of nonfiction about the space program, videos, interviews with eyewitnesses, and the implausibility of the alternative explanations and claims.[6]

Held propositions are provably finite,[7] but nonetheless troublesome to quantify and categorize. If direct observation is constantly contributing atomic beliefs it presumably outnumbers all other sources through its sheer flood of information: "I see a blob of redness," "I hear a rhythmic thud resembling a kick drum," "I feel through proprioception that my hand rests lapward," etc. However it also seems likely that only a small portion of this stream is faithfully retained as an enduring update to one's ontological commitments. In studies analyzing recall or witness testimony, details are frequently forgotten or altered even moments after an event.[8][9][10]

Transmission arguably contributes a significant part of our most meaningful ontological updates and therefore appeals to authority are prevalent in our justifications.

As to deduction, developmental psychologists as early as Piaget argued that our foundational beliefs are formed when we are unable to systematically employ deductive logic or scientific reasoning.[11] Though later developmental psychologists like Harris and Gopnik criticized Piaget for oversimplification and found evidence that scientific reasoning begins earlier than Piaget assumed, later authors still acknowledge a dominant role for testimony in forming our earliest beliefs.[12][13]

The conformity studies of Solomon Asch suggest people privilege transmission over even their direct observation, or at least are willing to behave as though they do.[14] Achen and Bartels have used a variety of studies to argue that voters form political beliefs through family and community transmission rather than through careful analysis, noting the high rate at which incumbents are punished for events clearly outside their control, such as inclement weather or shark attacks.[15]

IV. Testimonial Gettier Problems

The prominence of reliance on authority undermines the Platonic definition of knowledge in ways that evoke Gettier problems. Gettier problems can be conjunctive or disjunctive. If most of our beliefs are testimonial, that means that most of our beliefs are conjunctive justifications along the

following lines: "(P1) Person X claims Q & (P2) Person X is a reliable source of knowledge, therefore (3) Q." These sorts of conjunctive propositions are openings through which we can smuggle in a certain category of problem, the Testimonial Gettier. It's easy to imagine how one might misplace trust in a source of information, but this also highlights novel unexpected results in the context of debates.

V. Debate and knowledge laundering

Suppose two noted astrophysicists, Carl S. and Subrahmanyan C., live on a certain street. A neighbor, Bob, lives between them. Bob, although not an expert, knows their reputation, and has seen direct evidence to attest to their high level of domain expertise (say judging the coherence and insightfulness of responses in conversation, observations of throngs of reporters congratulating them for new discoveries, faces on magazine covers, or similar.)

During a neighborhood party, the topic of intelligent extraterrestrial life comes up. Carl explains the Drake equation, focusing on the extreme vastness of space and on the inconceivably high number of possible planets where it might have emerged. Carl explains that has convinced him that there is near certain likelihood of life arising somewhere other than Earth. Subrahmanyan responds by explaining the Fermi paradox, stressing that if life were not unique, it would almost certainly be common, at least enough such that we should have certainly detected its signs by now. Each continued to give compelling extrapolations and rebuttals, until half the neighbors left for the night firmly convinced that Carl had the better arguments, and half the neighbors left convinced that Subrahmanyan was correct. Every neighbor left with a thorough understanding of what happened that night, to include at least one strong case explained carefully by a relevant and highly competent authority, for one side or the other.[16]

Now it just so happens that in our universe, extraterrestrial intelligent life either does or does not exist.[17] Whichever the case, since all neighbors left with a strongly justified belief, according to the strict Platonic definition, it follows that exactly half the block actually knows whether intelligent life exists or not.[18]

It's troubling enough to think that, through the above chain of reasoning, we have now vested some dozen people with knowledge about a seemingly inscrutable secret of the universe almost at random. What's more troubling is that if we try to go back and add precision to this series of events, there's a risk of uncovering a more perverse and more general issue.

One way to resolve the somewhat similar paradox of the preface is to acknowledge that most of our beliefs are probabilistic, and accept that conjunctions should naturally lower our confidence. So we might expect that to help with these suspicious conjunctive conclusions. Carl and Subrahmanyan were almost certainly arguing from probabilistic beliefs, and maintained considerable uncertainty themselves. They might never claim so much as even a 90% certainty in either direction on the subject. Carl might be convinced by the conjunction of (C1 + C2 + … + Cn), while Subrahmanyan was more moved by (S1 + S2 + … + Sn). Assume each stated their strongest premises and arguments. The audience will have a unique benefit over either speaker. Team Carl will have the benefit of (C1 + C2…) but also add C(n+1): this view is endorsed by one of the smartest astrophysicists I know. Subrahmanyan's followers will similarly be able to add a premise to their justification. In other words, the audience's justification has the unique benefit of reference to a

competent authority, which might be more compelling than the underlying deduction that brought each expert to the original position. Should an audience's claim to knowledge be stronger than the authority that taught them the claim?.

VI. Possible limits on the persuasiveness of near-perfect oracles

Some time after the block party, one neighbor, Claude, finally completes his latest project, GPT-99, a superintelligent AI. For AI safety reasons, Claude only allows this oracle to communicate via a single bit of information, flashing a light once or twice respectively, in response to rate-limited yes or no questions. Claude included some undisclosed safeguards that lead to random answers in some situations.[20] Claude allows C and S shared use of the machine while he's on vacation. The astrophysicists determine who will use the machine by flipping a fair coin each morning.

After a few weeks of testing and experiments, the astrophysicists compare results. Carl insists that the machine is astounding, producing perfectly accurate responses on verifiable questions, at ~95% of the time. Subrahmanyan agrees. They both acknowledge that nonsense questions and paradoxes, including many questions about the future, seem to generate unpredictable results, as expected. Neil and Carl run into trouble when they compare notes on the machine's failures. They had each flagged a series of questions as receiving "certainly false" results. When reviewing their logs, they found they strongly disagreed about which items were actually answered incorrectly.

Some readers will struggle to entertain the possibility that two intelligent agents can ask a mostly accurate oracle the same question, and leave convinced opposite things are true. This probably happens more often that we would like to admit.[21]

Let's consider more realistic examples.

I am highly confident I can ask the Internet the birthdate of Miles Davis and get an accurate answer, gaining a justified true belief in the process.

I am also highly confident I can use the Internet to confirm that most global elites are not secretly bipedal lizards, a notable conspiracy theory. Adherents of conspiracy theories are generally similarly confident that the Internet is a source of confirmation, simply arriving at the different conclusions.

This should not be read to give credence or equal standing to conspiracy theories, whose beliefs frequently have the notable additional feature of being false. However, many true and false beliefs are nonetheless justified through roughly similar means: consultation with a probabilistic oracle, one that is mostly reliable but flawed.

VII. A reliable Internet invites epistemological collapse

It is tempting at this point to simply dismiss the Internet as unreliable. However, that seems hard to square with the Internet's effectiveness at transmitting and verifying uncontroversial information. I would trust anyone to ask Google the birthdate of Miles Davis or Joe Strummer and gain knowledge through that process. One can presume this reliability likely extends to facts beyond the birthdates of musicians. The overwhelming majority of straightforwardly verifiable accumulated human knowledge benefits from Internet transmission. Knowing which times the Internet is more or less likely to lead one astray is a difficult art, honed through embarrassing mistakes, encounters with trolls, and some

intuition about how much skepticism one ought to deploy while refining one's ontology to guard against the water hazard of credulity towards misinformation while avoiding the abject nihilism.

As I encounter uncertainty in my everyday life, I increasingly consult search engines or forums or other parts of the Internet for validation. The ready access to a bolstered justification I have found largely positive (save for wikipedia in our pockets tragically killing late night "bar debates" among friends). I know it can be risky to consult this unreliable oracle, but I also know the risks and read defensively to avoid them.

As a consequence, more and more of the things I believe, I believe because I have consulted a powerful but imperfect oracle on the topic. My justification for maintaining all my beliefs is converging to the same process, the same source. "Why do you believe that?" "Well, I looked it up on the Internet" (and, implied, "took all appropriate precautions while reading to not fall into a web of simple lies or biased sources or conspiracy theories").

If all my beliefs hold this same justification, though, then sorting which of my beliefs constitute Platonic knowledge and which do not collapses to determining which of my beliefs are true. No small feat, but collapsing "JTB" into "TB" deftly sidesteps all Gettier problems by eliminating half of the equation (albeit in a deeply unsettling way).

Seeing this spurious conclusion, one might revisit and reject the premise of a mostly reliable internet. Many skeptics already vocally insist it should not count as a reliable source of justification. If that holds though, then we end up in a different unsettling situation. If a growing share of our beliefs are derived from a flawed source and lack justification, then a growing share of our beliefs simply aren't knowledge, and again are merely true or false. Outside of our own personal experience, it's as if we have each just adopted a series of beliefs at random, some of which happen to be true by pure coincidence. If no one gains justification for their beliefs through unreliable authorities, this also resolves most Gettier problems, in an equally unsatisfying way.

VIII. Epistemological collapse begins at home

In the preceding sections, the internet is described a bit like a belief factory. One consults this oracle, and it manufactures new beliefs directly inside your brain. Obviously the process is more complex than that, involving constant scrutiny over signals of reliability or misdirection on each individual page or sentence. We can draw a box around that whole process and call it a belief factory that is either reliable or unreliable.

Each of us has some larger process for evaluating all incoming evidence about the world and sorting it into ontological commitments. Even though we might acknowledge cognitive biases and other flaws in that machinery, we nonetheless believe our beliefs. When asked why we believe a certain thing, we might answer with a conjunction of all inputs and situations that caused us to update. Or we could collapse that all to, "well, I have a sort of inquisitive and rational approach to fact finding, and I'm especially rigorous on claims in this general area, so I'm a trustworthy source on this topic. Since I believe it, and my ontology is generally trustworthy, it's probably true."

Since that justification could be applied to all of one's beliefs, it collapses JTB even more destructively, since it applies across all beliefs, not merely those gained from spending too much time online.

IX. Why get so hung up on knowledge anyway

One possible solution, or retreat, is that the platonic conception of knowledge simply does nto matter very much after all. Maybe once we consider carefully the strength of someone's justification, and preserve some full account of that, we can simply stop. Say someone believes P with a strong, very strong, or quite weak justification (or whatever more precise terms we prefer), and call that enough to meet all reasonable needs without drifting into the sorites paradox of what level of justification or confidence is strictly required before something gains the imprimatur of knowledge.

Maybe that's the Bayesian way out. More or less reject the concept of Platonic knowledge completely. Assign all beliefs some probabilistic measure of strength, then be done. When we claim to "know" something, it's just a particularly emphatic way of stressing high confidence in that belief. Those committed to knowledge realism might try to pick a certain threshold of confidence whereby anything above that should be considered knowledge, but may run into issues with such attempts at thresholds, per Kaplan, M. (1981). Rational acceptance. Philosophical Studies, 40(2), 129-145.

[1] Gettier, E. L. (1963). Is justified true belief knowledge?. Analysis, 23(6), 121-123.

[2] Or outright hostility. See Tamler Sommers' comments, in, for example, Very Bad Wizards #112. https://www.verybadwizards.com/112

[3] Irving, Christiano, and Amodei. AI safety via debate. https://arxiv.org/abs/1805.00899

[4] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. https://www.science.org/doi/10.1126/science.aap9559 The study examined the velocity of true and false information on Twitter across 4.5 million tweets, finding false information spread much faster and further than true information, likely due to its perceived novelty. Memes and comics periodically cite this as a universal experience: https://xkcd.com/386/

[5] Actually, this is a somewhat problematic example, because some have theorized that the harsh conditions have bleached the flags into unrecognizability. Those who value this sort of precision can rehabilitate the example by substituting "flag remnants" where appropriate.

[6] 5%-30% of readers may disagree with me, depending on the country surveyed. Skeptics include one Tennessean who ambushed Buzz Aldrin to call him a liar and received a punch in the face for his troubles. https://www.snopes.com/fact-check/buzz-aldrin-punched-conspiracist/ That Mitchell and Webb Look (Season 4, Episode 2) Moon Landing Sketch details an alternative scenario, and should be considered useful alternative history here. https://www.youtube.com/watch?v=P6MOnehCOUw

[7] Smullyan, R. M. (1978). What is the name of this book? The riddle of Dracula and other logical puzzles. Item 243, A proof that you are either inconsistent or conceited.

[8] Loftus, E. F. (1979). The malleability of human memory: Information introduced after we view an incident can transform memory. American Scientist, 67(3), 312-320.

[9] Schacter, D. L. (2001). The seven sins of memory: How the mind forgets and remembers.

[10] Garrett, B. L. (2011). Convicting the innocent: Where criminal prosecutions go wrong.

[11] Piaget, J. (1930). The child's conception of physical causality.

[12] Harris, P.L. (2012). Trusting what you're told: How children learn from others.

[13] Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. Science, 337(6102), 1623-1627.

[14] Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological monographs: General and applied, 70(9), 1-70.

[15] Achen, C. H., & Bartels, L. M. (2016). Democracy for realists: Why elections do not produce responsive government. Princeton University Press.

[16] Suppose for the sake of illustration that dark forests, grabby aliens, or recalculations of uncertain parameters did not come up at this, say, circa 1982 block party. (a) Paradis, Justine (18 February

2022). "Outside/In[box]: What is the Dark Forest Theory?". New Hampshire Public Radio. (b) Hanson, Martin, et. al., (2021). If Loud Aliens Explain Human Earliness, Quiet Aliens Are Also Rare. (c) Sanders, Drexler, Ord. (2018). Dissolving the Fermi Paradox. https://arxiv.org/abs/1806.02404v1

[17] Aristotle. (ca. -40). Organon, On Interpretation.

[18] But... obviously not? This superficially recalls the Wanamaker paradox, for the Department Store magnate who allegedly said something like, "I know half the money I spend on advertising is wasted, just not which half." It perhaps more strongly evokes the paradox of the preface[19], where authors suggest in the preface that they acknowledge there is likely some error left in the book, despite having reviewed and believing each of the statements in the book individually, because each situation involves a conjunction of claims where the confidence in the conjunction appears to diverge significantly from the confidence in the individual claims.

[19] Makinson, D.C. (1965). The Paradox of the Preface. Analysis, 25(6), 205-207.

[20] This is probably an unsafe protocol and is not included here to suggest that this protocol would be sufficiently effective. See, for example, Yudkowsky's AI-Box Experiments on escape. https://yudkowsky.net/singularity/aibox/ For an introduction to AI safety issues, see Yudkowsky's "Artificial Intelligence as a Positive and Negative Factor in Global Risk," 2008. Rate limiting might be a useful component of some future safety protocol, but notably it does not operate on the ratio of harms relative to benefits of any system.

[21] "They Saw a Game" is illustrative reading on the topic of differing conclusions by different groups in response to identical evidence and the impacts of bias, notwithstanding its eyebrow-raising endorsement of hard relativism. Hastorf, A. H., & Cantril, H. (1954). They saw a game; a case study. The Journal of Abnormal and Social Psychology, 49(1), 129–134. https://doi.org/10.1037/h0057880