COLLECTIVE WISDOM

AIDAN LYON[1,2,3]

1: Philosophy Department, University of Maryland, College Park. mailto:alyon@umd.edu alyon@umd.edu

2: Munich Centre for Mathematical Philosophy, Ludwig Maximilian University of Munich.

3: Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne.

* * * Draft Manuscript. Please do not quote or cite this version. * * *

I INTRODUCTION

When the Oracle of Delphi said that Socrates was the wisest of all, the philosopher asked himself:

> "Whatever does the god mean? What is his riddle? I am very conscious that I am not wise at all; what then does he mean by saying that I am the wisest?" *Apologia*[1] 21b

and he began an investigation to find the answer. His answer (as I shall argue) was that wisdom consists of knowing fine things and also knowing the limits of one's knowledge. Although Socrates apparently did not know many fine things, he also did not think he knew these things. Others, in contrast, often thought they knew when then they did not. And because of this, Socrates was the wisest individual of all—as the Oracle had proclaimed.

In more recent times, many authors have said that *collectives* of individuals can be wise. Indeed, these authors say that collectives can even be *wiser* than the individuals of which they are composed.[2] These authors are inspired by examples of collectives accomplishing tasks that typi-

[1] Translation by George Grube and John Cooper (2000). *The Trial and Death of Socrates: Euthyphro, Apology, Crito, Death Scene from Phaedo*. Hackett Publishing.

[2] See *e.g.,* James Surowiecki (2004). *The Wisdom of Crowds*. Doubleday; Scott Page (2008). *The Difference*. Princeton University Press; and many of the articles in Hélène Landemore and Jon Elster (2012). *Collective Wisdom: Principes and Mechanisms*.

cally challenge individuals. For instance, there is evidence that collectives can diagnose extremely rare illnesses, even though individual medical experts have failed to do so.[3] Collectives can also make very accurate estimates of quantities such as the weight of an ox and the number of jelly beans in a jar.[4]

As exciting as these examples might be, it would seem that they are not really instances of wisdom. Rather, they appear to be merely examples of collectives making accurate judgments—albeit perhaps *surprisingly* accurate judgments.[5] Although there is much to be said for this assessment, I shall argue that collectives can in fact be wiser than their individuals. One may wonder how this could be. In short, the answer comes from Socrates and from psychological research on human overconfidence.[6] Individuals typically overestimate how much they know, but, as collectives, they can avoid this error. This entails that collectives can be wiser than their individuals, in the Socratic sense of wisdom outlined above.

I shall begin by clarifying the focus of my discussion. This will involve critically examining a theorem known as the *Diversity Prediction Theorem*, which Page 2008 has claimed provides a "logic for the wisdom of crowds" (p. 209). Although it falls short of such a grand claim, Page's work nevertheless provides a useful framework for understanding collective wisdom.

[3] See *e.g.,* Rachel Nuwer (2013). "Software could make rare diseases easier to spot". In: *New Scientist* 218.2913, p. 21.

[4] See Francis Galton (1907b). "Vox Populi". In: *Nature* 75, pp. 450–1 and Surowiecki (2004).

[5] Reasoning more extensively along similar lines, Daniel Andler (2012). "What has Collective Wisdom to do with Wisdom?" In: *Collective Wisdom*. Ed. by Cambridge University Press, pp. 72–84 concludes that "the wisdom of crowds has not much to do with wisdom" (p. 94).

[6] See *e.g.* Sarah Lichtenstein, Baruch Fishhoff, and Lawrence D. Phillips (1982). "Calibration of Probabilities: The State of the Art to 1980". In: *Judgment under Uncertainty: Heuristics and Biases*. University of Cambridge; and Peter Juslin, Anders Winman, and Henrik Olsson (2000). "Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect." In: *Psychological Review* 107.2, p. 384.

II A FRAMEWORK FOR COLLECTIVE WISDOM

In this paper, I will be focused on situations that involve individuals producing numerical judgements that are aggregated into a collective numerical judgement of the same kind. Canonical examples include estimations of the number of jelly beans in a jar, interest rate predictions, and probabilistic assessments and forecasts. Moreover, I shall restrict my focus to a very simple method of aggregation: collective judgements will be defined as the unweighted linear average of the individual judgements.[7] This means that deliberation groups, economies, markets, elections, voting procedures, etc. all fall outside the purview of my discussion.

A classical example that falls within my focus was reported by Francis Galton.[8] Galton noted an interesting result of a contest of some 800 people at a county fair in Plymouth, 1906. The contest was to produce the most accurate estimate of the weight of an ox, slaughtered and dressed. Galton found that the average of the crowd's estimates was surprisingly close to the true weight of the ox, despite huge errors in many of the individual guesses.[9]

Although the average was extremely accurate, there were estimates—which Galton suspected were from experts—that were even more accurate. This raises the next issue of clarification: how well does the collective need to perform to exhibit "collective wisdom"? One answer that

[7] This is not to imply that other aggregation methods are not important or not worthy of study. It is simply that unweighted linear averages (sometimes known as "democratic opinion pools") are enough to raise interesting philosophical issues. Moreover, unweighted linear averages are often considered a good baseline by which to compare other methods—see e.g., J Armstrong (2001). "Combining forecasts". In: *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Ed. by J. Scott Armstrong. Norwell, MA: Kluwer Academic Publishers. However, they are not free of controversy—for a survey of their issues, see Christian Genest, James V Zidek, et al. (1986). "Combining probability distributions: A critique and an annotated bibliography". In: *Statistical Science* 1.1, pp. 114–135.

[8] Francis Galton (1907a). "Letters to the Editor: The Ballot-Box". In: *Nature* 75, pp. 900–1.

[9] In Galton 1907b he aggregated the estimates using the median, and in Galton 1907a he reported the mean to be more accurate than the median.

seems plausible, and which often appears in the literature, is that the performance of the collective must be at least as good as the average performance of the individuals. For example, Hong and Page[10] *define* collective wisdom as occurring when this condition is satisfied. Hong and Page therefore understand this condition as both a necessary and sufficient one. However, it is surely not a sufficient condition. If the vast majority of Galton's collective had estimated the weight of the ox to be around ten thousand tons, then the collective estimate would also be close to that clearly erroneous figure, and we would not think that a collective wisdom had been exhibited. There must, therefore, be a constraint on the collective's performance in absolute terms—although this may vary with context. Therefore, I take Hong and Page's definition only as a *necessary* condition for collective wisdom:

CW1 The performance of the collective is better than the average individual performance.

To analyse the "wisdom of crowds", Page presents what he calls the *Diversity Prediction Theorem* (DPT).[11] Consider n individuals, $i = 1, ..., n$. Each individual makes a judgement, j_i, about the true value of some real-valued quantity whose true value is τ. The accuracy of a judgement, j, is measured in terms of its *squared error*: $SqE(j) = (j - \tau)^2$. The collective judgement, c, is defined as the average of the individual judgements: $c = 1/n \sum_{i=1}^{n} j_i$. The *prediction diversity* of the collective, d, is defined as the variance of the collective's judgements: $d = 1/n \sum_{i=1}^{n} (j_i - c)^2$. With these terms defined, we can state the theorem: the squared error of the collective judgement

[10] Lu Hong and Scott Page (2012). "The Micro-Foundations of Collective Wisdom". In: *Collective Wisdom: Principles and Mechanisms*.

[11] Although I'm focusing on Page because his work has had a large impact on the "wisdom of crowds" literature, the theorem was proven earlier by Anders Krogh and Jesper Vedelsby (1995). "Neural network ensembles, cross validation, and active learning". In: *Advances in Neural Information Processing Systems*, pp. 231–238.

is the average squared error of the individual judements minus the collective's prediction diversity:

$$SqE(c) = \frac{1}{n}\sum_{i=1}^{n} SqE(j_i) - d$$

Note that there is nothing probabilistic about the DPT, and there is no assumption of any kind of independence between the j_i.

This seems like a striking result: no matter what the judgements are, no matter who produced them or how they produced them, and no matter what the quantity in question is, so long as there is some diversity to the judgements—i.e., so long as $d > 0$—the collective judgement will be more accurate than the average individual judgement. Indeed, Page calls this the *"Crowd Beats Averages Law"* (p. 209), and says that the DPT provides a "logic for the wisdom of crowds" (*ibid.*). Put in terms of an explanation, it would appear that we have a mathematical explanation of a large class of collective wisdom phenomena: effects like the one Galton observed at Plymouth happen because they are *mathematically guaranteed*.

However, there is a problem. The problem can be seen by noting that a particular error function was chosen as the measure of accuracy: the squared error function. Since the squared error function is convex, it follows from *Jensen's inequality*[12] that the error of the collective judgement is less than or equal to the average error of the individual judgements. However, if we had chosen to measure accuracy using a concave function, then the opposite follows: the error of the collective is *greater* than or equal to the average error of the individual judgements.

So the decision to measure accuracy by the squared error function was not an innocent one, and the problem is that the DPT does not hold for many other error functions. For example, Galton 1907b reports accuracies in terms of absolute errors and absolute percentage errors. According to the absolute error function, $AbsE(j) = |j - \tau|$, the error of the collective can be equal to the

[12] Johan Ludwig William Valdemar Jensen (1906). "Sur les fonctions convexes et les inégalités entre les valeurs moyennes". In: *Acta Mathematica* 30.1, pp. 175–193.

average error *even if d is positive*. This happens if the judgements j_i are either all greater than τ or all less than τ. Moreover, if either of these conditions obtain, and if accuracy is measured by a strictly concave error function, then the collective error will be *greater* than the average error, thus producing a kind of *anti*-collective wisdom effect.

Given that CW1 does not specify a particular way in which accuracy must be measured (and rightly so), it follows that any *general* result about when collective wisdom effects obtain should at least be robust with respect to a plausible range of performance measures. This does not mean that formal results concerning particular measures of performance are of little value. It is just that we should be careful when making general conclusions from particular results.

One might object that the squared error function is the one true measure of accuracy. However, our notion of accuracy is just an abstraction of the loss functions that we tend to employ in our decision making, and the loss functions that we use depend on what we happen to care about. For example, if we are manufacturing widgets to very precise specifications, then one widget that is far off may be considered almost just as bad as another widget that is closer to, but also far from, the specifications—i.e., they are almost equally useless. We may be able to make some general claims about measures of accuracy, but they will tend to be fairly weak—e.g., that they are non-decreasing functions[13]—and so it is unlikely that we can single out one function as *the* measure of accuracy.

Given that the set of plausible accuracy measures includes non-convex functions, it follows that a necessary condition for the accuracy (generally understood) of a collective judgement being

[13] Even this may not be true, for as L. J. Savage (1954). *The Foundations of Statistics*. New York: Wiley noted, there are examples of non-monotonic loss functions: “William Tell, for example, in estimating the angle by which to elevate his crossbow for the apple shot might have preferred a downward angle of 10 degrees to one of 1 degree; but such circumstances seem exceptional.” (pp. 230-231). Indeed, such cases do seem exceptional, and so a conceptual analysis of accuracy might appropriately disregard them.

better than the average accuracy (generally understood) of the individual judgements is that some of the individual judgements be too low and some be too high.

CW2 For the accuracy of a collective judgement to be better than the average accuracy of the individual judgements, some of the judgements must be below the true value and some must be above the true value.

This condition is sometimes called *bracketing*, because the judgements "bracket" the true value of the quantity in question.[14] If the individual judgements bracket the true value, then, when they are averaged, there is some cancellation of their errors. For linear error functions, this cancellation will be enough for the collective to be more accurate than the average individual. For concave error functions, the amount of cancellation *might* be enough, but it might not. So, to say that bracketing is a necessary condition for collective wisdom understood in terms of general accuracy is to say that there must be some cancellation of errors in the process of averaging.

To summarise. In this section, I have clarified and restricted my focus of discussion concerning certain key concepts, and the rest of the paper should be understood with these clarifications and restrictions in mind. I have also identified one necessary condition for collective wisdom understood in the most general terms: CW1. And for collective wisdom understood in terms of general accuracy, I have identified CW2 as a necessary condition. As I mentioned earlier, there may be other necessary conditions for collective wisdom. However, I shall leave this line of thought here, because it turns out that the above two conditions are enough to result in some interesting consequences for collective credences, which is what I will argue for in the next section.

[14] Jack B Soll and Richard P Larrick (2009). "Strategies for revising judgment: How (and how well) people use others' opinions." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.3, p. 780.

III COLLECTIVE CREDENCES

Consider a collective of individuals that assign credences to some arbitrary proposition. A standard way of measuring the accuracy of these credences is by the *Brier score*. The Brier score of a credal judgement j is: $Br(j) = (j - \tau)^2$, where τ is the truth value of the proposition in question. Even though we are now in the context of credences and truth values, the DPT still applies and so it follows that the Brier score of a collective credence is less than the average Brier score of the individual credences, if there is some difference between them. This looks like good news: collective wisdom effects are guaranteed for diverse credal judgements.

However, the Brier score is the probabilistic version of the squared error function, and so it, too, is a convex function, and so the points I made in the previous section apply here as well as they did before. If we measure the accuracy of a credence by the absolute error in probability, $|j - \tau|$, then, as before, we are not guaranteed a collective wisdom effect. And if we choose a concave score, then it is even possible for a collective credence to be less accurate than the average individual credence.

Therefore, as I argued in the previous section, in order for there to be a collective wisdom effect in terms of accuracy, the judgements must bracket the truth value of the proposition in question. That is, the judgements must be such that there is some cancellation of their errors when they are averaged. And here lies the problem: *it is impossible for credences to bracket truth values*. Truth values are always either 0 or 1, and coherent credences always lie between 0 and 1 (inclusively).[15]

[15] Assuming the credences are coherent—that is, they satisfy some reasonably standard set of probability axioms (e.g., Kolmogorov A. N. Kolmogorov (1933). *Foundations of Probability*. (1950) Chelsea Publishing Company, Popper Karl R. Popper (1938). “A Set of Independent Axioms for Probability”. In: *Mind* 47.186, pp. 275–277, Rényi Alfred Rényi (1955). “On a New Axiomatic Theory of Probability”. In: *Acta Mathematica Academiae Scientiarum Hungaricae* 6, pp. 286–335; Alfred Rényi (1970). *Foundations of Probability*. Holden-Day, Inc.).

So rather than having good news, it looks like we have bad news: far from being guaranteed, collective wisdom effects are *impossible* for credences.

One might object that, unlike the squared error of a quantity estimate, there really is something special about the Brier score. Patrick Maher[16] and Paul Horwich[17] have argued for the absolute error measure as the best measure of accuracy for credences, and in response, Jim Joyce[18] has argued that the Brier score has a number of desirable properties as a measure of "epistemic utility" (which Joyce seems to identify with accuracy). However, Joyce admits that the Brier score may not be the "right" measure in every situation (p. 293).Until an argument is given to show that, definitely, the Brier score is the one true measure of credal accuracy, it is best to not rely on the peculiarities of any particular accuracy measure and to make points that are robust against plausible candidates.

The point about coherent credences not being able to bracket truth values is a simple one, but it is not an idle philosophical curiosity. In psychology, there has recently been a surge of interest in the so-called *crowd within* phenomenon. Roughly speaking, by asking someone to make a second estimate about some quantity for which they have already made an initial estimate and by taking the average of their two estimates, one obtains a "collective" estimate that is more accurate than both of the individual's estimates.[19] Even more roughly speaking: it appears that each individual has a "crowd" within them. Herzog and Hertwig[20] have developed a judgement elicitation technique

[16] Patrick Maher (2002). "Joyce's argument for probabilism". In: *Philosophy of Science* 69.1, pp. 73–81.

[17] Paul Horwich (1982). *Probability and Evidence*. Cambridge University Press.

[18] James M Joyce (2009). "Accuracy and coherence: Prospects for an alethic epistemology of partial belief". In: *Degrees of Belief*. Springer, pp. 263–297.

[19] Edward Vul and Harold Pashler (2008). "Measuring the crowd within probabilistic representations within individuals". In: *Psychological Science* 19.7, pp. 645–647

[20] Stefan M Herzog and Ralph Hertwig (2009). "The wisdom of many in one mind improving individual judgments with dialectical bootstrapping". In: *Psychological Science* 20.2, pp. 231–237

nique called *dialectical bootstrapping*, by which an individual is prompted to generate their second judgement by considering an evidence set that is different from the one they used to generate their first judgement. Herzog and Hertwig have found that by averaging the two judgements elicited using this method, one obtains a "collective" judgement that tends to have a lower absolute error than the initial judgement made, even if the absolute error of the second judgement is greater than the first. However, Herzog and Hertwig only examine judgements that are quantity estimations (*cf.* section II). One might think that similar effects could be observed for credal judgements; however, because credences cannot bracket truth values, such effects are mathematically impossible. The average of two credal judgements will only have a lower absolute error than the first judgement if the second judgement has a lower absolute error than the first, and the average's absolute error will always be equal to the average absolute error of the two judgements. So dialectical bootstrapping will nit have the same effect on credences as it has on quantity estimates.

Since it is impossible for credences to bracket truth values, it is impossible for there to be any cancellation of errors and so it is impossible for there to be genuine collective wisdom effects for credences. It is important to note, though, that this conclusion is true only given that we have chosen to measure the performance of individual and collective credences in terms of *accuracy* (and the other restrictions I made in section II). There are, however, other measures of credal performance.

Another measure of credal performance is *calibration*. The basic idea behind calibration is that if you assign, say, 0.9 probability to a large number of propositions—that you consider to be independent[21]—then 90% of those propositions should be true, and you are *miscalibrated* otherwise. This basic idea can be generalised as follows. Let x_i be a type of credal judgement (e.g., 0.9),

[21] In everything that follows, I always assume that the propositions in question are considered to be probabilistically independent.

n_i be the number of judgements you make of that type, f_i the fraction of propositions that receive x_i credence that are true, and let N be the total number of judgements. Then your calibration is measured by: $\mathrm{Cal} = \frac{1}{N}\sum_{i=1}^{K} n_i|x_i - f_i|$. If $\mathrm{Cal} > 0$, then you are miscalibrated and the larger the value of Cal, the more miscalibrated you are.[22]

The calibration of a collective can be better than the average calibration of the individuals. For example, consider the following example:

Example 1

Truth values	1	0	1	0	1	0	1	0	1	0
Ann’s credences	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Bob’s credences	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Their collective credences	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Because 50% of the propositions are true, Ann and Bob are both miscalibrated by 0.3. But, as a collective, they are perfectly calibrated, since the average of their credences is always 0.5.

It is possible to for the collective calibration to be worse than the average individual calibration; indeed, it can be worse than *all* of the individual calibration scores. Some work has been made in identifying conditions under which the collective calibration is better than average individual calibration, but it is difficult to find conditions that are general and realistic.[23] At any rate, it is clear that the positive effect is not limited to contrived examples, such as Example 1, and it can be observed empirically. In Appendix A, I report the result of an experiment which elicited credal

[22] Sometimes calibration is measured in terms of the squared difference between x_i and f_i. I am focusing on the measure based on absolute differences because of my earlier points concerning convex measures. If there can be genuine collective wisdom effects for Cal, then it follows there can be genuine collective wisdom effects for calibration measures based on squared differences too. The converse is not true, hence my focus on absolute errors.

[23] For discussion and further references, see Dan Ariely et al. (2000). “The effects of averaging subjective probability estimates between and within judges.” In: *Journal of Experimental Psychology: Applied* 6.2, p. 130. (Note that Ariely *et al.* use a calibration score based on squared differences, not absolute differences.)

judgments using Amazon's Mechanical Turk.[24] The calibration of the collective was substantially better than the average individual calibration. On average, participants were miscalibrated by 0.25 (95% CI = [0.23, 0.26]), and the collective was miscalibrated by 0.12. Out of 135 individuals, only 2 were better calibrated than the collective.

The rest of the paper proceeds as follows. In the next section I shall give an interpretation of Socrates' thoughts on wisdom. In section V, I will show that Socrates' notion of wisdom can be generalised by understanding it in terms of calibration. Since the calibration of a collective can be better than the average individual, we will have the result that the wisdom of a collective can be greater than the average wisdom of the individual in the collective.[25]

IV SOCRATES' WISDOM

Not knowing what the Oracle meant when she said that he was the wisest of all, Socrates proceeded to interrogate the local Athenians who were renowned for their wisdom. Socrates said he hoped he could find someone wiser than he, so that he could present him[26] as a counterexample to the Oracle. However, from each interrogation Socrates discovered why he was wiser than his interrogatee.

Socrates first interrogated a politician, who had a reputation for being wise, and concluded:

> I thought that he appeared wise to many people and especially to himself, but he was not. I then tried to show him that he thought himself wise, but that he was not. As a result he came to dislike me, and so did many of the bystanders. So I withdrew and thought to myself: "I am wiser than this man; it is likely that neither of us knows anything worthwhile, but he thinks he knows something when he does not, whereas

[24] See Gabriele Paolacci, Jesse Chandler, and Panagiotis Ipeirotis (2010). "Running experiments on Amazon Mechanical Turk". In: *Judgment and Decision Making* 5.5, pp. 411–419 for how experiments are conducted on this service.

[25] Note that this conclusion does not depend on my point earlier about the Brier score not being the privileged accuracy measure. One can insist that the Brier score is in fact the one true measure of accuracy for credences but also accept that calibration is another important virtue that an individual's credences may have.

[26] In ancient Athens, only men were considered to be candidates for having a high degree of wisdom.

> when I do not know, neither do I think I know; so I am likely to be wiser than he to this small extent, that I do not think I know what I do not know." 21c–d.

After offending a few more politicians, Socrates moved on to the poets:

> "After the politicians, I went to the poets, the writers of tragedies and dithyrambs and the others, intending in their case to catch myself being more ignorant than they. [...] I soon realized that poets do not compose their poems with knowledge, but by some inborn talent and by inspiration, like seers and prophets who also say many fine things without any understanding of what they say. [...] At the same time I saw that, because of their poetry, they thought themselves very wise men in other respects, which they were not. So there again I withdrew, thinking that I had the same advantage over them as I had over the politicians." 22a–c.

Socrates then proceeded to interrogate the artisans, expecting them to have knowledge of many fine things and thus have much wisdom:

> "In this I was not mistaken; they knew things I did not know, and to that extent they were wiser than I. But, men of Athens, the good craftsmen seemed to me to have the same fault as the poets: each of them, because of his success at his craft, thought himself very wise in other most important pursuits, and this error of theirs overshadowed the wisdom they had, so that I asked myself, on behalf of the oracle, whether I should prefer to be as I am, with neither their wisdom nor their ignorance, or to have both. The answer I gave myself and the oracle was that it was to my advantage to be as I am." 22d–e.

These passages, I believe, contain the essential elements to Socrates' thoughts on wisdom.

Contemporary philosophers have given a number of interpretations of these passages. These interpretations include: S is wise iff S believes s/he is not wise; S is wise iff S believes S does not know anything; S is wise iff for all p, S believes S knows p iff S knows p; and S is wise iff for all p, S believes S knows p iff S's belief in p is highly justified. I will not delve into the details of these interpretations[27] because they are all defective in one important respect: they all treat wisdom as a simple binary property. However, in the above passages we find comparisons of wisdom ("I am

[27] See e.g., Sharon Ryan (2013). "Wisdom". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2013. `http://plato.stanford.edu/archives/sum2013/entries/wisdom/` for a discussion of them.

wiser than this man"), degrees of wisdom ("I am likely to be wiser then he to this *small extent*"), and perhaps even aggregations of wisdom ("this error of theirs *overshadowed* the wisdom they had"). So the above interpretations are textually inaccurate in these respects.

The following is a much more plausible interpretation of the above passages, and it is also a more plausible (but partial) account of our concept of wisdom:

D1 *Ceteris paribus*, A is wiser than B if A knows more about topics Γ than B.

D2 *Ceteris paribus*, A is wiser than B if the difference between what A thinks h/she[28] knows and what A really knows is smaller than the difference between what B thinks h/she knows and what B really knows.

The interpretation has the following properties that require some explanation: (i) it has two dimensions, D1 and D2, (ii) it leaves Γ unspecified, (iii) it has *ceteris paribus* clauses, and (iv) it refers to the difference between what one thinks one knows and what one genuinely knows.

D1 captures this component to the passages: *ceteris paribus*, if you have knowledge about certain topics, then you have some wisdom. Socrates' reason for going to the poets and the artisans was their reputation for having a special kind of knowledge. Socrates concluded that the poets did not deserve this reputation but that the artisans did: "they knew things I did not know, and to that extent they were wiser than I." Socrates does not state an exhaustive list of the topics conducive to wisdom. The poetic process and various crafts apparently fall into this list. And presumably there may be other such topics—e.g., science and morality. Let Γ include all, and only, those topics.

D2 captures what I take to be the main point of the above passages: wisdom has to do with

[28] Although Socrates was only concerned with making comparisons between men, I see no reason to think that this was essential to his concept of wisdom. Incidentally, if the main conclusion of this paper is true, then there is empirical evidence that women tend to be wiser than men. See J.B. Soll and J. Klayman (2004). "Overconfidence in interval estimates." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.2, p. 299.

knowing the limits of your knowledge. However, there is a discrepancy between D2 and the above passages. Socrates seems to have been concerned only with people *over*-estimating their knowledge, but D2 doesn't discriminate between over- and under-estimations of knowledge. For my purposes, the discrepancy doesn't matter and I could replace D2 with:

D2' *Ceteris paribus*, *A* is wiser than *B* if *A* overestimates what h/she knows less than the extent to which *B* overestimates what h/she knows.

Indeed, this modification actually makes it easier to argue for my main conclusion (see section VI). Nevertheless, D2 seems true of the concept of wisdom and in the spirit of the above passages, so I'll proceed with the presumption that it is true.

I have deliberately left D1 and D2 independent of each other. Socrates does seem to think that comparisons can be made between the rankings along these dimensions, but he does not tell us how such comparisons are to be made, e.g.: "each of them, because of his success at his craft, thought himself very wise in other most important pursuits, and this error of theirs overshadowed the wisdom they had". I suspect that Socrates was of the opinion that *any* overestimation of one's knowledge "overshadows" the rest of one's wisdom. However, he does not say this, and he says little about this issue of aggregation. Thus, keeping D1 and D2 independent with the *ceteris paribus* clauses allows me to give a minimal interpretation that does not commit to any details about how the different components of wisdom are to be aggregated. The *ceteris paribus* clauses also allow me to bracket the possibility that there are still further aspects to wisdom.[29]

I have just explained points (i), (ii), and (iii). Now for (iv): First, there is a subtlety concerning the original text and its translation. In the above passages, it seems that Socrates is wiser than his interrogatees, because, in contrast to them, he does not *think* he knows what he does not ac-

[29] *E.g.,* Sharon Ryan (1999). "What is wisdom?" In: *Philosophical Studies* 93.2, pp. 119–139 discusses the possibility that *actually* living well (as opposed to merely knowing how to live well) is important for wisdom.

tually know. However, it is not entirely clear that mere thinking is enough. Plato sometimes uses the word '*oiomai*', which roughly translates to English as 'think' or 'suppose', but it also has the connotation of being at the forefront of one's mind or a judgement that one has arrived at consciously through reason.[30] This suggests that it may be important that when he does not actually know some proposition, Socrates explicitly believes that he does not know it, rather than that he (merely) implicitly believes that he does not know it, or that he simply refrains from believing that he knows the proposition. To put the issue in terms of an analysis of wisdom: is it important to one's wisdom that one is explicitly aware of the limitations of one's knowledge, or is it enough to simply refrain from making the mistake of believing that one knows what one does not? I suspect that a good answer to this question is that not making the mistake contributes to your wisdom, and being aware of the limitation makes an *additional* contribution to your wisdom. However, this goes beyond an exegesis of Socrates, and the original text does not seem to determine a unique resolution to the exegetical issue. If one insists that the "explicit belief" interpretation is correct and crucial to Socrates understanding of wisdom, then one should read all instances of 'Socratic wisdom' in this paper as 'Socratic-like wisdom'.[31]

Next, there is the issue of how to measure the difference between what one thinks one knows and what one actually knows. One obstacle to this is that defining knowledge is notoriously difficult. If we had an acceptable definition of knowledge, then we could count how many propositions one thinks one knows and subtract the number of propositions that one actually knows. However,

[30] Many thanks to [name removed for blind review] for pointing this issue out to me—and also for a lengthy discussion about its possible implications.

[31] This doesn't matter too much because I'll be generalising Socrates' notion to a Bayesian framework and thus going beyond what he said about wisdom anyway. Moreover, it is useful to attribute judgements to collectives, even though collectives do not have mental states, and I see no reason to think the same is not true of properties of those judgements, such as their accuracy and wisdom.

although we cannot do that, we can calculate lower and upper bounds on the difference and make comparisons in simple cases. For example, consider Ann and Bob who have exactly the same epistemic state except that Bob thinks he knows a proposition that is actually false. Since it seems to be well-accepted that a proposition must be true for the one to know it, we can infer that Bob does not actually know the proposition in question, and so we can also infer that Ann is wiser than Bob. So, in this way, we can make estimates of the difference between what one thinks one knows and what one actually knows, and that may be enough to make the comparisons of wisdom described by D2.

Of course, we may be able to make even more refined estimates, by making additional assumptions about the nature of knowledge. If belief is conceptually prior to knowledge and justification is a necessary condition for a belief to count as knowledge, then if we find cases of Bob having true but clearly unjustified beliefs, we could factor that into our estimation. I shall not pursue this line of thought in any more detail, because such assumptions are controversial[32] and it turns out that psychologists (and others) have more sophisticated measures of misestimations of knowledge, and I shall use their work to generalise Socrates' notion of wisdom to a quantitative, Bayesian framework. This is the topic of the next section.

V WISDOM AS CALIBRATION—OF IMPRECISE CREDENCES

One of the most robust findings in the psychology of human judgement is that people tend to be too confident in their judgements. In a comprehensive review of research on overconfidence, Lichtenstein 1982 concluded that: "Overconfidence is found in most tasks; that is, people tend to overestimate how much they know" (p. iv). For example, it has been found that students frequently

[32] See e.g., Timothy Williamson (2002). *Knowledge and its Limits*. Oxford University Press.

overestimate their performance on exams.[33]

Overconfidence is closely related to calibration (section III)—roughly speaking, an individual is overconfident if the probabilities that they assign, x_i, are, on average, greater than the corresponding f_i. More precisely, to measure overconfidence, we first map the x_i to the $[0.5, 1]$ interval and flip the truth values of their corresponding propositions.[34] Then, the equation that measures overconfidence is:

$$\Delta C = \frac{1}{N}\sum_{i=1}^{K} n_i(x_i - f_i)$$

If $\Delta C > 0$ then the individual is, on average, overconfident, and if $\Delta C < 0$ then he/she is, on average, underconfident. On-average overconfidence is a kind of miscalibration whereby the credences tend to be too extreme—i.e., they are too close to the end points 0 and 1. On-average underconfidence is a kind of miscalibration in the other direction, such that the credences tend not to be extreme enough. Underconfidence is sometimes observed, but it appears that overconfidence is far more common—hence Lichtenstein's conclusion: people tend to overestimate how much they know.

The terms "underconfidence" and "overconfidence" are somewhat unfortunate, for they have normative connotations—e.g., if one is *over*confident, then one is *too* confident, and so one *should be less* confident—that are not necessarily appropriate. If someone has unluckily come across some misleading evidence and yet responded to it in a perfectly rational way, then they may turn out to be, say, overconfident. And yet it seems that they should not have been less confident, for they responded appropriately to the evidence that they acquired. This is analogous to Gettier-like failures of knowledge: one can correctly think that some evidence justifies a belief that one knows

[33] Dennis E Clayson (2005). "Performance overconfidence: Metacognitive effects or misplaced student expectations?" In: *Journal of Marketing Education* 27.2, pp. 122–129. For more examples of overconfidence and a discussion of some of the controversy surrounding the details of the phenomenon see Lichtenstein *et al.* (*ibid*) and Juslin *et al.* (*ibid*).

[34] This means that, for example, a probability of 0.1 in P is treated as a 0.9 probability in $\neg P$.

some proposition, even though the evidence itself is misleading.[35] And yet it may be perfectly rational for someone to have this belief; indeed it may even be irrational for them not to have this belief.[36] For this reason, overconfidence might be more aptly called "on-average over estimation of knowledge", and similarly for underconfidence.

If overconfidence is an (on average) overestimation of knowledge and underconfidence is an (on average) underestimation of knowledge, then we have a probabilistic version of D2:

$\mathrm{D2}_{\mathrm{prob}}$ *Ceteris paribus*, *A* is wiser than *B* if *A* is better calibrated than *B*.

If *B* is more overconfident than *A*, then, *ceteris paribus*, *A* is wiser than *B*. Similarly, if *B* is more underconfident than *A*, then, *ceteris paribus*, *A* is wiser than *B*. (It is important that $\mathrm{D2}_{\mathrm{prob}}$ involves Cal and not ΔC, because it is possible for $\Delta C = 0$ and $\mathrm{Cal} > 0$. This happens when one is sometimes overconfident, sometimes underconfident, but is neither on average.)

From $\mathrm{D2}_{\mathrm{prob}}$, it follows that a collective can be wiser, in this Socratic sense, than its individuals. Consider again Example 1 (from section III). Ann systematically overestimates what she knows—only half of the propositions she assigns 0.8 to are true. Similarly, Bob also systematically overestimates what he knows. Therefore, both Ann and Bob are overconfident. However, their collective credences (which are the averages of Ann's and Bob's credences) turn out to be perfectly calibrated (because the average of 0.2 and 0.8 is 0.5). Therefore, by $\mathrm{D2}_{\mathrm{prob}}$, the collective, in this case, is wiser than its individuals.

Since the calibration of a collective can be better than the calibration of its individuals, it can be wiser than its individuals. However, recall that CW1 compares the performance of the collective against the *average* performance of the individuals in the collective. This is weaker than

[35] Edmund Gettier (1963). "Is Justified True Belief Knowledge?" In: *Analysis*, pp. 121–123.

[36] Due to limitations of space, I'm glossing over many subtleties to do with externalism/internalism of reasons, rationality, and justification.

the condition that the performance of the collective be better than *all* of the individuals. However, since $D2_{\text{prob}}$ is only about comparisons of wisdom, we are not able to always identify whether CW1 has been satisfied. One final assumption needs to be made to do this: wisdom comes in *degrees* that are negatively proportional to calibration.[37]

† *Ceteris paribus*, A's degree of wisdom is negatively proportional to A's calibration score.

It follows that, *ceteris paribus*, the lower an individual's calibration score, the wiser they are. With this final assumption in place it would appear that we have found a probabilistic analogue of Socrates' notion of wisdom and that we have also found a way for the wisdom of a collective to be greater than the average wisdom of the individuals in the collective.

However, there is an important problem with all of this. Consider the following example, in which 90% of the propositions are true:

Example 2

Truth values	1	1	1	1	1	1	1	1	1	0
Ann's credences	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Bob's credences	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Socrates' credences	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Suppose that the propositions are about some craft, so that Socrates has no knowledge about them. Given his lack of knowledge, he should not suggest otherwise, and so he does the best thing he can do: he assigns 0.5 probability to each proposition—in order to express his ignorance. However, because 90% of the propositions happen to be true, Socrates is miscalibrated by 40 percentage points, in the direction of underconfidence. So, it would seem that, far from accurately expressing his ignorance, Socrates has drastically underestimated his knowledge. The only way for him to be

[37] Recall that Socrates spoke of being wiser than the politician to "a small extent", so this is not a dramatic departure from the original text.

well calibrated in this situation is for him to assign probabilities that are more extreme. However, this is incompatible with his total ignorance of the craft at hand. Therefore, Socrates is forced to either make some false claim to knowledge or to underestimate his knowledge. Therefore, $D2_{\text{prob}}$ is not the probabilistic analogue of D2 after all.[38]

This is a version of a well-known problem with the standard Bayesian framework: it does not distinguish between an assignment of 0.5 based on ignorance, and one that is based on extensive knowledge. For example, consider two coins such that you know that one could be of any bias (towards heads or tails) and that the other is symmetric and that it landed heads 50% of the time from a million previous flips. Most people, when asked to assign a probability to the second coin landing hands will say "0.5". And most people will also say "0.5" when asked for the probability of the first coin landing heads—but some of them will do so begrudgingly, or even only when they are coerced. This is because there is more uncertainty associated with the first coin than which a simple "0.5" statement of probability can capture.

A solution to this problem is to adopt a framework based on *imprecise probabilities*.[39] For individual probability assignments, this amounts to treating probabilities as intervals of real numbers, rather than single real numbers. In what follows I shall represent the imprecise probability

[38] There is evidence that people tend to try to be like Socrates in this example and run into similar problems. For example, J Frank Yates (1982). "External correspondence: Decompositions of the mean probability score". In: *Organizational Behavior and Human Performance* 30.1, pp. 132–156 found that "subjects were inclined to report forecasts of .5 when they felt they knew little about the event in question. This finding is problematic because self-reported knowledge was only minimally related to the actual external correspondence of the subjects' forecasts."

[39] See e.g., Isaac Levi (1980). *The Enterprise of Knowledge: An essay on knowledge, credal probability, and chance*. MIT Press: Cambridge, Massachussetts; Peter Walley (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman Hall, pp. 3–4; J.M. Joyce (2005). "How Probabilities Reflect Evidence". In: *Philosophical Perspectives* 19.1, pp. 153–178; James M Joyce (2010). "A Defense of Imprecise Credences in Inference and Decision Making". In: *Philosophical Perspectives* 24.1, pp. 281–323.

of a proposition as $[l,u]$ where l is the lower probability and u is the upper probability. It will also be useful to refer to the midpoint of $[l,u]$ as m. The problem is solved by representing total ignorance as the widest open interval of probabilities $(0,1)$.[40] The midpoint of this interval is 0.5 and we model someone who is forced to assign a probability to a proposition that they know nothing about as saying 0.5 but with the widest interval $(0,1)$ around that "best guess" probability. And in cases in which more information is possessed, we may model the person with a narrower interval. For example, for the symmetric coin that has been flipped a million times, we might model the individual's response as 0.5 surrounded by an interval such as $[0.499, 0.501]$.[41]

Imprecise probabilities allow for richer and more complete representations of uncertainty. However, we need measures of overconfidence and calibration that can accommodate these richer representations. Some work in this direction has been done by [work co-authored by the author of this manuscript] *et al.*,[42] who have generalised the standard measure of overconfidence to account for imprecise credences, and to see if people appear less overconfident when the imprecision of their credences are accounted for. They begin their generalisation by noting that the standard measure is a weighted average of unweighted averages of the differences between credences and truth values:

$$\Delta C \quad = \quad \frac{1}{N}\sum_{i=1}^{K} n_i(x_i - f_i) = \frac{1}{N}\sum_{i=1}^{K} n_i \left(\frac{1}{n_i}\sum_{j=1}^{n_i}(x_i - \tau_j) \right)$$

where τ_j is the truth value of the j proposition that received a credence of x_i. They then generalise the measure by replacing the types of x_i with the types of m_i,[43] and weighting the differences

[40] There is an issue here concerning whether or not an agent with credence (0,1) can learn by conditionalistion. (See Susanna Rinard (2013). "Against Radical Credal Imprecision". In: *Thought: A Journal of Philosophy* 2.2, pp. 157–165.) This issue is beyond the scope of this paper.

[41] The exact interval, from a normative perspective, will be determined by the prior probabilities and the evidence on which the individual has conditionalised (for more details on how this can work see Walley *ibid.*).

[42] [Reference for work co-authored by the author of this manuscript].

[43] [work co-authored by the author of this manuscript] *et al.* (*ibid.*) let the m_i be anywhere between l_i and u_i. Nothing I

between the m_i and τ_j with the precision of the imprecise probability that was assigned to the jth proposition:[44]

$$\Delta C_{\text{weighted}} = \frac{1}{N}\sum_{i=1}^{K} n_i \left(\frac{1}{\mathbf{n_i}} \sum_{j=1}^{n_i} (m_i - \tau_j)(1-(u_j - l_j)) \right)$$

where $\mathbf{n} = \sum_{j=1}^{n_i}(1-(u_j-l_j))$. This has the consequence that the m_i that come from more precise probabilities have more contribution to the overconfidence score than those m_i that come from wider intervals. For example, consider:

Example 3

Truth values	1	0	1	0	1	1	1	1
Ann's credences	[0.4, 0.6]	[0.4, 0.6]	[0.4, 0.6]	[0.4, 0.6]	[0.1, 0.9]	[0.1, 0.9]	[0.1, 0.9]	[0.1, 0.9]
Bob's credences	[0.4, 0.6]	[0.4, 0.6]	[0.4, 0.6]	[0.4, 0.6]	[0.4, 0.6]	[0.4, 0.6]	[0.4, 0.6]	[0.4, 0.6]

According to ΔC (where the x_i have been set equal to the m_i), Ann and Bob are equally underconfident (because their x_is are always equal to 0.5). However, according to $\Delta C_{\text{weighted}}$, Ann is less underconfident than Bob because some of the differences between her m_i and τ_j contribute less to her underconfidence than does Bob's.

There is one complication that needs to be discussed before these ideas can be applied to collective credences and wisdom. The complication comes from the fact that if someone always assigns the same imprecise probability, then they will be treated by $\Delta C_{\text{weighted}}$ as though they had assigned precise probabilities. For example, consider the following:

Example 4

Truth values:	0	1	0	1
Ann's credences:	$[0.75, 0.75]$	$[0.75, 0.75]$	$[0.75, 0.75]$	$[0.75, 0.75]$
Bob's credences:	$[0.5, 1]$	$[0.5, 1]$	$[0.5, 1]$	$[0.5, 1]$

say here is incompatible with that more general framework.

[44] As [work co-authored by the author of this manuscript] *et al.* note, this is not the only way to introduce such a weighting scheme, and it is open question as to what the best way is.

Because Bob assigned imprecise probabilities of constant width, there are no differences between the weights $(1-(u_j-l_j))$, and since they are normalised by $\mathbf{n_i}$, $\Delta C_{\text{weighted}}$ is simply equal to ΔC, and so $\Delta C_{\text{weighted}}$ treats Ann and Bob as equally overconfident.

A solution to this problem comes from noting that although the original measure ΔC can be seen as a weighted average of unweighted averages, it can also be viewed as an unweighted average of unweighted *sums*:

$$\Delta C \quad = \quad \frac{1}{N}\sum_{i=1}^{K} n_i \left(\frac{1}{n_i} \sum_{j=1}^{n_i} (x_i - \tau_j) \right) = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_i-\tau_j)$$

The solution, then, is to make the unweighted sum a weighted one:

$$\Delta C_{\text{IP}} = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(m_i-\tau_j)(1-(u_j-l_j))$$

This new measure has the desirable result of treating Bob as being less overconfident than Ann in Example 4 while also treating Ann as less underconfident than Bob in Example 3.[45] It also collapses into the original measure, ΔC, when all of the probabilities are maximally precise.

Given ΔC_{IP}, we immediately get a corresponding measure of calibration by taking the absolute values of the weighted sums:

$$\text{Cal}_{\text{IP}} = \frac{1}{N}\sum_{i=1}^{K}\left|\sum_{j=1}^{n_i}(m_i-\tau_j)(1-(u_j-l_j))\right|$$

This means that underconfidence for one kind of m_i does not cancel out with any overconfidence for another kind of m_i. In what follows, I shall call the calibration defined by Cal_{IP} *IP-calibration*.

Now we can solve the problem with Socrates not being able to express his ignorance. Consider again the following example, but this time a new individual, Cam, has the most uninformative precise credences and Socrates has the most uninformative imprecise credences:

[45] It has the somewhat undesirable consequence that comparisons with ΔC do not make sense. [work co-authored by the author of this manuscript] *et al.*'s primary concern was to be able to make comparisons between measures that take into account imprecise probabilities and the standard measure ΔC, hence their focus on $\Delta C_{\text{weighted}}$.

Example 5

Truth values	1	1	1	1	1	1	1	1	1	0
Ann's credences	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Bob's credences	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Cam's credences	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Socrates' credences	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)

where a single number x_i, is shorthand for the maximally precise probability $[x_i, x_i]$. According to Cal$_{\text{IP}}$, Socrates is perfectly calibrated because he is maximally noncommittal with his credences, Ann is slightly mis-calibrated, Cam is worse again, and Bob is the most mis-calibrated.[46]

Given that we have moved to an imprecise probability framework to measure wisdom as IP-calibration, the definition of the collective credence also needs to be modified so that it can be imprecise and defined in possibly imprecise individual credences. A simple way of doing this, which is in the spirit of the restrictions made in section I, is to define the collective imprecise probability as $[l_c, u_c]$, where l_c is the unweighted linear average of the individual lower probabilities and u_c is the unweighted linear average of the individual upper probabilities. As before, it can turn out that a collective is better IP-calibrated than the average individual in the collective. It is also empirically feasible—see Appendix B for an example.[47]

[46] One might worry that Cal$_{\text{IP}}$ makes it too easy to be perfectly wise—because one can just be truly ignorant of everything and suspend judgement about everything. However, recall that here we are only concerned with a probabilistically analogue of D2. There is still D1 to account for, and a probabilistic analogue of that aspect of wisdom could be given. For example, we could understand D1 in terms of the Brier score over the Γ–propositions. Perfect wisdom, then, could only be attained when one's Brier score for the the Γ–propositions is 0 and one's Cal$_{\text{IP}}$ score is also 0.

[47] There are many other ways of aggregating imprecise probabilities and these may result in different empirical properties (see e.g., Walley 1991, pp. 186–8; Serafin Moral and Jose Del Sagrado (1998). "Aggregation of imprecise probabilities". In: *Aggregation and Fusion of Imperfect Information*. Springer, pp. 162–188; Robert F Nau (2002). "The aggregation of imprecise probabilities". In: *Journal of Statistical Planning and Inference* 105.1, pp. 265–282).

VI WISDOM AS CALIBRATION—OF PRECISE CREDENCES

Because I have presented the arguments of the previous sections alongside expositions of various ideas that are crucial to the arguments, I shall explicitly state the main argument here:

P1 *Ceteris paribus*, A's degree of wisdom is negatively proportional to the degree to which *A* mis-estimates his/her knowledge.

P2 The degree to which *A* mis-estimates his/her knowledge is A's IP-calibration score.

P3 From P1–2: *Ceteris paribus*, A's degree of wisdom is negatively proportional to A's IP-calibration.

P4 The IP-calibration of a collective can be better than the average IP-calibration of the individuals in the collective, and this happens empirically—see Appendix B.

C Therefore, the degree of wisdom of a collective can be greater than the average degree of wisdom of the individuals in the collective, and this happens empirically.

The support for P1 comes from D2, along with the assumption that wisdom is not merely a comparative concept and that it comes in degrees. P2 is motived by the fact psychologists and others interpret overconfidence in terms of overestimations of knowledge, and that imprecise probabilities are a way of improving the standard Bayesian framework so that epistemic states of ignorance can be appropriately modeled. P3 follows from P1 and P2. The support for P4 comes from the examples in section V and Appendix B. The conclusion, C, follows from P3 and P4.

Although I believe this is a sound argument, I acknowledge that D2 goes beyond what Socrates said about wisdom and P2 could be controversial since it involves a modification of the standard concept of calibration in terms of a non-standard model of uncertainty (*viz.* imprecise probabilities). Therefore, I will argue that the main conclusion follows from a weaker set of premises. If we interpret Socrates according to D2', then the move to imprecise probabilities that I made in the previous section is not required.

The alternative argument is as follows:

P1 *Ceteris paribus*, A's degree of wisdom is negatively proportional to the degree to which A overestimates his/her knowledge.

P2 The degree to which A overestimates his/her knowledge is A's *total degree of overconfidence* (explained below).

P3 From P1 and P2: *Ceteris paribus*, A's degree of wisdom is negatively proportional to A's total degree of overconfidence.

P4 The total degree of overconfidence of a collective can be less than the average total degree of overconfidence of the individuals in the collective, and this happens empirically—see Appendix C.

C Therefore, the degree of wisdom of a collective can be greater than the average degree of wisdom of the individuals in the collective, and this happens empirically.

The support for P1 comes from D2', along with the assumption that wisdom is not merely a comparative concept and that it comes in degrees. I will argue for P2 below. P3 follows from P1 and P2. The support for P4 comes from Examples 1, 2, and 5 and Appendix C. The conclusion, C, follows from P3 and P4.

As I mentioned earlier, on-average overconfidence is a systematic overestimation of one's knowledge. However, a measure such as ΔC allows for the possibility that some overconfidence is cancelled out by some underconfidence—this is because ΔC is a measure of *average* overconfidence (or underconfidence). (As I mentioned earlier, it is possible for $\Delta C = 0$ and $\text{Cal} > 0$.) So we need a way of measuring the total amount of overconfidence that is unaffected by instances of underconfidence. This can easily be done by modifying ΔC as follows:

$$\Delta C_{+} = \frac{1}{N}\sum_{i=1}^{K} n_i(x_i - f_i)\sigma_i, \text{ where } \sigma_i = 0 \text{ if } x_i - f_i < 0, \text{ and } \sigma_i = 1 \text{ if } x_i - f_i \geq 0$$

That is, ΔC_{+} is the total degree of overconfidence. P2 is thus the premise that the extent to which someone overestimates their knowledge is measured by ΔC_{+}. Importantly, Example 2 is a not a problem for ΔC_{+}. Recall that in Example 2, Socrates was counted as unwise, even though he did everything he could do to honestly express his ignorance using precise probabilities. This happened because ΔC for Socrates was positive. However, ΔC_{+} for Socrates in Example 2 is equal to 0, and so if degree of wisdom is negatively proportional to ΔC_{+}, then Socrates is counted as being perfectly wise (in the example, and *ceteris paribus*).

To summarise: if D2' is the correct interpretation of Socrates and a true (but partial) account of wisdom, and if the extent to which someone overestimates their knowledge is measured by ΔC_{+}, then it follows that a collective's Socratic wisdom can be greater than the average individual wisdom. I prefer D2 as an interpretation of Socrates, and, so, I believe that IP-calibration is a better measure of wisdom. However, these may be considered somewhat controversial assumptions, and so I have given an alternative argument that does not rely on them.

VII CONCLUSION

It might appear that the so-called "wisdom of crowds" and "collective wisdom" have very little to do with wisdom. However, I have argued that there is a sense in which collectives can be genuinely wise. This sense of wisdom is the one we find from Socrates in Plato's *Apologia*. According to Socrates, there are two ways to be wise: (i) to have knowledge about important or valuable matters, and (ii) to be aware of the limitations of one's knowledge. I have argued that the degree to which one is aware of the limitations of one's knowledge (as understood by psychologists) is measured by the calibration of one's imprecise credences (or, alternatively, ones total degree of overconfidence). Since the calibration (and total degree of overconfidence) of a collective can be better than the average calibration of the individuals in the collective, it follows that collectives can have a degree of wisdom greater than the average degree of wisdom of its individuals.

APPENDIX A

This is a re-analysis of data reported in [work co-authored by the author of this manuscript] *et al. (ibid.)* to show that for credences about future events elicited from real people, the collective credences are better calibrated than the average calibration of the individual credences.

METHODS

Amazon's Mechanical Turk was used to recruit 150 participants. All participants were required to be residents of the U.S., to have approval rates of previous Human Intelligence Tasks (HITs) above 90%, and to have had at least 50 prior HITs approved. Participants were asked to assign probabilities to 30 general geo-political-economical events and they were paid 1 USD for successfully doing so. Participants were asked to answer the questions by specifying a single best guess probability:

> Q What is your probability that E?

where the E is the geo-political-economical event in question.[48] Any participant who failed to answer all of the questions or agree to the IRB consent form was also rejected. In the end, there were 135 participants who gave usable responses.

[48] See [work co-authored by the author of this manuscript] (*ibid.*) for the full list of questions.

Calibration was calculated according to Cal, as defined in section III. And to estimate the calibration scores, all credences were put into bins of size 0.05.

RESULTS

The average calibration score for the individuals was 24.58, 95% CI = [23.15, 26.02]. The calibration of the collective was 12.06. Out of the 135 individuals, only 2 were better calibrated than the collective, with the best calibration score being 9.5.

APPENDIX B

This is also a re-analysis of data reported in [work co-authored by the author of this manuscript] (*ibid.*) to show that the collective imprecise credences are better IP-calibrated than the average IP-calibration of the individual imprecise credences.

METHODS

The methods were the same as those described in Appendix A, except that 150 new participants were recruited to answer the same set of questions by specifying lower, upper, and best guess probabilities—in that order. Any participant who failed to answer all of the questions or agree to the IRB consent form or who gave incoherent responses (i.e., the lower probability was above upper probability or the best guess probability was lower than the lower probability or higher than the upper probability) was rejected. In the end, there were 105 participants who gave usable responses.

Also, instead of calibration, IP-calibration was used. IP-calibration was calculated according to Cal_{IP} as defined in section V. As before, the best guess probabilities were put into bins of size 0.05 to form the m_i.

The collective lower probability is defined as the unweighted average of the individual lower credences, the collective best guess is the unweighted average of the individual best guesses, and the collective upper credence is the unweighted average of the individual upper credences.

RESULTS

Setting the m_i to be the midpoints of the intervals: The average IP-calibration of the individuals was 12.49, 95% CI = [11.28, 13.71]. The IP-calibration of the collective was 9.88. Out of the 105 individuals, 38 were better IP-calibrated than the collective, with the best IP-calibration score being 2.52.

Setting the m_i to the best guess probabilities: The average IP-calibration of the individuals was 12.88, 95% CI = [11.71, 14.05]. The IP-calibration of the collective was 8.56. Out of the 105 individuals, 29 were better IP-calibrated than the collective, with the best IP-calibration score being 2.90.

APPENDIX C

Here I show that the results from Appendix A also confirm that a collective's ΔC_+ score is better than than the average ΔC_+ score of the individuals.

METHODS

The methods are exactly those reported in Appendix A. The ΔC_+ score of a set of credences is defined as it was in section VI.

RESULTS

The average ΔC_{+} score of the individuals was 21.18, 95% CI = [19.60, 22.77]. The ΔC_{+} score of the collective was 0.69. No individual had a better ΔC_{+} score than the collective.