Imprecise Probability and the Measurement of Keynes's "Weight of Arguments"

William Peden
Centre for Humanities Engaging Science and Society, 50 Old Elvet, DH1 3HN, Department of Philosophy, Durham University, UK
w.j.peden@durham.ac.uk

Abstract

Many philosophers argue that Keynes's concept of the "weight of arguments" is an important aspect of argument appraisal. The weight of an argument is the quantity of relevant evidence cited in the premises. However, this dimension of argumentation does not have a received method for formalisation. Kyburg has suggested a measure of weight that uses the degree of imprecision in his system of "Evidential Probability" to quantify weight. I develop and defend this approach to measuring weight. I illustrate the usefulness of this measure by employing it to develop an answer to Popper's Paradox of Ideal Evidence.

Introduction

One of the oldest systems of imprecise probability in the literature is Henry E. Kyburg's system of "Evidential Probability". In this article, I shall defend and develop a proposal by Kyburg to use his system to measure what John Maynard Keynes called the "weight of arguments" [1]. The analysis of weight has been challenging for probabilistic theories of reasoning, but I shall argue that Kyburg's measure can address this issue. My discussion will illustrate the advantages of Evidential Probability for argument analysis. Additionally, to illustrate the usefulness of this measure, I shall use this measure to answer Karl Popper's "Paradox of Ideal Evidence". While there are many proposed methods for measuring weight, I shall focus entirely on Kyburg's measure.

In Section 1, I describe Keynes's concept of weight and explain its importance. In Section 2, I explain Kyburg's system and develop his proposal for measuring weight using his system; I also examine whether this measure can withstand a range of existing and novel objections to using imprecise probabilities to measure weight. In Section 3, I apply this measure to Popper's paradox.

1 Keynes and the Weight of Argument

Keynes uses the term "weight of argument" to refer to his concept. In the literature, "weight of evidence" is also common. I shall stick to Keynes's terminology, because the phrase "weight of evidence" sometimes refers to other concepts, such as the degree of confirmation of a hypothesis by the evidence. Additionally, in law, the phrase "weight of evidence" is typically used to mean the balance of evidence with respect to a hypothesis of guilt or innocence [2, p. ix]. Thus, "weight of argument" helps avoid some potential ambiguities.

The "weight of argument" of a statement H given another statement E is the extent to which E (which might be the conjunction of many distinct statements in natural language) provides information that is relevant to H. As Rod O'Donnell emphasises, the adjective 'relevant' in front of 'evidence' is very important [3]. Keynes is not referring to the sheer number of statements on the right hand side of a conditional probability $P(H \mid E)$ or the sheer bulk of information that these statements contain. By "relevant evidence", Keynes is only referring to the extent that E provides information that is pertinent to H in particular.

The informal concept of weight has a long history. I. J. Good notes that, as a metaphor, it can be traced back at least to the Ancient Greek goddess Themis and her 'scales of justice', in which the scales involve both a balance and a quantity of evidence [4]. Charles Sanders Peirce provides one of the earliest philosophical discussions [5]. He gives a simple example of weight's significance: suppose there is a bag that we know consists of red beans and/or black beans, in some unknown proportion. Intuitively, there is a difference between extrapolating from half of a sample of 1,000 beans drawn from the bag and half of a sample of only 2 beans. If someone is extrapolating from the former bag, then they seem to be entitled to a greater degree of confidence (in some informal sense) in estimating that half of the beans of a bag are red. According to Peirce, the difference is the greater quantity of evidence that the report of the larger sample provides. Thus, an argument from that report and the relevant background knowledge to a hypothesis about the proportion of red beans in the bag has more weight.

Keynes firmly distinguishes weight and conditional probability, because they can move in different directions. For example, imagine that you hear a weather report stating that the weather in your area tomorrow will not be windy. The conditional probability that tomorrow *will* be windy, given your total evidence, has decreased. In contrast, the weight of argument has *increased*. Jochen Runde notes that Keynes presupposes that weight always increases monotonically with additions of relevant evidence to a body of evidence [6]. This monotonicity of weight means that, if E is relevant to H given background knowledge K, then $(E \wedge K)$ must have greater weight

with respect to H than K alone. In this respect, weight differs from conditional probability, because conditional probability does not always increase monotonically as relevant statements are added to the conditional evidence.

However, some philosophers who have discussed weight have questioned this monotonicity presupposition. Runde argues that, under some circumstances, new evidence might reveal that our evidence is more limited than we thought [6]. James M. Joyce implicitly rejects the monotonicity presupposition in his discussion of weight [7]. Brian Weatherson argues against the monotonicity presupposition and provides an elegant example to support his denial [8]. Imagine that you are playing poker with several people. You are wondering if another player, whom I shall call "Lydia", has a straight flush. A straight flush in poker consists of five cards of sequential rank in the same suite. For instance, a hand consisting of a two, a three, a four, a five, and a six, all Diamonds, would be a straight flush. Since a dealt poker hand consists of 5 cards from a normal deck of 52 cards, there are $\frac{52!}{(5!)(47!)} = 2,598,960$ possible hands that Lydia might have. There are 10 possible straight flushes for each suit and 4 suits. Therefore, there are 40 possible straight flush. In the absence of any additional evidence that makes a particular hand more likely, the Keynesian probability that Lydia has a straight flush is $\frac{40}{2,598,960}$. In the Keynesian theory of evidence and probability, this initial total evidence provides substantial information about your conjecture that she has a straight flush. However, the subsequent combination of this initial evidence with Lydia's facial expressions, tone of voice, past bluffing behaviour, and other sources of information regarding her hand might leave you with much more vague total evidence. In Keynes's probability theory, this vagueness might make it impossible to specify a precise conditional probability for the conjecture that Lydia has a straight flush.

In Weatherson's example, the subsequent evidence is qualitative. However, there are also possible scenarios in which the evidence is entirely quantitative. For instance, imagine that I am about to make a selection from an urn containing 100 lottery tickets. 90 of the tickets are red and 10 are blue. You are wondering whether I shall draw a blue ticket. If there is no additional evidence that suggests that any one ticket is more likely to be selected than another, then the initial Keynesian probability of the hypothesis that I shall draw a blue ticket is $\frac{10}{100} = 0.1$. Suppose that you then learn that (1) I am drawing the ticket from the top of the urn and you also learn that (2) the proportion of blue tickets at the top of the urn is 0.2, 1, or somewhere in between. In other words, you now know that the ticket will be selected from a subset of the urn whose proportion of blue tickets does not match the proportion for the urn as a whole, but you only have imprecise statistical data about the subset: the proportion of blue tickets x has a value $0.2 \geq x \leq 1$. While you have acquired new evidence about the hypothesis that the ticket will be blue, it

seems that the weight of your total evidence has been reduced, because the pertinent evidence of proportions has become imprecise.

Weatherson argues that, in such situations, it is intuitive to say that the weight of argument has fallen after you have acquired this new evidence. Metaphorically, Weatherson is pointing to the possibility that additional relevant information can result in a murkier overall picture, such that the total evidence provides vaguer information regarding a proposition and therefore less weight. On the grounds of such examples, I shall *not* share Keynes's presupposition (which he does not defend) that weight is monotonic.

Keynes did not doubt that weight was an important concept, but he was unsure about precisely *how* weight has practical significance. However, there have been a considerable number of suggested applications. In formal epistemology, Janina Hosiasson utilises Keynes's concept as part of an analysis of evidence and its value [9]. Weatherson argues that probabilistic theories of reasoning, such as Bayesianism, can incorporate weight to distinguish (a) propositions for which our total evidence provides considerable information and (b) propositions for which our total relevant evidence is highly exiguous, which is a distinction that is not captured by conditional probabilities [10]. In the philosophy of law, Barbara Davidson and Robert Pargetter use weight in their analysis of the legal phrase 'beyond reasonable doubt'[11]. James Franklin has also employed the concept in the philosophy of law [12,13]. These are just a few examples of how Keynes's concept has been fecund in its philosophical applications.

Keynes doubted that weight is quantitatively measurable. However, a quantitative formalization of weight might be useful, given the employment of the concept of weight by formal epistemologists and philosophers of law. Imprecise probabilities offer one possible basis for such measurement.

2 Evidential Probability and the Weight of Argument

The oldest suggestion of using imprecise probabilities to measure weight seems to be Kyburg's proposal in a 1961 monograph. He proposed using the degree of imprecision of the intervals in his probability system, called 'Evidential Probability' [14]. (I shall use the upper-case 'Evidential Probability' to refer to Kyburg's system and lower-case phrases like, 'evidential probability' for the probability values in his system.) The probabilities in this system take the values of intervals in the range from 0 to 1. For example, the intervals can cover the whole range, such as $[0, 1]$, or some subinterval of the range, such as $[0.5, 0.75]$, or a degenerate interval such as $[0.5, 0.5]$. In a 1968 article, Kyburg restricts his claim to the measurement of weight where the

evidence is of the same sort [15]. His proposed measure is a very simple function:

Definition 1 *Weight of argument for H given E and K* $= (y - x)$ where H is a hypothesis, E is some evidence, K is the relevant background information, x is the lower bound of the evidential probability interval for H given $(E \wedge K)$, and y is the upper bound of the interval for H given $(E \wedge K)$.

However, this measure is linguistically awkward, since the values will be high when weight is low and low when weight is high. A simple modification makes the definition more verbally felicitous:

Definition 2 *Weight of argument for H given E and $K = WK = 1 - (y - x)$*

Thus, insofar as the Evidential Probability intervals are more imprecise (wider) the weight will be low; insofar as the intervals are precise, the weight will be high. Thus, when the value of H given $(E \wedge K)$ is a degenerate interval like $[0.5, 0.5]$, then the WK measure of weight will take a maximal value: $1 - (0.5 - 0.5) = 1$. When the interval is the maximally wide $[0, 1]$, the measure's value will take a minimal value: $1 - (1 - 0) = 0$. When the interval is a non-degenerate sub-interval, the measure's value will take some value between 1 and 0, depending on the imprecision of the evidential probability.

I shall argue that Kyburg's measure seems to offer a very widely applicable measure of weight, which goes at least beyond his 1968 claim. I shall begin by describing Kyburg's system of Evidential Probability. I shall then detail some simple examples of the WK measure in action. I shall finish this subsection by examining a range of existing and novel challenges for this measure.

2.1 Evidential Probability

Kyburg developed his system over 40 years, but I shall discuss his system in terms of its final version, as expressed in [16]. Kyburg's system of Evidential Probability maps interval values onto statements relative to other statements. Such values are intended to represent objective evidential relations between the two *relata*: the interval values are interpreted as a form of (non-deductive) logical relations, rather than in terms of belief. Thus, Evidential Probability is a member of the family of logical interpretations of probability. However, unlike many philosophers who use such an interpretation, Kyburg argues that the probabilities should be imprecise, in many contexts.

For the formal application of Evidential Probability, the statements in the probability relation are expressions within a formal language that models the reasoning of some agent(s). An Evidential Probability function EP has a value

$EP(H \mid E \wedge K) = [x, y]$, where the range of the function is interpreted as representing the objective degree of support that the conjunction of E and K provide for H and the interval is a closed interval with two real fraction as its limits.[1]

Instead of symmetry principles (as in Objective Bayesianism) or subjective credences (as in Subjective Bayesianism), Kyburg uses direct inference as the foundation of his probabilities. To use a simplified example, imagine that all you know about a single object a is that it is a fish. You also know that 1.5% of fish are goldfish. Kyburg (and many other philosophers) would say that the probability of the hypothesis Fa, given what you know, is 1.5%. However, this situation is extremely unrealistic. Our statistical data is typically imprecise: the proportions of predicates like 'fish' in populations are estimated with margins of error. And, as Hans Reichenbach argued [17], we always have competing statistics about reference classes for any given individual: we would not just know that a is a fish, but a fish that is differentiated from other fish via some definite description, such as 'That orange fish in Tank 7'. There is no reason why our data for the proportion of goldfish among orange fish or among fish in Tank 7 must match our data for fish in general. Thus, for Kyburg, the foundational issue in probability is formulating rules for ignoring and/or combining such competing reference class data.

To determine the range of the EP function for a hypothesis H and a body of evidence K, Kyburg proposes the following procedure. Where H concerns a single object a and a predicate F, one begins by enumerating the reference classes to which (i) a is known to belong and for which (ii) K contains data about the frequency of objects with F in that reference class. Such data can be imprecise: for example, K might include a statement R_i that only specifies that F has a relative frequency of 0.4 ± 0.05 in a particular reference class that includes the object a. For these reference classes, there will be a set of statements about frequencies in reference classes: $\{R_1, R_2, R_3 \ldots R_n\}$. These are the possible reference class statements, which report relative frequencies that might supply the interval values for $EP(H \mid K)$. Some of these reference class statements might *conflict*, in the sense that the relative frequency reported by one statement is not a subinterval of the relative frequency in the other: for example, perhaps you know that $1.5\% \pm 0.1\%$ of fish are goldfish, whereas you know that $45\% \pm 0.01\%$ of fish in Tank 7 are goldfish. Thus, there is a reference class problem: what is the appropriate reference class statement for the probability of a hypothesis, given competing reference class statements in K? Or, if there are no reasons to favour any single reference class statement to determine the

[1] I am expressing the evidential probabilities in a formalism that somewhat differs from Kyburg's, in that it is closer to the formalism of Bayesian conditional probabilities. This choice is to make it more familiar to readers who are unfamiliar with Evidential Probability. The differences from Kyburg are only superficial.

probability of the hypothesis, how should one combine conflicting statements into a single interval value?

Kyburg's answer is to sequentially apply the following rules to this set:

(1) Sharpening by Richness Compare each statement in the set via a pairwise comparison. Suppose that R_j has been deduced from a full joint distribution for a random variable Q, whereas R_i has been deduced from a marginal distribution for Q. Suppose that the interval of the relative frequency reported in R_j *conflicts* with the relative frequency reported in R_i. Sharpening by Richness requires that one ignores R_i. Put another way, Sharpening by Richness enables us to favour reference class statements that more fully articulate our statistical data.

For example, for a hypothesis about a selection from a set and given the choice between (1) statistical data about the *raw proportion* in the set and (2) statistical data about the *proportion of selections* from that set, one should favour the latter. Imagine that you are about to make a random selection from one of two decks of cards. You wondering whether to bet that the card you draw will be an Ace of Spades. Let H be the hypothesis that the card you will draw is the Ace of Spades. You know that Deck 1 is a normal deck and that Deck 2 is a normal deck *minus* the Spades. Before making your random selection, you must toss a six-sided die that you know to be normal and fair. If the die lands on 1 or 2, you must draw from Deck 1; if the die lands on a 3, 4, 5, or 6, you must draw from Deck 2. Thus, if the die lands on 2/3rds of the equiprobable possibilities, then you are certain to draw a card other than the Ace of Spades. Should you use (a) your information about the proportion of Aces of Spades in the total number of cards (1 out of 103) or (b) your information about the relative frequency of tossing the die *and* selecting the Ace of Spades? The former implies an evidential probability of $[\frac{1}{103}, \frac{1}{103}]$, whereas the latter joint distribution is the relative frequency of a 1 or 2 multiplied by the conditional frequency of selecting an Ace given that the die has landed favourably: $\left(\frac{1}{3}\right)\left(\frac{1}{52}\right) = \frac{1}{156}$, for an interval of $\left[\frac{1}{156}, \frac{1}{156}\right]$. Sharpening by Richness requires that you ignore the relative frequency about the cards alone, in favour of the joint distribution concerning the random variables of the card draw *and* the die toss. Assuming no other relative frequency statements survive competition against the latter data, the evidential probability of H will be $\left[\frac{1}{156}, \frac{1}{156}\right]$. In general, Sharpening by Richness means that statistical statements in the set of candidates for an evidential probability which are based on joint distributions must be favoured over those that are based on marginal distributions, in those circumstances where they conflict.

(2) Sharpening by Specificity Having applied the previous rule, one compares the surviving statements by a pairwise comparison. Suppose that the interval of the relative frequency reported in R_j conflicts with the relative frequency reported in R_i. If a statement R_j describes a proper subset of R_i, then R_j is preferred over R_i. Therefore, data that is more specifically about the individual in question will be favoured over more general conflicting data.

For example, suppose you are wondering if a flower in your greenhouse will bloom in March. If you know the general relative frequency of this species blooming in March and the more specific data for this species blooming in March when stored in a greenhouse, and these frequencies conflict, then one should ignore the data about the species in general. Similarly, in the earlier example of wondering if a is a goldfish, given choice between data for the proportion of goldfish among fish in general and data for the proportion of goldfish among fish in Tank 7 (the container of a) Sharpening by Specificity requires that one use the latter statistic, because you know that all the fish in Tank 7 are fish, but not all fish are in Tank 7. (The fish in Tank 7 are a proper subset of fish.) To use a gambling example, if you are selecting a ball from the upper region of an urn and you have data for the proportion of red balls on the top of the urn and for the proportion for the urn in general, then you should use your data for the upper region when determining the evidential probability that the ball is red.

Suppose that you have applied these two rules. If a single reference class statement R_j remains after one or both of these rules have been applied, then the evidential probability is $[x, y]$, where x and y are the lower and upper limits of the relative frequency data stated in R_j. For example, if R_j states that the relative frequency of F in a reference class is 0.5 ± 0.03, then $EP(Fa \mid K) = [0.47, 0.53]$. If a set of reference class statements remain, then the following rule must be applied:

(3) Sharpening by Precision If there is a statement R_j that states a relative frequency that is a proper subinterval of every other remaining statements' relative frequency data, then the evidential probability is this proper subinterval. If there is no such statement, then the evidential probability is the shortest possible cover of the intervals from the surviving members of the set.

For example, if R_1, R_2, and R_3 are the surviving statements and they provide intervals of [0.1, 0.2], [0.15, 0.3], and [0.5, 0.55], then the evidential probability is [0.1, 0.55].

Essentially the same reasoning for single-case probabilities extends to plural-case probabilities, because the Sharpening rules can be multi-member sets (a sample of

balls from the top of an urn; the penguins of Antarctica; the stars in the Known Universe etc.) as well as hypotheses about particular individuals. Kyburg develops a range of general formal properties for values of the EP function: see [16,18], and a summary of salient results on page 49 in [19]. One axiom, that will be particular important later (in Subsection 2.2) is that if $EP(\Phi \leftrightarrow \Psi \mid \Omega, 1]$, then $EP(\Phi \mid \Omega) = EP(\Psi \mid \Omega)$, for any statements Φ, Ψ, and Ω. This has the consequence that if one knows that one statement is true if and only if a second statement is true, then the two statements must have the same evidential probability relative to one's total evidence.

Despite Kyburg's references to relative frequencies, he does not identify probabilities with relative frequencies: evidential probabilities take their values from accepted statements about frequencies, but they are not identified with frequencies. Instead, they are logical probabilities, akin to those of Keynes in [2] and Rudolf Carnap in [20]. As a result, he does not deny that probabilities can be meaningfully ascribed to single-case hypotheses, such as the hypothesis that a die will land on 6 in a particular toss. This feature of Evidential Probability enables Kyburg to obviate the standard frequentist problems with single-case probabilities. Indeed, as described above, all evidential probabilities are either single-case or ultimately derived from single-case probabilities.

Since Kyburg's proposal for measuring weight is my focus in this article, I shall not discuss the various general strengths and weaknesses of Kyburg's system. (The special issue on Kyburg in [21] contains further discussion, including critical assessments of Evidential Probability from Bayesian perspectives.) Instead, I shall restrict the scope my discussion to his measure of weight.

2.2 Examples

In Section 1, I mentioned a simple type of example of weight in reasoning, which was developed by Peirce. Suppose that you are considering hypotheses about the composition of an urn full of beans. You know that the urn consists of red and/or black beans in some unknown proportion. Assume that your background information is suitable for standard statistical inferences involving sampling without replacement: the selections are apparently random, they are not independent in your model of the sampling set-up, the population of beans is finite with well-defined means for the proportions of red beans and black beans, and so on. Intuitively, as your sample of beans increases, the quantity of evidence that is available to you (the weight) increases, *ceteris paribus*. I shall now explain how the WK measure treats this scenario.

Kyburg discusses this type of sample-to-population inferences in his system at

great length in Chapter 11 of [16]. Since my focus is the WK measure, rather than the details of statistical inferences in his system, I shall be brief in my description of how such extrapolations work in Evidential Probability. Let H be the hypothesis that 49.5-50.5% of the beans in the urn are red. Let E_1 be the sample report that 2 sampled beans are red. Let E_2 be the sample report that 3,000 beans, including the beans described in E_1, also are red. Let K be your relevant background knowledge. Assume that there is sufficient information in K to calculate, using combinatorics, that between 2% and 100% of the 2-fold samples of any large finite population of beans in the urn will be matching samples, in the sense that the samples will match the population within a margin of error of 1%. The Law of Large Numbers is one obvious source of combinatoric information that might be useful, under suitable conditions, for calculating the proportion of matching samples. However, you might know other combinatoric principles that could help with the estimation; additionally, background information about the urn, urns in general, and other contingent facts might be relevant to the estimation of the proportion of matching samples in the urn. In short, the probability of an interval-valued hypothesis for a population, given some sample data and background information, will depend on the statistical estimate of matching samples of that size for that population, relative to the background information.

In this particular scenario, we have assumed that you know that the proportion of possible samples that match the population in redness, within a margin of error of 1%, is in the closed interval [0.02, 1]. If the rules of Sharpening select this statement about the set of possible compositions of the beans in the urn as your best information about H given the conjunction of E_1 and K, then the evidential probability will be $EP(H \mid E_1 \wedge K) = [0.02, 1]$. The WK measure gives the output of $(1-(1-0.02)) = 0.02$ for the weight of E_1 and K with respect to H, which reflects Peirce's judgement that E_1 provides *some* evidence about H, but almost none.

In contrast, suppose that you can calculate that the relative frequency of 3,000-fold samples that match (within a margin of error of 1%) any large finite population of beans in the urn is some proportion from 72.665% to 100%. Assume that the conditions match those described in the paragraph, *mutatis mutandis*, such that $EP(H \mid E_1 \wedge E_2 \wedge K) = [0.72665, 1]$. The WK measure for your total evidence with respect to H is now $(1 - (1 - 0.72665)) = 0.72665$. In accordance with the intuition that Peirce is trying to convey, the WK measure gives us the result that your total evidence is much greater given a 3,000-fold sample report than a 2-fold sample report. In general, in the extremely simple sort of sampling model that Peirce is implicitly describing, if the evidence simply consists of increasingly larger sample reports, and if the evidential probabilities are simply derived from the combinatoric properties of possible samples of a large finite population, then the evidential probability intervals

will narrow as the reported samples are larger. Thus, on the WK measure, the weight will increase in a linear fashion in such (idealized) circumstances.

I shall now give a different example in which the WK measure seems to work well. Imagine two different object-tossing scenarios: one in which the object you are tossing is an ordinary 1 coin; another in which the object you are tossing is a gömböc. A gömböc is a homogenous three-dimensional solid that has just two equilibria on a flat surface: (1) a stable equilibrium and (2) an unstable equilibrium. The gömböc that you are tossing has the Coptic letter '  ' on the side with its stable equilibrium point and the Coptic letter '  ' on the side with its unstable equilibrium point. You are wondering whether the next toss of the gömböc on a flat table in front of you will land '  '. In the coin tossing case, you are wondering whether the next toss of the coin will land 'heads'. You might know that 1 coins land 'heads' with a relative frequency somewhere between 49% and 51%. Assume that, given your total evidence, this data about 1 coins in general is your best information (according to the rules of Sharpening) about whether the next toss will land 'heads', such that the evidential probability interval is [0.49, 0.51]. The weight of your total evidence, on the WK measure, is the very high value of $(1 - (0.51 - 0.49)) = 0.98$.

In contrast, imagine that you are almost entirely unfamiliar with the dynamics of the gömböc and gömböcs in general. Since gömböcs are rare (the first was produced in 2007 and the shape itself was only first conjectured in 1995 by the mathematician Vladimir Arnold) there might be little readily available information on how frequently they end up on any particular side. Nonetheless, imagine that you are able to conduct a brief investigation and learn that they land on their stable equilibrium point in somewhere between 0.01% and 99.9% of tosses on a flat table. Assume that this is your best information according to the rules of Sharpening. The evidential probability that the gömböc will land '  ', given your total evidence, is [0.01, 0.999]. The weight of your total evidence, on the WK measure, is the very low value of $(1 - (0.999 - 0.01)) = 0.011$. This value reflects the intuition that you have a much greater quantity of evidence about the hypothesis that the 1 coins will land 'heads' than the hypothesis that the gömböc will land ''.

These examples suggest that Kyburg's measure works well, at least in some simple cases of reasoning. There are many more complex cases where a measure of weight might be useful, and it would be preferable to have a broad survey of them. Yet such a survey would require a very wide-ranging analysis of many types of reasoning using Evidential Probability. In its place, I shall look at some *prima facie* problems for measures of weight that use imprecise probabilities, and examine whether the WK measure is vulnerable to them.

2.3 Dilation

Some imprecise probability systems have a feature called "dilation", in which updating on apparently irrelevant evidence can cause the probabilities to become more imprecise. Some philosophers try to use dilation as a general objection against such probability systems [22] and others have sought to defend imprecise probability systems by responding to such objections [23]. For my purposes in this article, the significant feature of dilation is the challenge that they can pose to measures of weight such as the WK measure. Seamus Bradley has recently raised such an objection: in dilation scenarios, the weight of argument for a hypothesis given the total evidence will fall upon the addition of apparently irrelevant evidence; consequently, a measure that identifies weight with the degree of imprecision will register a decrease upon the addition of apparently irrelevant evidence [24]. This raises the question of whether dilation is a feature of Evidential Probability. I shall set aside both (a) whether dilation is problematic and (b) whether dilation is problematic for all the types of weight measures that Bradley discusses. I shall focus purely on the WK measure. Arthur Paul Pederson and Gregory Wheeler [14], as well as Teddy Seidenfeld [25], have noted that dilation is not a feature of Evidential Probability, but they do not argue for this point in detail. In this subsection, I shall explain why the standard dilation scenarios are impossible in Kyburg's system.

I shall adapt Bradley's dilation scenario to Evidential Probability. Imagine that I am about to randomly select a card from a pile of 40 cards. Let H be 'The card is red', X be 'The card is even', Y be 'The card is red if and only if it is even', and K be my total evidence. Suppose that my best reference class information for H's probability is that 1/2 of the cards in the pile are red, whereas I only know that the proportion of even-numbered cards in the pile is between 0 and 1. In some imprecise probability systems, learning Y results in a widening of the intervals of the probability of X given Y and K, in comparison to the probability of X given K alone.

In Evidential Probability, the value of $EP(X \mid Y \wedge K)$ is determined by the rules of Sharpening to the relative frequency data about X that $(Y \wedge K)$ implies. By assumption, the information about the relative frequency of cards in the pile is initially my best information about H. Since this data tells me that exactly half of the cards are red, such that $EP(H \mid K) = [0.5, 0.5]$. Learning Y gives me a new potential reference class statement: I now know that the card is a member of a subset of the cards in the pile (the subset of cards that are red if and only if they are even) and I must now apply the rules of Sharpening to check if I should favour my relative frequency data about this reference class (the data that states that the proportion of red cards in this class is between 0 and 1) over my information for the

pile as a whole.

The intervals [0.5, 0.5] and [0, 1] do not conflict, in the sense described in Subsection 2.1, because [0.5, 0.5] is a proper subinterval of [0, 1]. Therefore, neither Sharpening by Richness nor Sharpening by Specificity is applicable. (The latter rule is not applicable, even though [0, 1] is derived from the evidence that the card in question is a member of a proper subset of the pile, because the intervals do not conflict and Sharpening by Specificity only favours more specific reference classes when they conflict with more general reference classes.) The remaining rule is Sharpening by Precision, which requires selecting a proper subinterval of the possible intervals that have survived the first two rules, *if* such a subinterval is available; otherwise, the evidential probability is the cover of the intervals. Since [0.5, 0.5] is a proper subinterval of [0, 1], Sharpening by Precision requires selecting [0.5, 0.5] as the value of $EP(H \mid Y \wedge K)$. Therefore, the Evidential Probability of H does not become more imprecise upon learning Y. The WK measure provides the output that learning Y has not affected the weight, because $WK(H \mid K) = 1 - (0.5 - 0.5) = 1$ and $WK(H \mid Y \wedge K) = 1 - (0.5 - 0.5) = 1$.

Additionally, Evidential Probability provides the potentially intuitive result that the weight of argument for X given Y and K is greater than the weight for H given K alone. As I noted towards the end of Subsection 2.2, if $EP(\Phi \leftrightarrow \Psi \mid \Omega) = [1, 1]$, then $EP(\Phi \mid \Omega) = EP(\Psi \mid \Omega)$, for any Φ, Ψ, and total evidence Ω. Therefore, since $EP(H \leftrightarrow X \mid (H \leftrightarrow X) \wedge K) = [1, 1]$, it follows that $EP(X \mid (H \leftrightarrow X) \wedge K) = EP(H \mid (H \leftrightarrow X) \wedge K)$. Using the reasoning in the earlier paragraph and the fact that Y is equivalent to $(H \leftrightarrow X)$, we can infer that the evidential probability that the card is even, given my new total evidence, must be [0.5, 0.5]. We can also infer this directly: the card is now known to be red if and only if it is even, such that the earlier [0, 1] interval is now competing with the [0.5, 0.5] interval from the data for red cards in the pile. Neither interval conflicts, because [0.5, 0.5] is a proper subinterval of [0, 1], and therefore Sharpening by Richness and Sharpening by Specificity are both inapplicable. Once again, via Sharpening by Precision, the new evidential probability must be [0.5, 0.5], because this interval is a proper subinterval of [0, 1]. Learning Y has increased the weight for X according to Kyburg's measure, because $WK(X \mid K) = 1 - (1 - 0) = 0$ and $WK(X \mid Y \wedge K) = 1 - (0.5 - 0.5) = 1$.

For some, these are both intuitive results. The WK measure tracks the intuition that Y adds to the weight for X and does not reduce the weight for H. However, those who believe that the imprecise probabilities *should* become more imprecise in dilation scenarios and/or that the weight should decrease will not share this intuition. To comprehensively defend the result in this subsection in depth would require a full defence of Evidential Probability against alternative approaches, and this defence would be outside the scope of my argument. However, there are two arguments that

can be made for this output of Evidential Probability. Firstly, learning Y tells us that the card is a member of a subset of the pile, but it is a subset for which I have no (non-vacuous) data in Bradley's scenario; for all I know, the relative frequency of red cards in this subset is no different from the relative frequency of cards in the pile as a whole. In contrast, I *do* know that all the members of the subset are members of the pile and I have no reason to believe that they are unrepresentative of the pile with respect to their colour. As I do not seem to have any reasons to disregard the data for the pile, I should still use it.

Secondly, the claim that one should always use statistics for the most precise reference class has an apparent counterexample of singleton sets. I always know that the card is a member of the singleton containing that specific card. In Bradley's scenario, I am ignorant about the relative frequency of redness in this singleton. (Otherwise, I would have no reason to consider the pile as a whole, the subset, or any other broader reference class.) However, it seems unreasonable to say that, until I know whether the card is red or not red, my evidential probability for the hypothesis that it is red must be [0, 1], even though the singleton is the most specific possible reference class. This example suggests that, when conjecturing that a particular object a satisfies a predicate F, the mere fact that a is a member of a reference class A and that A is a proper subset of B is insufficient to require that the data for A must be favoured over the data for B. Kyburg avoids this consequence by limiting Sharpening by Specificity to comparisons in which two possible intervals are in conflict. Therefore, the evidential probability for a single-case hypothesis will be derived from the data about a singleton set *if* that data conflicts with any other candidates. For instance, if I know that the card is red, then the [1, 1] interval from my data for the singleton set will be selected over the [0.5, 0.5] interval from my data for the pile. However, such data must be ignored in Kyburg's system in cases where it does not conflict with more precise general data, such that the [0, 1] interval will not always dominate alternative data. Thus, in some circumstances and within the context of Kyburg's interpretation of probability (where probabilities are always derived via relative frequency data) it seems reasonable for relatively non-specific relative frequency information to be the basis of a probability. The selections of the [0.5, 0.5] value in the dilation scenario discussed above are instances of Kyburg's rules for adjudicating it is reasonable to use data for a reference class that is not the most specific.

One might wonder whether the evidential probabilities will widen in some variations on Bradley's scenario. For example, suppose that I learn that the card is a member of a subset of cards from a second pile. I do not know the proportion of red cards in this second pile. Will the evidential probability that the card is red, given my new total evidence, have a wider interval? If I am selecting from this subset,

then some of my original evidence be contracted, because I have learned that it is false that the card will be selected from the first pile.[2] Suppose that (a) my interval for this subset is imprecise and (b) according to the rules of Sharpening, this interval must be selected over any competing possible interval. Under such circumstances, the evidential probability for the hypothesis that the card is red will widen. Yet, while this scenario exemplifies how evidential probabilities can become imprecise upon learning new data, it is not analogous to dilation: I acquired evidence that some of my earlier information was false, and (in contrast to scenarios featuring dilation) no-one thinks that such evidence is irrelevant. The WK measure will give a relatively imprecise value in this scenario, but that is an unsurprising feature, because we have assumed that the relatively precise frequency data has become unavailable.

However, while the classic form of dilation does not occur in Evidential Probability, there are circumstances in which learning new evidence can cause a widening of evidential probability intervals. Seidenfeld examines such a scenario in [22] and I shall now discuss this possibility.

2.4 The Hollow Cube

Seidenfeld developed his "Hollow Cube" scenario in order to demonstrate that Evidential Probabilists cannot always make use of Bayesian statistical methods. In particular, joint distributions of frequencies can only be utilised for determining evidential probabilities when the evidence contains those joint distributions. In some circumstances, this inability to feature of Evidential Probability results in the intervals widening. In Seidenfeld's scenario, we are measuring the volume of a hollow cube. We hypothesise that the cube has a volume of V millilitres, where V is an interval-value. Assume that we have two available measurement methods:

- (1) We can fill the cube with a liquid. We can then calculate the probability that the cube will have a volume V given that it has been filled by the measured quantity of that liquid.
- (2) We can cut a rod which has a length equal to the cube's edge and measure the length of this rod. We can then calculate the probability that the cube will have a volume V given the result of this cutting and measuring procedure.

[2] If I am not selecting from this subset and my data for the subset is relatively imprecise, then Sharpening by Precision will require ignoring the subset data in favour of the data for the pile as a whole.

Seidenfeld notes that Bayesians can always combine these results to calculate a posterior probability for the hypothesis: it is simply a matter of using the relevant priors and likelihoods from our existing full distribution to calculate the conditional probability of the hypothesis given the conjunction of the measurements. (This assumes, of course, that we happen to share the relevant Bayesian probabilities.) In contrast, using Evidential Probability, we can only use these Bayesian methods if we know a joint distribution for the frequency that the hypothesis will be true given the results of *both* measurements. However, Seidenfeld notes that such rich information might be unavailable. Potentially, we might just have the separate frequencies for the hypothesis's truth for each measurement; these frequencies might be very different. If the report of (2) is added after the report of (1) or *vice versa*, then the evidential probability can become wider after acquiring more evidence. As a result, the degree of imprecision will increase, and the WK measure decrease, such that learning the report of (2) results in less weight for an argument for/against the hypothesis given our total evidence.

For example, it is possible that using method (1) produces a measurement that strongly indicates that the cube's volume is in the interval V, whereas using method (2) produces a measurement that strong indicates that the volume is *not* in the interval V. I shall use the following abbreviations: E_1 is the estimate for the cube's volume given method (1), E_2 is the corresponding estimate given method (2), and H is the hypothesis that the volume of the cube is in the interval V. Our background knowledge is represented by K, and by assumption K lacks a sufficiently rich full distribution to use a joint distribution for H given E_1 and E_2 for Bayesian statistical methods. However, it does contain relative frequency data for the conditional probabilities of H given E_1 and H given E_2.

Since it is assumed that the joint distribution is not available, Evidential Probability will require using confidence-interval methods from classical statistics (if possible) as an alternative [16]. Suppose that we are using a confidence level such that the inference that $\neg H$ has a $\pm$ 2% confidence interval. Assume that H is extremely unlikely given E_1 and that $EP(H \mid E_1 \wedge K) = [0.01, 0.03]$. With comparable assumptions, suppose that E_2 provides an evidential probability that is almost a mirror image, such that H is very likely given our measurement using the rod: $EP(H \mid E_2 \wedge K) = [0.97, 0.99]$.

Sharpening by Richness is not available in this case, since we have assumed that there is no available joint distribution that provides frequency data for H given both measurements in our background knowledge. Sharpening by Specificity is not available, since neither $EP(H \mid E_1 \wedge K)$ nor $EP(H \mid E_2 \wedge K)$ is based on a reference class that we know to be the subset of the other. Only Sharpening by Precision remains. Since there is no available proper subinterval, this rule requires that we use

the cover of the two potential interval, such that $EP(H \mid E_1 \wedge E_2 \wedge K) = [0.01, 0.99]$. Therefore, if we learned E_1 and obtained [0.01, 0.03] as the probability of H, then subsequently learning E_2 would result in a wider probability.

Kyburg accepts Seidenfeld's example without objection [26]. In itself, Seidenfeld's example is not problematic for Evidential Probability: if we discover that our initial probability for H was radically dependent on our choice of measurement method, then a less precise interval seems to provide an appropriate representation of our greater uncertainty upon learning this dependence. It also does not present a direct challenge to Kyburg's 1968 claim about the possible use of evidential probabilities to measure weight, because it involves different kinds of evidence; in his 1968 article, Kyburg only claims that the imprecision of evidential probabilities can measure the weight when the evidence is "of the same sort". Yet if one shares Keynes's assumption that weight must increase monotonically, Seidenfeld's scenario implies that the WK measure does not always capture the general informal concept of weight.

However, once we drop this monotonicity assumption, the Hollow Cube is no longer a problem for the WK measure. What is happening in Seidenfeld's scenario is analogous to Weatherson's poker example, which I discussed in Section 1. Adding new information to our total evidence about a hypothesis can result in a new body of evidence that is less informative about that hypothesis. Put figuratively, the combination of relevant evidence might have less weight than the sum of its parts. For instance, if we were estimating appropriate betting odds that the volume of a cube is within the interval V, then one might think that the range of reasonable betting odds is greater once we know the result of the second measurement. In other words, there is a greater degree of arbitrariness in any such decision according to Evidential Probabilists. (See Chapter 14 of [19] for Kyburg's account of how this idea can be explicated in his system and how evidential probabilities can guide betting decisions.) Using the WK measure, one can register this shift in the extent to which the evidence provides information once the discordance between the results of the two measurements becomes apparent.

One might object that a report of the second measurement is intuitively relevant to one's hypothesis, and therefore that the weight (which is the quantity of evidence in one's total information with respect to a hypothesis) should be greater. However, there is an ambiguity in "quantity of evidence" here. This phrase might mean (1) the quantity of relevant statements that are available to us, which *has* increased once the second measurement is reported. Yet it might also mean (2) the extent to which the *conjunction* of our total data (after learning the result of the second measurement) provides evidence about the hypothesis. In Seidenfeld's scenario, this latter quantity seems to have fallen, because the evidence is more ambiguous: the two measurements

gave very different results, and we lack the background information to amalgamate them via Bayesian statistical methods. It is meaning (2) that Keynes meant by "weight". The WK measure formalises this fall in this quantity.

In general, the WK measure seems to work well even in those cases where evidential probabilities widen upon learning new evidence. Firstly, learning new evidence can cause one to abandon old evidence. For example, imagine that you are studying the empirical basis for a particular homeopathic medicine's efficacy. You learn that the putative studies of this medicine are fraudulent, because the tests never actually occurred. It is possible that your subsequent evidential probabilities for hypotheses like "This homeopathic treatment alleviates nasal congestion symptoms in 83–88% of adults" will be much wider due to the contraction of relative frequency data on the medicine's efficacy. The quantity of evidence has decreased in this scenario, even though the information that the results were fraudulent is relevant evidence for such hypotheses, and this decrease will be reflected in higher values of the WK measure.

Secondly, Sharpening by Richness and Sharpening by Specificity can require that relatively imprecise intervals must be used after new evidence is acquired. For instance, recall the card selection set-up discussed in Subsection 2.2: I am making a random selection from a pile of 40 cards and considering the hypothesis that the selected card will be red. Imagine that I learn that the card is from the top of the pile. Suppose that I know that 10–20% of cards at the top of the pile are red. The [0.1, 0.2] interval conflicts with the [0.5, 0.5] interval that I could derive from my data about the pile. Sharpening by Specificity requires that I use the wider interval. Therefore, the WK measure will indicate a loss of weight. However, this fall in the WK measure reflects the intuition that (a) I have learned that the card belongs to an unrepresentative subset of the pile and (b) my evidence about this subset is less precise than my data for the pile. By contrast, in the version of this scenario that I discussed in Subsection 2.2, condition (a) was not satisfied, since I did not know that the subset is unrepresentative of the pile in its colour. Assuming that conditions (a) and (b) entail a loss of weight, the WK measure provides an adequate analysis of this possibility.

These particular examples do not provide a general proof that there are no circumstances in which evidential probabilities widen and yet the weight does not seem to decrease. Since the intervals in Kyburg's system can widen under a variety of possible situations, there is always the possibility that such a widening might not correspond to the informal concept of weight. Nonetheless, if Kyburg's system is a generally adequate theory of epistemic probability (and I have not argued, in this article, that this claim is true) the intervals for a hypothesis given new total evidence should not widen or contract except via some important change in the total evidence. Furthermore, they suggest that the WK measure is widely applicable,

even if it cannot address every possible circumstance in which we want to formalise weight.

2.5 The Problem of Corroborating Evidence

Evidential Probability intervals can be maximally imprecise: there is no wider value than [0, 1] in Kyburg's formalism. Concomitantly, the WK measure can have a minimal value: when $EP(\Phi \mid \Omega) = [0, 1]$, then $WK(\Phi \mid \Omega) = (1 - (1 - 0)) = 0$. Evidential Probability intervals can also be maximally precise: if the intervals are degenerate, then no additional evidence can increase their precision. This entails that the WK measure can have a maximal value: when $EP(\Phi \mid \Omega) = [x, y]$ and $x = y$, then $WK(\Phi \mid \Omega) = (1 - (y - x)) = 1$. The Problem of Corroborating Evidence is that even once a hypothesis is inconsistent with our total evidence, such that its evidential probability is the degenerate interval [0, 0], it seems possible to corroborate the evidence that is inconsistent with that hypothesis. Thus, the quantity of relevant evidence seems to be increasing while the WK measure stays at 1. This is a *prima facie* problem for the WK measure.

For instance, imagine that we are hypothesising that "All metal rods expand when they are heated" and that we are testing this hypothesis in a laboratory. In an experiment, we apply a heat source to the rod, but we compress the rod as it is heated. Let H be "All metal rods expand when heated", E_1 be "A metal rod was heated but did not expand" and E_2 be "A different metal rod was heated but did not expand." Suppose we accept E_1 after performing the experiment many times. (Due to measurement error and the fallibility of any instruments we could use, a single experiment will not be sufficient to establish E_1.) Since E_1 is inconsistent with H, our acceptance of E_1 reduces the evidential probability of H given the total evidence to [0, 0], such that $WK(H \mid E_1 \wedge K) = 1$. This is because we know that the hypothesis is a member of the reference class of universal generalisations and we know from deductive logic that 0% of universal generalisations with counterexamples are true. Subsequently learning E_2 and conjoining it with $(E_1 \wedge K)$ will not provide an interval other than [0, 0]. Therefore, the WK measure value for H given $E_1 \wedge E_2 \wedge K)$ will not be different from 1. Yet one might think that our total evidence with respect to H has increased.

It is important to note that there is only a problem if E_1 has been fully accepted as evidence. This does not require that we regard E_1 as irrefutable, but it does mean that if we are Evidential Probabilists, then we must accept its immediate deductive consequences, including $\neg H$. If E_1 and E_2 merely become more probable at each stage, then the evidential probability of H relative to our total evidence will just become a lower and narrower interval, such that the WK measure increases. If

accepting E_2 is genuinely adding apparently relevant information about H beyond the acceptance of E_1, then it must be via increasing the apparent reliability of E_1 and any other accepted statements that are inconsistent with H, *without* affecting the probability of H. However, the weight that a body of evidence supplies and the apparent reliability of that body of evidence are two different aspects of the strength of arguments. The WK measure is only intended to explicate weight; there is no apparent reason to require that it *also* captures apparent reliability.[3]

Once the distinction between the reliability of evidence and the quantity of evidence is made, the Problem of Corroborating Evidence does not pose a problem for the WK measure. However, it does raise the question of the relationship between weight and evidential relevance. It also illustrates that the "relevance" of evidence can be ambiguous. I shall discuss these issues later, in Section 3.

2.6 Inertia

As noted in the previous subsection, the WK measure will indicate a minimum weight when the Evidential Probability interval is the maximally wide [0, 1] interval. For example, imagine that there is a card that will be randomly drawn from a deck of unknown composition. Suppose that the only information provided by your total evidence, K, about the card's redness is that it will be drawn from a deck of cards and either 0%, 100%, or some intermediate proportion of the cards in this deck are red. Let H be the hypothesis that the card is red. Therefore, by the rules of Sharpening, $EP(H \mid K) = [0, 1]$, and the WK measure gives a value of 0 for the weight in this example. This illustrates the potential usefulness of the [0, 1] interval to enable the representation of minimal weight.

However, for some imprecise probability updating rules, it is not possible for additions of apparently relevant evidence to shift from the [0, 1] interval, unless either (a) the hypothesis has a probability of 1 given the new total evidence or (b) the hypothesis has a probability of 0 given the evidence. This feature, called "inertia", seems to have been first discovered by Peter Walley [30]. For the WK measure, the potential problem is that inertia could result in the measure failing to register apparent increases of evidence when (a) and (b) are not satisfied. For example, in the card selection scenario from the previous paragraph, suppose you learn the statement E, that '99% of the cards in the deck are red'. E seems to be relevant to H, but adding it to the total evidence will not change H's probability in some imprecise probability systems. Thus, if inertia is a feature of Evidential

[3]Kyburg does have an explication of the reliability of evidence, which he develops and applies in [16], [18], and [19].

Probability, then the WK measure would fail to register an apparent increase in weight.

As Isaac Levi has noticed, Evidential Probability does *not* feature inertia [31]. To illustrate this claim, I shall explain what the rules of Sharpening require when E is added to K. Since [0.5, 0.5] is a proper subinterval of [0. 1], Sharpening by Precision requires using the information from E such that $EP(H \mid E \wedge K) = [0.5, 0.5]$. More generally, any evidence that can be combined with K to derive a relatively precise reference class statement that is a candidate for the probability of H will increase the weight, because the interval from the new reference class statement will not conflict with the [0, 1] interval and it will be more precise. In any such scenario, the rule of Sharpening by Precision will require that one uses the more precise interval. Furthermore, this rule was developed independently of the problem of inertia, and therefore it is not just an *ad hoc* response to this issue.[4]

Levi criticises Kyburg's avoidance of inertia as *creatio ex nihilo*. His worry seems to be that Evidential Probability makes it possible to move from a state of complete ignorance to a precise probability distribution for H, without conditionalizing on a prior distribution. For the measurement of weight, the problem would be that Kyburg avoids inertia only by supposing evidence that is not present. However, regardless of whether Kyburg's overall system is satisfactory, his method of avoiding inertia is grounded in evidence: *if* we have information about a relative frequency that is more precise than the [0, 1] interval and *if* this is a suitable basis for the probability of a hypothesis H, then according to Kyburg this more precise relative frequency should be used. The Evidential Probabilist cannot shift from a [0, 1] interval for a hypothesis without suitable evidence that provides pertinent relative frequency data. One can approve or disapprove of Kyburg's method of creation, but it is incorrect to say that it proceeds from nothing, because the relative frequency data is essential.

Section Summary. In this section, I have reviewed some cases in which the WK measure works well, as well as some potentially difficult scenarios. The measure's successes in these areas does not prove that it is universally applicable. Indeed, it is plausible that there will be some forms of argumentation (perhaps mathematical argumentation) that are not amenable to analysis using Evidential Probability. Conceivably, the WK measure will be unable to measure weight in these areas. Nonetheless, the measure is promising and it seems broadly applicable. For example, unlike Kyburg in his 1968 article, I have not restricted the scope of an Evidential

[4]It might be *ad hoc* in other senses, but it was not specifically formulated to avoid the issue that Walley raises.

Probability-based measure of weight to arguments featuring evidence "of the same sort". However, one might wonder whether such the WK measure offers any worthwhile fruits. I shall now turn to an instance in formal epistemology where the WK measure can aide philosophical analyses.

3 The Paradox of Ideal Evidence

Popper's Paradox of Ideal Evidence (PIE), in [33] is a criticism of what he calls "subjective" theories of probability. By "subjective", Popper means theories that interpret probability (at least in some cases) as an epistemic concept concerning rational belief, rather than as a mind-independent concept like relative frequencies or propensities. Thus, theories like Subjective Bayesianism *and* Objective Bayesianism are "subjective" in Popper's sense. (He mentions the probability theories of Keynes and Carnap in particular.) Popper develops the following scenario: suppose that you have a coin, Z, when you have no knowledge of the bias or fairness of Z. Let H be the conjecture that "the nth unobserved toss of Z will be heads". Your background knowledge has no empirical evidence regarding H and you assign a probability of 0.5 to this conjecture. You subsequently learn E, which is a statistical report that is "ideally favourable" to this assignment of 0.5, such as a report stating that 1,500 out of the 3,000 tosses landed heads.[5] In your probability model for the coin, the tosses are not independent, such that earlier tosses of the coin can be relevant to subsequent tosses. Assume that the conditional probability of H given your new total evidence is equal to the prior probability, such that $P(H \mid E \wedge K) = 0.5$. In some "subjective" probability theories, these values for the prior probability and the conditional probability will be mandatory, whereas in other theories (such as Subjective Bayesianism) they are merely permissible. In either case, Popper's example can be formulated.

There are actually two different problems that Popper develops from this scenario. I shall call these the 'Relevance Paradox' and the 'Representation Paradox'. Both are objections to "subjective" theories of probability, including Evidential Probability. I shall discuss them separately, and argue that the WK measure can help facilitate an answer to both paradoxes.

3.1 The Relevance Paradox

One of the putative applications of "subjective" theories of probability is that they can be used to define evidential relevance. The standard definition dates back to

[5] Popper's argument can also be made in terms of margins of error, but I shall simplify matters by focusing on the case where the sample frequency exactly matches the prior probability.

Keynes in [2]. Branden Fitelson notes that it has been popular among probabilists (including Bayesians) ever since Keynes's analysis [35]. On this definition, for some suitable "subjective" *precise* probability function P, any statement Φ, any hypothesis Ψ, and background knowledge Ω:

Evidential Relevance (1) Φ is evidentially relevant to Ψ, relative to Ω, if and only if $P(\Psi \mid \Phi \wedge \Omega) \neq P(\Psi \mid \Omega)$.

Put another way, in this analysis of evidential relevance, Φ is *evidentially* relevant to Ψ (relative to Ω) if and only if Φ is *probabilistically* relevant to Ψ (relative to Ω), for some suitable precise probability function P.

Popper's scenario raises a problem for this definition. The report of 3,000 tosses seems to be relevant to H, relative to the implicit background knowledge. However, since $P(H \mid E \wedge K) = P(H \mid K)$, the standard definition implies that E is not relevant to H in Popper's example. Thus, the first sub-paradox within Popper's PIE, which I have called "the Relevance Paradox", challenges analyses of evidential relevance that are based on "subjective" probability functions.

I shall not engage in a full critical discussion of the various responses that have been made to the PIE, since my objective is to use the WK measure to argue that there is *at least one* adequate answer to the Relevance Paradox, rather than that there is *only one* adequate answer. Nonetheless, I shall mention that Keynes seems to anticipate problems with his initial definition of relevance, and develops a "stricter and more complicated definition" on page 55 of [2] to address such problems. Unfortunately, Carnap, in section of [20], proved a trivialization problem for Keynes's strict definition: almost any arbitrary statement will be evidentially relevant to almost any other arbitrary statement. If a statement A that has an implication E that is probalistically relevant to H, then A is relevant to H on Keynes's "stricter" definition. Yet Carnap notes that, in classical logic, any statement A implies $(A \vee H)$, and this disjunction will be probabilistically relevant to H probability functions, except in special cases. (I explain the details of these special cases when I return to Carnap's trivialization problem.) Thus, one obvious adequacy condition for any answer to the Relevance Paradox is that it also addresses the issues that Carnap raises. I shall develop an answer that can address both challenges.

Consider what happens if we alter the standard definition to one using Evidential Probability:

Evidential Relevance (2) Φ is evidentially relevant to Ψ, relative to Ω, if and only if $\mathrm{EP}(\Psi \mid \Phi \wedge \Omega) \neq \mathrm{EP}(\Psi \mid \Omega)$.

(The only difference from Evidential Relevance (1) is that we have switched from a subjective precise probability function to the Evidential Probability function EP.)

Unlike precise probability functions, evidential probabilities can differ in two ways. Firstly, the means of the limits of the intervals can differ. For example, if $EP(H \mid E_1 \wedge K) = [0.8, 0.85]$ and $EP(H \mid E_2 \wedge K) = [0.5, 0.55]$, then the mean of the limits of the first probability is 0.825, whereas the value for the second is 0.525. Informally, the value of $EP(H \mid E_1 \wedge K)$ is "greater" than the value of $EP(H \mid E_2 \wedge K)$. Secondly, the degree of imprecision of the intervals can differ. For example, if $EP(H \mid E_1 \wedge K) = [0.1, 0.3]$ and $EP(H \mid E_2 \wedge K) = [0, 0.4]$, then the latter value is more imprecise, even though one cannot say that either probability value is "greater". Of course, it is also possible two intervals to contrast in both respects, such as [0.1, 0.5] and [0.9, 0.95]. Therefore, if $EP(\Psi \mid \Phi \wedge \Omega) \neq EP(\Psi \mid \Omega)$, the difference might be a difference in mean value of the intervals, a difference in the precision of the intervals, or both.

The WK measure will increase or decrease when an increase or decrease in precision occurs, because the WK measure is proportionate to the degree of imprecision of evidential probabilities. We can use the WK measure to interpret how differences of precision can constitute evidential relevance: the additional information has either increased or decreased the weight of an argument from one's total evidence to the hypothesis in question. This enables an answer to the PIE that is intuitive and informally grounded, yet also systematic and formal: in Popper's scenario, the weight has increased. I shall now sketch this answer in more detail.

Suppose that you are almost totally ignorant about the coin in question. (Perhaps it has an extremely unusual shape, like a gömböc, that could result in a long-run frequency of 'heads' that is very high or very low.) In particular, suppose that you have no relative frequency knowledge that can be used to determine an evidential probability value with any precision. Therefore, the probability of the hypothesis H, that the coin will land heads, given your background information K, is $EP(H \mid K) = [0, 1]$. Such a position of extreme ignorance seems to be the sort of situation that Popper had in mind[6].

By combining E with your background knowledge, you might be able to use to infer that the long-run relative frequency of the coin landing heads is $0.5 \pm \epsilon$, where ϵ is 3%. Whether this value is correct will depend on whether it is acceptable, given your background knowledge, that the sample of tosses is representative, in the sense of matching the long-run relative frequency within a margin of error of 3%. (Your background knowledge might give you reasons to doubt that the sample is representative of the long-run relative frequency.) Assuming that you can extrapolate from

[6]My discussion on this point would not differ significantly if we assume that your have stronger background information, such as the knowledge that coins with this shape land at a relative frequency somewhere in the interval [0.01, 0.99] and that you can use this information to determine that $EP(H \mid K) = [0.01, 0.99]$.

E to the long-run relative frequency, you have learned from $(E \wedge K)$ that:

R_1 The long-run relative frequency of heads for this coin is between 0.47 and 0.53.

Since the nth toss is a member of the reference class of long-run tosses of the coin, it follows that the set of reference class statements for H has expanded, because it now includes R_1 and this statement is one of the statements that must be considered when determining $EP(H \mid E \wedge K)$. If R_1 is the appropriate reference class statement for H according to Sharpening, then $EP(H \mid E \wedge K) = [0.47, 0.53]$. Therefore, E can be relevant to H, relative to K, according to **Evidential Relevance (2)**. Using the WK measure, we can describe what has happened in Popper's example: E is relevant, but its relevance consists in increasing the weight.

One important aspect of this definition of evidential relevance and its interpretation using the WK measure is that evidence can be relevant by *decreasing* the weight of argument, as well as by increasing it. For example, in Seidenfeld's Hollow Cube scenario from Subsection 2.4, the second measurement is relevant to the hypothesis, but its relevance consists in decreasing the weight, as well as disconfirming the hypothesis that the volume is the cube is the high initial interval from the first measurement and the background knowledge.

As I mentioned earlier, Carnap proved that some definitions of evidential relevance suffer from a triviality problem. For example, on Keynes's strict definition, A is evidentially relevant to H, relative to K, if A classically entails E and E is probabilistically relevant to H given K. Carnap proved that, if $P(H \mid K)$ is neither 0 nor 1 and $P(A \mid K) \neq 1$, then A is evidentially relevant to H on Keynes's strict definition, for *any* statement A. Thus, 'There is a planet in the Solar System beyond Pluto' is relevant to 'Dark energy exists', because the former implies 'There is a planet in the Solar System beyond Pluto or dark energy exists', the this disjunction is probabilistically relevant to 'Dark energy exists', and neither statement has a value of 0 or 1 given current scientific knowledge.

In Evidential Probability, this trivialization proof will not work, because evidential probabilities can only differ if there is relative frequency data that alters that intervals via the Rules of Sharpening. Thus, if $(E \wedge K)$ does not imply any new reference class statements, then $EP(H \mid E \wedge K)$ will not differ from $EP(E \mid K)$, and E will not be relevant according to **Evidential Relevance (2)**. There are further issues for analyses of evidence that I have not addressed. For example, while the PIE suggests that probabilistic definitions of evidential relevance might be too narrow, there are some paradoxes in confirmation theory in which definitions of evidence are *arguably* too broad. These include the New Riddle of Induction (developed by Nelson Goodman in [37] and [38]) and the Paradox of the Ravens (developed by

Carl Hempel [39]). Neither paradox was originally targeted at probabilistic definitions of evidence, but they are further potential problems for any theory of evidence. Naturally I have not addressed these challenges and others in this paper. However, my argument does evince how one particular problem regarding evidential relevance can be addressed using the WK measure. In particular, the WK measure enables us to formalise how evidence can be relevant to a hypothesis *even if* that evidence neither confirms not disconfirms the hypothesis.

3.2 The Representation Paradox

One issue that Popper raises in light of this scenario is that, on standard "subjective" theories, an agent's degree of belief in H should not change upon learning E, and he claims that this entails that "subjective" probabilities cannot represent the change in weight, and therefore "subjective" probabilities cannot adequately represent evidential relations. However, the representation of evidential relations was part of the original motivation for the development of "subjective" probability theories. In contemporary terminology, Popper is claiming that probabilities cannot distinguish *ignorance*, the absence of relevant evidence, from *equivocation,* the presence of evenly-balanced evidence. This charge against theories like Bayesianism is still common today: see [40] and [41]. I shall call this issue the "Representation Paradox."

There are a number of possible lines of response for a "subjective" probabilist, but the WK offers a simple and direct reply to Popper: since weight is formalizable in terms of the imprecision of Evidential Probability intervals, it *is* possible to represent an increase in weight using "subjective" probabilities. The Evidential Probability intervals will narrow as the weight of argument for H given the total evidence increases. Thus, in Popper's scenario, prior to acquiring the report of tosses of the coin, you are in a state of complete ignorance with respect to the nth coin toss; once you have acquired the evidence from the report (and assuming that it can be combined with your background knowledge to enable you to utilise relatively precise frequency data about the coin) your total evidence leaves you in an equivocal position regarding the nth toss.

This shift from ignorance towards equivocation could have practical significance. Suppose that you were trying to estimate a precise expected value for H. (The context of the estimate might be gambling, but Popper's PIE is adaptable to a variety of possible situations.) In Chapter 14 of [17], Kyburg argues that the range of rational expected values is equal to the width of the Evidential Probability intervals. Thus, prior to learning any evidence about the coin, your choice of expected value is very arbitrary. However, if the interval for H given your total evidence becomes

narrower, then only a value within this relatively narrow interval will be rational in Kyburg's decision theory. There is much more to be said here, as Kyburg's decision theory is fairly inchoate. Nevertheless, for the Representation Problem, the salient point is that if we assume that Evidential Probability is an adequate "subjective" probability system and that the WK measure is an adequate measure of weight, then "subjective" probabilities can represent what is happening in the PIE, and there will be *some* connection between the greater quantity of evidence and practical decisions.

Answering the Representation Paradox by appealing to the weight of argument is not an original move (for instance, see [7] and [42]) but the particular use of the WK measure is novel. Furthermore, there are some ancillary features of Kyburg's system which might make the WK particularly attractive as a means of formulating this response to Popper. For example, the width of Evidential Probability intervals are objectively determined, in the sense that they are always the same for given evidence. (See Chapter 12 of [16] for a detailed exposition of this point.) Put another way, two agents using Evidential Probability who have identical relevant evidence and a shared domain (the statements of a language) must have identical intervals for a given hypothesis, regardless of their subjective opinions. Since the WK measure is a function whose only variables are the limits of Evidential Probability intervals, this form of objectivity is also a feature of the WK measure. Consequently, on the WK measure, weight will be significantly independent of opinion. (There can still be a subjective element in what evidence we accept or what domain we choose.) Therefore, the WK measure can play a useful and novel role in answer to the Representation Paradox.

One possible objection to my answer is that a person might know that the precise relative frequency of 'heads' for tosses of a type of coin. In particular, they might know that the statistical statement that '50% of tosses of this type of coin land heads'. Suppose that this knowledge is the person's best reference class data, such that $EP(H \mid K) = [0.5, 0.5]$. No subsequent evidence for the frequency of 'heads' reference class will result in a more precise probability for the hypothesis given the total evidence. (This subsequent evidence could take the form of sample data for the type of coin, confirmation of a physical model for shapes with the coin's symmetries and asymmetries, analogical evidence from similar types of coins, and so on.) Yet their epistemic state does seem to have changed in some respect when such evidence is acquired.

My response to this objection is that (1) probabilities given evidence and (2) weight are collectively insufficient to characterise *all* the important aspects of someone's epistemic state, *but* that Evidential Probability can characterise the other aspects as well. In particular, the "subjectivist" can say that additional evidence can increase the reliability of one's existing evidence. This reliability can also be

explicated using Evidential Probability. I mentioned Kyburg's formalisation of this concept at the end of Section 2.5: Kyburg develops a probabilistic model of the reliability of evidence as part of a broader model of scientific knowledge, in which evidence is ranked depending on its reliability given one's foundational evidence. Alternative probabilistic analyses of this aspect of reasoning include Richard Jeffrey's version of conditionalization [43] (in which we can model the evidence as merely probable) and the modelling of reliability by Luc Bovens and Stephan Hartmann in Chapter 3 of [44]. In the coin-tossing case, a probabilist can argue that it is the statistical statement that '50% of tosses of this type of coin' land heads that becomes more probable.[7] By using formalisations of relative probabilities *and* weight *and* reliability, a "subjectivist" can represent the evidential relations in such variations on Popper's PIE.

To formulate these answers to the Relevance Paradox and the Representation Paradox, I have made strong assertions and assumptions about the adequacy of both (a) Evidential Probability and (b) the WK measure. Since Evidential Probability currently has a very marginal position within imprecise probability theory (let alone the philosophy of science in general) I expect that both (a) and (b) will be objectionable to most readers. As my aim in this section has been to exemplify how the WK measure is philosophically useful, the fact that its adequacy plays an important part in my arguments above supports that aim, rather than undermines it. I have also not proven that these answers are superior to alternative responses to the PIE. However, my objective has been to develop a novel answer to the PIE, rather than to prove that this answer is uniquely adequate; it is plausible that there are multiple viable responses that defenders of "subjective" theories might make to Popper's paradox.

4 Conclusion

Weight seems to be an important part of argument strength. Keynes was sceptical about its quantitative measurement, but I have argued that the WK measure offers a propitious formalization of weight. Kyburg thought (at least from 1968 onwards) that measures of weight based on Evidential Probability would be limited to evidence "of the same sort", but once one drops the assumption that weight must increase monotonically, his approach can be expanded to a much more general measure. I have also illustrated how the WK measure can be used to formulate new arguments

[7] There could be further responses, such as denying that precise statistical generalisations can be acceptable, but I have accepted the details of the objection so that my defence is independent of such controversies.

within formal epistemology. I have not discussed alternative measures, such as Walley's measure in [24]. It might be the case that different measures of weight are preferable for analysing different arguments and inferences, so that alternative measures could be attractive even for ardent Evidential Probabilists. However, my aim has been to promote the WK measure, rather than to claim that other measures are unsatisfactory. I hope that my discussion has furthered this aim and illustrated the potential of Kyburg's system for the analysis of evidential relations.

Acknowledgements. I am grateful to Julian Reiss, Nancy Cartwright, Wendy Parker, Rune Nyrup, and the rest of the team at CHESS, for their assistance in the development of this article. I was also helped by a very encouraging and insightful group of referees.

References

[1] C. S. Peirce. The Probability of Induction. In C. Hartshorne and P Weiss, editors, *Elements of Logic*, volume 2 of *Collected Papers of Charles Sanders Peirce*, pages 82–105, Harvard, Mass. 1932. Harvard University Press.

[2] J. M. Keynes. *A Treatise on Probability.* Macmillan, London,

[3] D. A. Nance. The Burdens of Proof: Discriminatory Power, Weight of Evidence, and Tenacity of Belief. Cambridge University Press, Cambridge, 2016.

[4] R. M. O'Donnell. *Keynes: Philosophy, Economics and Politics.* Macmillan, Press, Basingstoke, 1989.

[5] I. J. Good. Weight of Evidence: A Brief Survey. *Bayesian Statistics*, 2: 249–270, 1985.

[6] J. Runde. Keynesian Uncertainty and the Weight of Arguments. *Economics and Philosophy*, 6 (2): 275–292, 1990.

[7] J. M. Joyce. How Probabilities Reflect Evidence. *Philosophical Perspective* 19 (1): 153-178, 2005.

[8] B. Weatherson. Keynes, uncertainty and interest rates. *Cambridge Journal of Economics*, 26 (1): 47-62, 2002.

[9] J. Hosiasson. Why Do we Prefer Probabilities Relative to Many Data? *Mind*, 40 (157): 23-26, 1931.

[10] B. Weatherson. The Bayesian and the Dogmatist. *Proceedings of the Aristotelian Society*, 107 (1pt2): 169–185, 2007.

[11] B. Davidson and R. Pargetter. Guilt Beyond Reasonable Doubt. *Australasian Journal of Philosophy*, 65 (2): 182–187, 1987. H.

[12] J. Franklin. Resurrecting Logical Probability. *Erkenntnis*, 55 (2): 277–305, 2001.

[13] J. Franklin. Case comment – United States vs. Copeland, 369 Supp. 2d 275 (E.D.N.Y. 2005): quantification of the 'proof beyond reasonable doubt' standard. *Law, Probability and Risk*, 5 (2): 159–165, 2006.

[14] Kyburg. *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middletown Connecticut, 1961.

[15] H. E. Kyburg. Bets and Beliefs. *American Philosophical Quarterly*, 5 (1): 54–63, 1968.

[16] H. E. Kyburg and C. M. Teng. *Uncertain Inference*. Cambridge University Press, Cambridge, 2001.

[17] H. Reichenbach. *The Theory of Probability*. Berkeley, University of California Press, 1949.

[18] H. E. Kyburg. *The Logical Foundations of Statistical Inference*. Dordrecht, Reidel, 1974.

[19] H. E. Kyburg. *Science and Reason*. Oxford University Press, New York, 1990.

[20] R. Carnap. *The Logical Foundations of Probability*. University of Chicago Press, Chicago, 1962.

[21] *Synthese*, 186 (2), 2012.

[22] R. White. Evidential Symmetry and Mushy Credence. In T. Szabó Gendler and J. Hawthorne, editors, *Oxford Studies in Epistemology*, pages 161–186, Oxford. 2010. Clarendon Press.

[23] A. P. Pedersen and G. Wheeler. Demystifying Dilation. *Erkenntnis*, 79 (6): 1305–1342, 2014.

[24] Bradley, Seamus, "Imprecise Probabilities", *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.), `http://plato.stanford.edu/archives/sum2015/entries/imprecise-probabilities/`

[25] T. Seidenfeld. Forbidden Fruit: When Epistemological Probability may *not* take a bite of the Bayesian apple. In W. Harper and G. Wheeler, editors, *Probability and Inference: Essays in Honour of Henry E. Kyburg, Jr.*, pages 267–279, London. 2007. College Publications.

[26] H. E. Kyburg. Bayesian Inference with Evidential Probability. In W. Harper and G. Wheeler, editors, *Probability and Inference: Essays in Honour of Henry E. Kyburg, Jr.*, pages 281–296, London. 2007. College Publications.

[27] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman, London, 1991.

[28] I. Levi. Probability Logic and Logical Probability. In W. Harper and G. Wheeler, editors, *Probability and Inference: Essays in Honour of Henry E. Kyburg, Jr.*, pages 255–261, London. 2007. College Publications.

[29] K. Popper. *The Logic of Scientific Discovery*. Hutchinson, London. 1980.

[30] B. Fitelson. Goodman's "New Riddle". *Journal of Philosophical Logic*, 37 (6): 613-643, 2008.

[31] N. Goodman. A Query on Confirmation. *The Journal of Philosophy* 43 (14): 383-385, 1946.

[32] N. Goodman. *Fact, Fiction, and Forecast*. The Athlone Press, University of London,

1954.

[33] C. G. Hempel. Studies in the Logic of Confirmation (I.). *Mind*, 54 (213): 1-26, 1945.

[34] J. D. Norton. Challenges to Bayesian Confirmation Theory. In S. Bandyopadhyay and M. R. Forster, editors, *Philosophy of Statistics*, pages 391–440, Oxford. 2011. Elsevier B. V.

[35] J. Reiss. What's Wrong With Our Theories of Evidence? *Theoria: An International Journal for Theory, History and Foundations of Science* 29 (2): 283-306, 2014.

[36] R. O'Donnell. Keynes's Weight of Argument and Popper's Paradox of Ideal Evidence. *Philosophy of Science*, 59 (1): 44-52, 1992.

[37] R. C. Jeffrey. *The Logic of Decision*. New York: McGraw-Hill, 1965.

[38] L. Bovens and S. Hartmann. *Bayesian Epistemology*. Oxford: Clarendon Press, 2003.

Received 14 May 2017