Rational Confidence with Unknown Unknowns

Isaac Swift*

October 12, 2023

Updated version at: Link

Abstract

Overconfidence is among the most well studied behavioral biases. When graphing confidence against ability or knowledge, many have found it to be too flat, downward sloping, or even hump-shaped. In surveys, as many as 90 percent of respondents may report they believe they are "above average" at a given skill/topic and even above the median. This paper characterizes all rational confidence-knowledge graphs and what fraction of the population can rationally believe they are above the average in the presence of unknown unknowns. When learning any topic, one does not initially know how information there is on the topic to potentially learn. The unknown unknowns are the things you do not know that you do not know. I find, that in a world with unknown unknowns, there is no restriction on what confidence-knowledge graphs can be generated by fully rational Bayesian agents or what fraction of these agents can think they are above average, even when they are not. This result forces us to question if any studies documenting overconfidence are actually evidence of a behavioral bias.

JEL Classification: D83, D91, C44

*Hong Kong Baptist University, isaacswift@hkbu.edu.hk. I'm grateful for helpful comments from Benson Tsz Kin Leung, Kim-Sau Chung, and Balazs Szentes.

1 Introduction

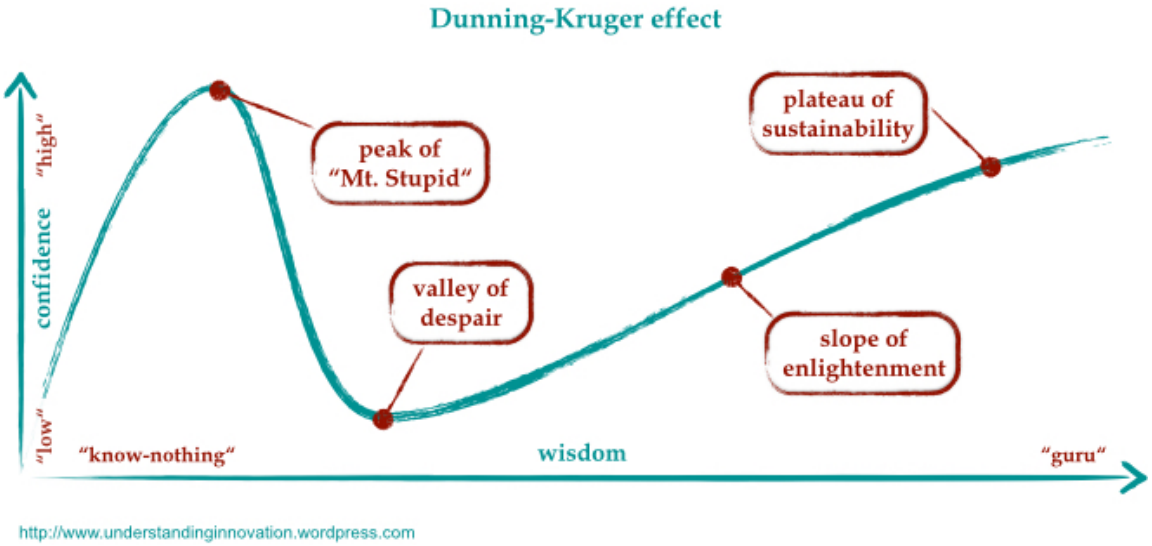
Among the most well documented behavioral biases is overconfidence. This paper studies the limits of what a rational level of confidence can look like. Though it sounds simple, measuring confidence becomes more complicated in situations with what I will call unknown unknowns, or situations where you don't know how much information there is available to potentially learn. Before I elaborate on how I model unknown unknowns, consider the following studies of confidence that have led to two types of seemingly irrational behavior.

First, it has been frequently observed that people with a small amount of knowledge about a subject overestimate what they know relative to their peers, while people with a lot of knowledge underestimate what they know. This type of result is often understood by a graph with confidence on the y-axis and knowledge or skill on the x-axis. The most well-known study of this type is Kruger and Dunning (1999). They have all the participants in their study take a test. Before telling them the results of the test, they have the participants guess what score they got. This plot of guesses vs actual scores is an example of what I'll call the confidence-knowledge graph. If everyone's guess was correct, they would simply fall on the 45 degree line. Of course, not everyone will be correct. Some will guess too high. Some will guess too low. Nonetheless, one might expect that if the participants are rational, these mistakes should even out in a way. The best fit line should still be the 45 degree line. However, the 45 degree line is not what they found. Participants with low scores on the test systematically overestimated their ability. Conversely, participants with high scores underestimated their ability. The result was a knowledge-confidence graph that was significantly flatter than the 45 degree line (though still upward sloping), and shifted up. This result of people with low knowledge or ability being overconfident and people with high knowledge or ability being underconfident has been documented in many situations and dubbed the "Dunning-Kruger effect," one of the most well cited behavioral biases.

Common opinion is that the confidence-knowledge graph can look quite counter-intuitive. Charles Darwin noticed it and remarked, "Ignorance more frequently begets confidence than

does knowledge." Mark Twain made a similar observation, "When I was a boy of 14, my father was so ignorant I could hardly stand to have the old man around. But when I got to be 21, I was astonished at how much the old man had learned in seven years." In *As You Like It*, William Shakespeare put it clearly, "the fool doth think he is wise, but the wise man knows himself to be a fool." All these quotes seem to point to the idea that the confidence-knowledge graph can be, not only too flat, but even downward sloping.

Any quick search on the topic will lead to many confidence-knowledge graphs that look even more counter-intuitive. To illustrate, consider the journey most of us go on as we learn economics. You take ECON 101. The class is easy for you. You come out of it feeling like you are pretty smart an learned quite a bit of economics. You take the rest of the the econ major. You ace this as well. You now have a degree in economics and feel like you must know nearly everything there is to know in the field. You start you PhD coursework. You realize, there's a lot more to economics that you weren't aware of. You don't know as much as you thought, but you still get through it. You start to work on your dissertation and wonder if you learned anything in all your classes. You get your first academic job and research without any supervisor. You are in the pit of despair as you realize you know nothing. This hump-shaped confidence-knowledge graph is abundant.



Are these observed confidence-knowledge graphs evidence of a behavioral bias? Does the

confidence-knowledge graph of rational Bayesian agents need to look like the 45 degree line. I show in this paper that it does not. The first objective of this paper is to bound what confidence-knowledge graphs can be generated by Bayesian updating of beliefs.

The second type of seemingly irrational behavior I want to talk about is what is known as illusory superiority, see Buunk and VanYperen (1991). It is also referred to as the Lake Wobegon effect, as Lake Wobegon was referred to as a place where "all the women are strong, all the men are good looking, and all the children are above average." We've all heard of the studies showing 88% of American drivers are above average Svenson (1981), or countless other similar results. This result persists whether they compare themselves to the mean or median. Seemingly by definition, the subjects in the study must be irrational. On the flip side, in certain fields the opposite of the Lake Wobegon effect is also observed. This is exemplified by the fact that nearly everyone thinks they're a below average speller. In Kruger (1999) they show this opposite effect for juggling, riding a unicycle, and other tasks. The second objective of this paper is to bound what fraction of the population can believe they are above average. Can rational Bayesians be systematically wrong about being above average?

In some fields we observe that a majority of people think they are above average, while in other fields a majority think they are below average. We also see in many cases, people with a small amount of information are more confident than people with a large amount of information. These facts appear at first glance to be cognitive biases inconsistent with rationality. However, I will present a model with fully rational Bayesian agents with common priors where these facts can still arise. The key ingredient of the model is what I will call unknown unknowns. The term comes from a quote by United States Secretary of Defense Donald Rumsfeld. He was talking about terrorist threats when he said, "We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know." The idea is that not only do we not have all the information, we don't always know how much information

is available to be learned in the first place. An example will make the logic clear.

2 Example

You come up with some new research topic you want to study. Since you've never worked on this topic before, you don't know how much work has been done previously on the topic or even if anyone has studied this topic. It could be completely new. So, you learn about the topic by looking for and then reading papers written on the topic. You'll continue to read papers and learn about the topic until you are stopped for one of three reasons. First, you'll stop if no more papers have been written on the topic. You don't know ex ante how many papers there are on this subject. Second, you'll stop if your search fails. Even if there are more papers on the topic, there is a possibility that you don't find them. Note that these two reasons are indistinguishable from your perspective. If you don't find another paper, it could be that another paper doesn't exist or that you just failed to find it. Third, you only have a limited amount of time and don't want to be reading papers forever. So, you stop if your run out of budget, which in the example you can think of as a time budget.

This example has the known knowns, known unknowns, and unknown unknowns from the quote in the introduction. The known knowns are papers that you find and read. It's knowledge or skills that you acquired. The known unknowns are papers that you know about because you found them, but you don't know the content of because you ran out of budget and didn't read them. This is knowledge that you don't know, but you know the knowledge exists and that you don't know it. There are also unknown unknowns. When you stop reading papers, you don't know if there are more papers out there that you didn't find. If there is such a paper, that is knowledge you don't know that you also aren't even aware of the existence of. Of course, if you're Bayesian, you realize the possibility and assign some probability to the event, but you don't know if such paper even exists.

Now let's put numbers to the example, so we can compute rational confidence levels and

get a flavor of the results. Let's say there could be zero, one, two, or three papers on the topic. You don't know ex ante how many papers there are, but your prior is that each of these are equally likely. You have a budget of two. This means that even if there are three papers, you will only read two of them. Also, your search is imperfect and modeled as a constant hazard rate of failure. So at each stage (zero, one, or two), if there is another paper, you have a 2/3 chance of finding it.

2.0.1 Confidence-knowledge graph

Let's compute the confidence-knowledge graph using Bayes' rule. We'll measure confidence as the posterior probability that you read all the papers. Knowledge will be measured by the number of papers you read. Someone who read zero papers doesn't know if there actually aren't any papers (25 percent chance ex ante) or there are papers that they just didn't find (1/3 chance conditional on papers existing). Bayes' rule gives the confidence level posterior probability.

$$\pi(0) = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{3}{4}\frac{1}{3}} = 50 \text{ percent}$$

Someone who has read zero papers believes there is a 50 percent chance that they know every paper on the topic (because there aren't any). In the same way you can compute that the confidence level is higher for someone who has read one paper.

$$\pi(1) = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{2}{3}\frac{1}{3}} = 60 \text{ percent}$$

So more knowledge, at least initially, led to higher confidence.

When considering someone who has read two papers, there are two different approaches we could take. Both give the same result. First, since they have exhausted their budget, you could say they will not search for a third paper. They wouldn't have time to read it anyway. Two or three papers existing were equally likely ex ante. They have received no information that helps them distinguish between the two situations. Thus, their confidence

level is 50 percent. Alternatively, you could suppose that they still search for the third paper even though they won't read it. They just want to know if it's there. If they do find a third paper, their confidence level is 0. They know for sure that they don't know everything. If they don't find a third paper, applying Bayes' rule in the same way gives a confidence level of 75 percent. Now using the probability of there being a third paper and the fraction of the population that would find it, you can compute the average confidence level of people who read two papers.

$$\pi(2) = \frac{1}{2}\left(\frac{3}{4}\right) + \frac{1}{2}\left(\frac{2}{3}(0) + \frac{1}{3}\left(\frac{3}{4}\right)\right) = 50 \text{ percent}$$

We arrived at the same average confidence-knowledge graph either way. Despite the agent being fully rational, the confidence-knowledge graph is not monotonically increasing, but hump-shaped.

2.0.2 Comparison statistics

Now we want to compute the fraction of the population that believes they are above average. So that the question makes sense, suppose there is a unit mass of agents that had the common prior described and all learned independently according to the described process. After the learning is complete, imagine asking each agent, "Are you strictly better than average?" If the agent guesses correctly, they get a payoff of one. If they guess incorrectly, they get a payoff off zero. In other words, the agent needs to pick if they think it is more likely they are above average or not.

First consider people who haven't read any papers. At best, there aren't any papers and they're tied for the average. At worst there are papers and they are strictly below average. These people would have to guess "no" they are not above average. If there actually aren't any papers, we would have 100 percent of the population guessing that they are not above average. This isn't actually very strange yet, because they are all correct.

Consider now the people who have read one paper. We computed already that there is a 60 percent chance of only one paper existing. If there is only one paper the average number of papers read in the population will be strictly lower than one. There is only a 40 percent chance of there being more than one paper. Thus, people who have read one paper must guess "yes" they think they are above average. If there is actually one paper, these people would make up two thirds of the population. So two thirds of the population thinks they are above average, and one third thinks they are below average. Again everyone's guess is correct.

Finally consider people who have read two papers. These people know they are above average. Even if there is another paper that they haven't read, they know that no one will have read it because they are at the maximum of the budget constraint. If there are actually two or three papers, 33.3 percent of the population will have read zero, 22.2 percent will have read one paper, and 44.4 percent will have read two papers. This means that about 67 percent of the population will say that they believe they are above average. This one is a little strange because only 44 percent of the population is actually above average. The actual average here is about 1.1 papers. This means that the people who read one paper believed they were above average but were actually below average.

2.0.3 Conclusion

The example had rational agents learning in a situation with unknown unknowns. Because of the presence of unknown unknowns, the confidence-knowledge graph did not need to be the 45 degree line or something analogous. In fact the graph does not even need to be upward sloping and can be hump-shaped as is often assumed in the world. In the more general model in the next section I will show which confidence-knowledge graphs can be rationalized in this way and which are actually evidence of bias. The fraction of the population that believes they are above average did not need to be 50 percent. More tellingly, it did not even need to be equal to the fraction that are actually above average. With unknown unknowns a significant

fraction of the population can appear ex post to be overconfident or underconfident. In the next section we will see what bounds there are on the fraction of the population that can believe they are above average.

3 Model

The general model is still very simple and analogous to the example. The results are just a basic statistical property.

Time is continuous. There is a mass of rational agents that learn about some subject. There is some total amount of information available, t^I. The total information, t^I, is unknown, but everyone has a common prior over its distribution, $\mathcal{I}(t)$. Each agent has an alternate stopping time, t_j^S. This time is drawn independently for each agent, but from an identical distribution, $\mathcal{S}(t)$. This time, t_j^S, is also unknown. t_j^S represents the time at which you would stop learning due to any other reason, conditional on not having already stopped due to t^I. So t_j^S is the time when you would stop because your search failed, you exhausted your budget, or any other reason. Simply put, t_j^S is the minimum of all reasons to stop other than t^I. It may seem at first odd to group them like this, particularly because your search failing is indistinguishable from t^I but exhausting your budget is not. You might know your budget from the very beginning. I show in the appendix however that it does not effect any of the results. You could separate the budget out as a different stopping time that is known to the agent and distinguishable from t^I, and it won't change anything in the end. As long as at least one of the variables is unknown, that is sufficient.

You stop learning at the minimum of the times t^I and t_j^S. The key point is that when you stop, you don't know which of the two stopping times stopped you. You've learned until you don't find anything else to learn. You can't tell if you've learned everything on the subject or if there is more to learn and your search was unsuccessful. This is the entirety of the model. We now want to compute the confidence-knowledge graph and the fraction of the

population that believes they are above average.

Your knowledge can be measured by the amount of time you spent learning, $\hat{t} = \min\{t^I, t_j^S\}$. Call the distribution of $\hat{t}$ the stopping time $F(t)$. $F(t)$ is derived from $\mathcal{I}(t)$ and $\mathcal{S}(t)$. Your confidence level could be measured in many different ways. For now let us consider the posterior probability that you have all available information as your confidence level, but other measures will be considered later.

Definition 1. Given some distributions, $\mathcal{I}(t)$ and $\mathcal{S}(t)$, the **confidence-knowledge graph**, $\pi(t) : \mathbb{R}_+ \to [0, 1]$, is the posterior probability of having all information conditional on stopping at time t.

So, the confidence-knowledge graph is just the probability that $\hat{t} = t^I$. We'll define the confidence knowledge-graph to be equal to zero at any t that cannot be reached.

We want to say that an allowable function is rationalizable if it could be the confidence-knowledge graph generated by fully rational Bayesian agents with a common prior.

Definition 2. A function $g : \mathbb{R}_+ \to [0, 1]$ is **rationalizable** if there exists some distributions $\mathcal{I}(t)$ and $\mathcal{S}(t)$ such that $g(t)$ is almost everywhere equal to the confidence-knowledge graph of those distributions.

3.1 Results

The first result is about what functions are rationalizable as a confidence-knowledge graph.

Proposition 1. *Any function $f : \mathbb{R}_+ \to [0, 1]$ is rationalizable.*

The posterior probability of having all information is simply the ratio of hazard rates.

$$\pi(t) = \frac{\sigma_I(t)}{\sigma_I(t) + \sigma_S(t)} \tag{1}$$

where $\sigma_G(t) = \frac{g(t)}{1-G(t)}$ is the notation for the hazard rate of distribution $G(t)$. Since hazard rates are nearly unconstrained, there will always exist distributions that give the desired

confidence-knowledge graph. For a given function, $f(t)$, all we need to do is find distributions such that the hazard rates satisfy the following condition.

$$\sigma_I(t) = \frac{f(t)}{1 - f(t)}\sigma_S(t) \tag{2}$$

There are infinitely many such distributions. The simplest is

$$\mathcal{S}(t) = 1 - e^{-t} \tag{3}$$

and

$$\mathcal{I}(t) = 1 - e^{-\int_0^t \frac{f(\tau)}{1-f(\tau)}d\tau}. \tag{4}$$

Additional details needed for the proof are given in the appendix.

This result means that flatter the confidence-knowledge graphs found in Kruger and Dunning (1999) are not evidence of a behavioral bias. Neither is a downward slope or a hump-shaped confidence-knowledge graph as is often believed. No results based on observed confidence-knowledge graphs can be taken as evidence of irrationality in a world of unknown unknowns.

3.1.1 Intuition

Consider $\mathcal{I}(t)$ and $\mathcal{S}(t)$ both being exponential distributions. Since these are memory-less distributions, the probability that you are stopped due to having found all possible information is constant across t. This constant level can be any level between zero and one. The level is simply, the ratio of the arrival rate of $\mathcal{I}(t)$ to the sum of the two arrival rates. If $\mathcal{I}(t)$ has a higher arrival rate (which is equal to the hazard rate) than $\mathcal{S}(t)$, this level is above 50 percent. If $\mathcal{I}(t)$ has a lower arrival rate, this level is below 50 percent. Thus a "bias" up or down in the confidence-knowledge graph or a graph that is very flat needn't be thought of as unusual at all.

For the confidence-knowledge graph to slope up (or down) all we need is for the hazard rate of $\mathcal{I}(t)$ to be increasing (decreasing) relative to the hazard rate of $\mathcal{S}(t)$. A hump-shaped confidence knowledge graph necessarily comes from distributions where the hazard rate of $\mathcal{I}(t)$ is first increasing and then decreasing relative to the hazard rate of $\mathcal{S}(t)$. This was the case in the example from the introduction. The hazard rate of the total amount of information at a given number of papers (probability of having only that number of papers conditional on having at least that many papers) increased from 25 percent at zero, to 33 percent at one, to 50 percent at two, and finally 100 percent at three. The hazard rate of stopping for other reasons was constant at 33 percent for zero or one papers, but jumped to 100 percent at two papers due to the budget. This explains the hump-shape confidence-knowledge graph in the example.

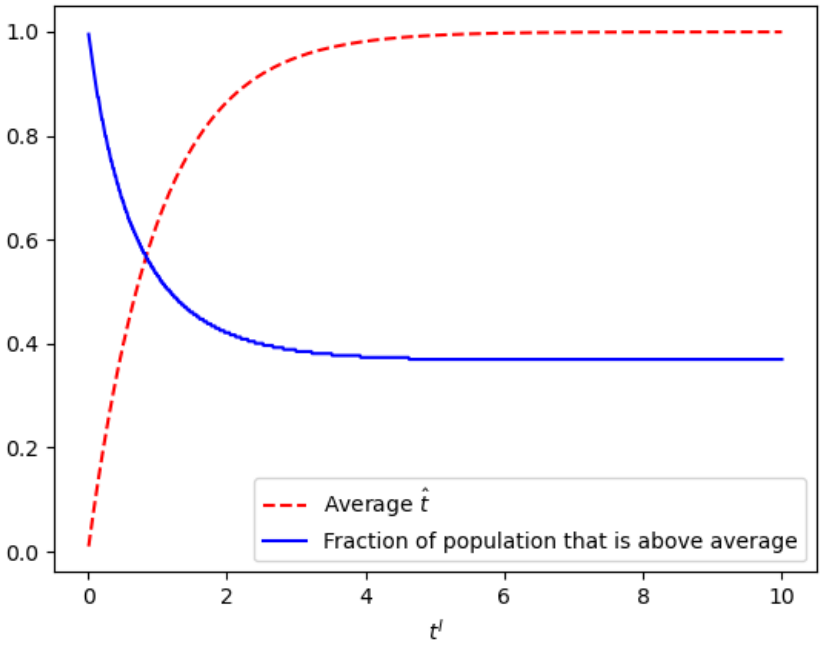
3.1.2 Comparison Results

The confidence-knowledge graph is about how people rate their knowledge in an absolute sense, or relative to the total amount of information available. The way people compare their knowledge level to that of their peers may also give the illusion of overconfidence, as seen in the next result.

Proposition 2. *For any number $p \in [0,1]$, there exists distributions $\mathcal{I}(t)$ and $\mathcal{S}(t)$ such that p fraction of the population believes they are above average.*

Proof. This can be proved by giving one simple example. Take $\mathcal{I}(t) = 1 - e^{-2t}$. Now let $\mathcal{S}(t)$ have a mass of probability $1 - p$ at point $t = 0$. Have the remaining probability distributed exponentially with an arrival rate of 1. This is $\mathcal{S}(t) = p - pe^{-t}$ for $t > 0$. Everyone stopped at $t = 0$ will know they are below average. Everyone stopped at a later point will have a two out of three chance of being stopped at $\hat{t} = t^I$. More likely than not, they are therefore above average. □

This results implies that 88 percent of drivers believing they are above average is not



irrational. Any fraction of the population may believe they are above average, even if they are rational. Take the example where $\mathcal{I}(t) = 1 - e^{-2t}$ and $\mathcal{S}(t) = 1 - e^{-t}$. Here everyone believes they are above average because there is a two thirds chance that they have all available information. This is not a mean verses median trick. A key difference is that here a large portion of the population is wrong in their guess. Also, everyone knows that everyone thinks they are above average and many of them are wrong.

The fraction of the population that is actually above average depends on the realized value of t^I. As t^I grows, this increases the length of the right tail. Thus, the average is brought up without changing anyone other than the right tail that is already above average. So the fraction of the population that is above the average goes down. The fraction of the population that is above average is approaching $\frac{1}{e}$, or about 37 percent. The fraction of the population that believes they are above average is constant at 100 percent for any realized value of t^I.

The fraction of the population that is wrong in their guess, is itself unconstrained.

Proposition 3. *For any $\epsilon > 0$, there exist distributions $\mathcal{I}(t)$, $\mathcal{S}(t)$, and a realization, t^I, such*

that the difference between the fraction that believe they are above average and the fraction that are actually above average exceeds $1-\epsilon$.

This can be done with the actual fraction exceeding the believed fraction by $1-\epsilon$, or the believed exceeding the actual by $1-\epsilon$. Both examples are shown in the appendix.

This result shows that even when we have an objective measure of knowledge/skill and people rank themselves as above or below average, we still do not have sufficient evidence to conclude irrationality regardless of how wrong they may be.

The next result gives some intuition for the source of people incorrectly believing they are above or below average. A large fraction of the population will incorrectly believe they are above average for a topic for which there is a lot of potential knowledge to learn. A large fraction of the population will incorrectly believe they are below average for topic for which there is not a lot of potential knowledge to learn. Imagine we all go try to look up papers on some new topic. If we find very few papers or none at all, we are going to believe we are below average. However, if there actually are very few papers or none at all, everyone will then believe they are below average, even though they are not.

Proposition 4. *The fraction of the population that wrongly believes they are above average is weakly increasing in the realization* t^I*, and the fraction that wrongly believes they are below average is weakly decreasing in* t^I.

Proof. Consider the realized value increasing from $\tilde{t}$ to $\tilde{t}+\epsilon$. Two things change. First, some people at the very top of the distribution will learn more as they were previously stopped only because there was no more information. Second, the average in the population will go up because the people at the top learned more. Anyone who stopped before $\tilde{t}$ will still be stopped at the same time and have the same guess of whether they are above or below average. Since the average went up, some people in the middle were above average and are now below average. If they thought they were above average, they were correct and now believe they are above average while they are actually below. If they thought they were

below average, they were incorrect but are now correct. The people at the new top level may think they are above or below average. They are actually above average. So if at $\tilde{t} + \epsilon$ they think they are above average, they are correct. This is a decrease in the fraction that incorrectly think they are below average if they think they are below average at $\tilde{t}$ and no change in the fractions in each group otherwise. If at $\tilde{t} + \epsilon$ they think they are below average, they are incorrect. This is an increase in the fraction that incorrectly believe they are below average if they think they are above average at $\tilde{t}$ and no change in the fractions in each group otherwise. At the realized value of t^I grows the fraction that incorrectly believe they are above average can only shrink or stay constant, while the fraction that incorrectly believe they are below average can only grow or stay constant. □

This result gives a possible story to explain the motivating results. Most people believing they are above average at driving can be explained by saying there is actually more to driving than people think there is. Perhaps people think along the lines of "I know how to drive. I've done it for a long time. I don't get in wrecks. I don't get tickets. I'm probably above average," while the don't realize it's possible to be much better or drive more smoothly or safely. Perhaps there is a lot more that goes into being the CEO of a company or any other field with these overconfidence results than people expect there to be. Conversely, everyone thinks they are below average at fields such as juggling. It looks hard and people think there would be a lot to learn. To the contrary, juggling is actually relatively simple to learn and many can pick it up very quickly.

4 Changing Units

In this section, I extend the main results to different types of measurement. This section is meant to show that the results will hold generally and don't depend on how knowledge or confidence is measured.

4.1 Percentiles

Here I address whether the results still hold if the variables are put into percentile terms. The short answer is that they mostly hold.

First, if we measure confidence as a function of knowledge percentile rather than amount of knowledge, the results do not change at all. In other words, we showed that $\pi(t)$ is unconstrained. I can now explain how this implies that $\pi(p)$ is also unconstrained when p is the percentile ranking of t. p is an increasing function of t. This means that rescaling the function to be in terms of p won't change the general shape of the curve. The rescaling will only contract or expand the the curve horizontally (or potentially contract some areas and expand others). Whether the curve slopes up or down or is hump-shaped or anything like that is unchanged. Recall, that there was an extra degree of freedom in choosing the distributions to match the confidence levels. Only the ratio of hazard rates is determined.

$$\sigma_I(t) = \frac{f(t)}{1 - f(t)}\sigma_S(t) \tag{5}$$

If $\sigma_I(t)$ and $\sigma_S(t)$ are both doubled, this didn't change the confidence. What it does is make those levels of t for which the hazard rates are doubled occur more frequently. So, in percentile space is is simply stretching that region of the graph horizontally. Since we're free to multiply both hazard rates by any positive values for any t, we can undo any stretching or contracting done by transforming the knowledge into percentile space in the first place.

Taking the y-variable, confidence, to be in percentile form may have a bit more bite. There are two different reasonable ways to do this. You could take everyone's confidence level and just rescale them to be percentiles (percentiles of perceived confidence). This does create some restriction on the confidence-knowledge graphs that can be obtained, but I will contend that they aren't meaningful restrictions. For example, in the baseline model, it is very easy to have a constant knowledge graph at any level. Everyone in the population can have 80 percent confidence. By definition, everyone's guess cannot be in the 80^{th} percentile of

confidence. Now if the confidence knowledge graph is flat, it means that everyone's confidence is tied for the same level. Depending on how you calculate ties, this will become a flat line at 0, 50 or 100 percentile. The problem comes from the fact that after rescaling to percentiles, apart from ties, every value is assigned only once. If there are no ties, every y value from 0 to 1, will be assigned exactly once. There can still be upward or downward slopes and weird shapes, but you can't have values that are skipped over entirely or spent too much time on. However, these restrictions do not give a way to measure irrational overconfidence, because they are simply from the definition of percentiles. When measuring the data you want to rationalize and scaling it into percentile terms, all the same restrictions still apply. So any confidence-knowledge graph that can't be rationalized by the model scaled into percentile terms, also cannot be generated by data that has been converted into percentile terms. The model is still able to rationalize any observed data measured in percentiles in this way.

A different way you could convert your y-variable to percentile is by having the agents all guess what percentile they are, then graphing that guessed percentile against the their actual level of knowledge (in percentile terms or not). Confidence is now measured as their guess of what percentile they are. In the model there was a single path of learning. This means that every agent is free from any uncertainty about what came before them. So they know perfectly how many people are at each knowledge level below them, even though they can be very uncertain about how much people above them know. But, for calculating the percentiles only what's below you matters. This means that, apart from ties, everyone can correctly guess exactly what percentile they are. So, the confidence-knowledge graph would become the 45 degree line.

Getting any other confidence-knowledge graph requires you to either be creative with how ties are handled, or relax the assumption of independence. Say that an agent is the p^{th} percentile if they know as much or more than p fraction of the population. Each agent is asked to guess their own percentile and faces a quadratic loss. The optimal guess for an

agent with knowledge t is the expected percentile.

$$\pi(t) + (1 - \pi(t))\mathcal{S}(t)$$

It has been shown already that $\pi(t)$ is unconstrained. Here the search cdf forms a lower bound on the optimal guess. However, $\mathcal{S}(t)$ can be chosen to be arbitrarily low without effecting $\pi(t)$. Thus, any confidence-knowledge graph can still be obtained.

If you aren't satisfied by a solution relying so heavily on ties or with the choice of tie-breaking rule, the baseline model will not be able to generate the desired result. What you would need is for the distribution $\mathcal{S}(t)$ to depend on the total amount of information, t^I. This is a very natural assumption. The distributions were only restricted to be independent earlier because the results didn't require any dependence and the presentation was simpler. The following example shows that even with percentile guesses, any confidence-knowledge graph is rationalizable by some joint probability function of t^S and t^I. Take the marginal distribution of t^I to be $\mathcal{I}(t) = 1 - e^{-t}$. Now define the conditional distribution of t^S for each level of t^I to be a triangle distribution over the interval $[t^I - 1, t^I]$. The most likely outcome is that you stopped learning at the tip of the triangle. The placement of the tip across this interval also gives the fraction of people that will be below the tip. So, if the tip of the triangle is at $t^I - 1 + p$, the most likely percentile rank to be is the p^{th} percentile. Therefore, any confidence-knowledge graph can be obtained by taking the marginal distribution for t^I to be any distribution with full support, and taking the conditional distribution for t^S to be a triangle distribution with the peak at the desired level of confidence.

4.2 Other Confidence Measures

As an alternative measure of confidence, let's now use the negative of the amount of available information the agent doesn't have. So now the confidence level can be expressed as,

$$\rho(t) = \mathbb{E}\left[-\left(t^I - \hat{t}\right) \mid \hat{t} = \min\{t^I, t^S\}\right]. \tag{6}$$

The agent's confidence is now not only a function of if they know everything, but on how much information they don't know. With $\pi(t)$ being the probability the agent knows all information, the problem can be simplified.

$$\begin{aligned}\rho &= \hat{t} - \mathbb{E}\left[t^I \mid \hat{t} = \min\{t^I, t^S\}\right] \\ &= \left(1 - \pi(\hat{t})\right)\left(\hat{t} - E\left[t^I \mid t^I > \hat{t}\right]\right)\end{aligned}$$

Let's choose $\mathcal{I}(t) = 1 - e^{-\lambda t}$. This choice makes the term in parenthesis equal to $-\frac{1}{\lambda}$. Now $\rho(\hat{t})$ is a linear function of $\pi(\hat{t})$, $\rho(\hat{t}) = \frac{\pi(\hat{t})-1}{\lambda}$. As shown above, by choice of $\mathcal{S}(t)$, we can get a $\pi(\hat{t})$ equal to any function we want. Thus, we now see that $\rho(\hat{t})$ is also similarly unconstrained.

4.2.1 General Confidence Measures

If you allow for joint probability distributions over t^I and t^S the result will hold for any general confidence measure. Take your confidence measure to be any functional $\psi : \mathbb{R}_+ \times \Delta(\mathbb{R}_+) \to \mathbb{R}$ mapping the amount of information you have and the posterior distribution of the amount of information you don't have into a real number. For each t, call $\mathcal{R}(t)$ the range of $\psi(t, F)$.

Proposition 5. *For any confidence measure, ψ, any function $g : \mathbb{R}_+ \to \mathcal{R}(t)$ can be rationalized by some joint probability distribution $F(t^S, t^I)$.*

This is trivially done by taking the marginal distribution of t^S to be any full support distribution. Then for each level of t^S set the conditional distribution for $t^I - t^S$ equal to

any distribution in the preimage of the desired confidence level, $\tilde{\mathcal{I}}(t^I - t^S|t^S) \in \psi^{-1}(t^S)$. The other confidence measures used in the paper are special cases of this.

5 Conclusion

5.1 Implication

This paper calls into question our ability to measure overconfidence. Unknown unknowns are virtually everywhere. In learning any subject or skill, there is an always unknown amount to be learned. We see from this paper that in the presence of unknown unknowns, overconfidence results based on confidence-knowledge graphs or conjectured comparisons to others don't work completely as imagined. While these results may reveal ex post overconfidence in the subjects, they aren't evidence of a behavioral bias or any kind of irrationality. Even a study with objective results isn't immune to this critique. Consider CEOs that hold out-of-the-money options too much. It can be shown that they lose money and ex post we see they were overconfident. However, They may view the stock price as a reflection of the market's belief of an average CEO's ability to manage the company or current projects. If the ability to run the company is learned in a world with unknown unknowns, every CEO may rationally believe they are above average. This means the company will do better than expected by the market and the stock price will go up. These CEOs turn out to be wrong, but that doesn't necessarily mean they were irrational.

Overconfidence is used to explain many phenomena, such as the no trade theorem, Milgrom and Stokey (1982). Can unknown unknowns as in this model explain information based trading? No, it cannot. While unknown unknowns lead to several facts that look like overconfidence, the model agents are still rational Bayesians. They will still work through the logic of "Why do they want to sell to me?" and "Why do they still want to sell to me knowing that I want to buy knowing that they want to sell knowing ..." and be unable to come to a trade they strictly prefer based on information alone. Overconfidence is such a

salient answer to puzzles such as this one because it is believed to be so prevalent. However, if many of our estimates of overconfidence are called into question, it's ability to solve problems loses some of that luster.

Outside the lab, these "overconfident" people see or communicate with many others like themselves. The CEO can look at what CEOs before them have done. Even if they rationally believe they are better than average, should the fact that all the other CEOs were ex post overconfident change their mind? Do these overconfidence-like results go away when agents can communicate? The answer is sometimes. Take the example where $\mathcal{I}(t) = 1 - e^{-2t}$ and $\mathcal{S}(t) = 1 - e^{-t}$. Here everyone believes they are above average. Since they are rational, if each agent shared their exact knowledge level, $\hat{t}$, the results would go away. Everyone would know if they are above or below average and whether they have all information or not. In most situations your exact level of knowledge or skill in a subject is something that is difficult to communicate. If asked, I'd have a hard time describing exactly how much economics I know. From watching past CEOs, perhaps you can infer that they all believed they were above average. That information is not sufficient to change your belief at all. From the prior distributions, you knew that everyone would think they were above average. You think that everyone is weakly below you and you don't change your mind, because people that are below are also going to think they're above average. If you tell the drivers in Sweden that 88 percent of them think they are above average, that doesn't need to change anyone's self-evaluation.

5.2 Literature Review

There are many papers that document overconfidence in a variety of settings, see Moore and Schatz (2017), Malmendier and Tate (2005), Camerer and Lovallo (1999) for a few examples. Those that most closely relate to this paper are ones that use confidence-knowledge graphs and those that use comparisons. The most well known using confidence-knowledge graphs is Kruger and Dunning (1999). For a review of related results and explanations, see Dunning

(2011). The most accepted explanation of the Dunning-Kruger effect is a "better-than-average" effect and some randomness in the test, but it is still a behavioral bias in such explanation. A few papers finding a better-than-average effect are Svenson (1981) and Buunk and VanYperen (1991). Kruger (1999) even finds a worse than average effect for some subjects.

There are several rational explanations of some overconfidence results. Benoît and Dubra (2011), Köszegi (2006), and Moore and Healy (2008) all give rationalizations of more than 50 percent of the population believing they are above average, but but none of these results apply as widely to different measures of overconfidence. There are also strategic and evolutionary explanations of overconfidence, see Zábojník (2004) and Johnson and Fowler (2011). The current paper is not attempting to provide a strategic reason for people to be overconfident, but rather to state a statistical property which my be confused with overconfidence in observations.

My talk of unknown unknowns sounds similar in motivation to unawareness papers, such as Modica and Rustichini (1999). Methodologically, there is no similarity. My agents are fully rational Bayesians. The model is more a statistical model of the distribution of information.

5.3 Future Work

There are many avenues of future work in unknown unknowns that seem promising. It could potentially be interesting to endogenize the search distribution. That was not done in this paper because I wanted to emphasize the statistical properties leading to apparent overconfidence, but as a separate project studying how individuals choose to acquire information in the presence of unknown unknowns may yield interesting results. In addition to the known and unknown unknowns, it may be interesting to add known and unknown knowns. You don't always know what you know.

References

Benoît, J.-P. and Dubra, J. (2011). Apparent overconfidence. *Econometrica*, 79(5):1591–1625.

Buunk, B. P. and VanYperen, N. W. (1991). Referential comparisons, relational comparisons, and exchange orientation: their relation to marital satisfaction. *Personality and Social Psychology Bulletin*, 17(6):709–717.

Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: an experimental approach. *American Economic Review*, 89(1):306–318.

Dunning, D. (2011). The dunning-kruger effect: On being ignorant of one's own ignorance. In Zanna, M. and Olson, J., editors, *Advances in Experimental Social Psychology*, volume 44, chapter 5, pages 247–296. Academic Press.

Johnson, D. and Fowler, J. (2011). The evolution of overconfidence. *Nature*, 477:317–320.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Kruger, J. (1999). Lake wobegon be gone! the "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, 77(2):221–232.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134.

Malmendier, U. and Tate, G. (2005). Ceo overconfidence and corporate investment. *The Journal of Finance*, 60(6):2661–2700.

Milgrom, P. and Stokey, N. (1982). Information, trade, and common knowledge. *Journal of Economic Theory*, 26(1):17–27.

Modica, S. and Rustichini, A. (1999). Unawareness and partitional information structures. *Games and Economic Behavior*, 27(2):265–298.

Moore, D. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502–517.

Moore, D. and Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, 11(8).

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, 47(2):143–148.

Zábojník, J. (2004). A model of rational bias in self-assessments. *Economic Theory*, 23(2):259–282.

6 Appendix

Budget:

Proof. Suppose here that we added a third stopping time to the model to represent the budget. The difference between this time, $t^B \sim \mathcal{B}(t)$, is that it is known. You know your value of t^B from the beginning. So, $\mathcal{B}(t)$ represents the distribution of budgets across the population, rather than a prior belief. Let's compute your confidence again. If you are stopped by your budget, your confidence level will be zero. As long as there aren't atoms in the distribution, there is zero probability that the total amount of information coincides perfectly with your budget. If you get stopped and you aren't at your budget constraint, your confidence level is the same as in the main model. You know you were stopped by either t^I or t^S.

Now we can find the average confidence for each time t by weighting those with the fraction of the population that stops for each reason. This gives the confidence-knowledge graph.

$$\begin{aligned}\pi(t) &= 0\ \frac{\sigma_B(t)}{\sigma_B(t)+\sigma_S(t)+\sigma_I(t)} + \frac{\sigma_I(t)}{\sigma_S(t)+\sigma_I(t)}\frac{\sigma_S(t)+\sigma_I(t)}{\sigma_B(t)+\sigma_S(t)+\sigma_I(t)} \\ &= \frac{\sigma_I(t)}{\sigma_B(t)+\sigma_S(t)+\sigma_I(t)}\end{aligned}$$

You'll notice that only the sum of the hazard rates for the budget and search enter the confidence formula. You can define a new distribution $\hat{\mathcal{S}}$ with a hazard rate equal to that sum, $\sigma_{\hat{s}}(t) = \sigma_B(t) + \sigma_S(t)$, and the confidence formula is unchanged from the main model in the text. □

Proposition 1:

Proof. We can take the distributions as defined in the text.

$$\mathcal{S}(t) = 1 - e^{-t} \quad \text{and} \quad \mathcal{I}(t) = 1 - e^{-\int_0^t \frac{f(\tau)}{1-f(\tau)}d\tau} \tag{7}$$

This works in most cases. I will just clarify and add a few points.

First, it seems in the text like I was presupposing the distributions $\mathcal{I}(t)$ and $\mathcal{S}(t)$ to have a pdf. In fact, we are choosing $\mathcal{I}(t)$ and $\mathcal{S}(t)$ to rationalize a function $f(t)$, and as shown, we can always choose distributions that have a pdf.

The boundary cases do create a bit of a complication. There is no complication with the distribution definitions when $f(t) = 0$. In this case, the hazard rate of $\mathcal{I}(t)$ becomes zero while the hazard rate of $\mathcal{S}(t)$ remains positive. This gives the desired confidence level. For any t such that $f(t) = 1$, we need to redefine the distributions. Define

$$\lambda_I(t) = \begin{cases} \frac{f(\tau)}{1-f(\tau)} & \text{if } f(t) \neq 1 \\ 1 & \text{if } f(t) = 1 \end{cases} \quad \text{and} \quad \lambda_S(t) = \begin{cases} 1 & \text{if } f(t) \neq 1 \\ 0 & \text{if } f(t) = 1 \end{cases}. \tag{8}$$

Now the functions

$$\mathcal{S}(t) = 1 - e^{-\int_0^t \lambda_S(t)d\tau} \quad \text{and} \quad \mathcal{I}(t) = 1 - e^{-\int_0^t \lambda_I(t)d\tau} \tag{9}$$

give the desired confidence levels for all values of t. This is the same as the original distribution definitions, except when $f(t) = 1$ we need $\mathcal{S}(t)$ to have a hazard rate of zero and $\mathcal{I}(t)$ to have a positive hazard rate.

On further note to verify is that the distribution functions indeed approach one. This is equivalent to the hazard rates integrating to infinity. We don't actually need both distributions to have this property, only the distribution describing the minimum of the two stopping times. It's reasonable that there is positive probability of an unlimited amount of information being available as long as the search will still always stop in finite time. It's also reasonable that the search doesn't fail until acquiring all information. While not necessary, it is sufficient that the hazard rate of at least one of the distributions integrates to infinity as is the case in my definition. Typically, the hazard rate of $\mathcal{S}(t)$ is constant at one. So it clearly integrates to infinity. For it to not integrate to infinity, $f(t)$ would need to be equal

to one for an increasing frequency as t grows. However, in that case, $\mathcal{I}(t)$ has a constant hazard rate of one and would thus integrate to infinity. So, the distribution function of the minimum of the two stopping times as described will always approach one in the limit.

Note that a change in the desired confidence level on a measure zero set would change the distributions as described here. As the definition of rationalizability only requires the posterior to be almost everywhere equal, no problem arises. □

Proposition 3:

Proof. Fix $\epsilon > 0$. Take the distribution $\mathcal{S}(t)$ to have mass at three points.

$$t^S = \begin{cases} 0 & \text{with probability } \frac{\epsilon}{4} \\ 1 & \text{with probability } 1 - \frac{\epsilon}{2} \\ 3 & \text{with probability } \frac{\epsilon}{4} \end{cases}$$

Take the distribution $\mathcal{I}(t)$ to have mass at two points.

$$t^I = \begin{cases} 1 & \text{with probability } q \\ 3 & \text{with probability } 1 - q \end{cases}$$

Everyone stopped at zero will know they are below average and everyone stopped at three will know they are above average. Most people will be stopped at one. If $t^I = 1$, these people will all be above average. If $t^I = 3$, they will all be below average. Of course, they don't know if t^I is 1 or 3. So if q is high, they will guess they are above average, and if q is low they will guess they are below average. The cutoff is

$$q^* = \frac{4 - 2\epsilon}{8 - 3\epsilon},$$

which is just a hair under 50 percent.

Almost half the time, nearly the entire population will be wrong in their self-assessment. If $q > q^*$ and $t^I = 3$, $1 - \frac{\epsilon}{4}$ fraction of the population will think they are above average while only $\frac{\epsilon}{4}$ actually are. This is a difference of $1 - \frac{\epsilon}{2}$. If $q < q^*$ and $t^I = 1$, no one will think they are above average while $1 - \frac{\epsilon}{4}$ actually are. This is a difference of $1 - \frac{\epsilon}{4}$. □