Information and Inaccuracy

William Roche and Tomoji Shogenji

Abstract

This paper proposes a new interpretation of mutual information (MI). We examine three extant interpretations of MI by reduction in doubt, by reduction in uncertainty, and by divergence. We argue that the first two are inconsistent with the epistemic value of information (EVI) assumed in many applications of MI: the greater is the amount of information we acquire, the better is our epistemic position, other things being equal. The third interpretation is consistent with EVI, but it is faced with the problem of measure sensitivity and fails to justify the use of MI in giving definitive answers to questions of information. We propose a fourth interpretation of MI by reduction in expected inaccuracy, where inaccuracy is measured by a strictly proper monotonic scoring rule. It is shown that the answers to questions of information given by MI are definitive whenever this interpretation is appropriate, and that it is appropriate in a wide range of applications with epistemic implications.

1. Introduction
2. Formal Analyses of the Three Interpretations
2.1 Reduction in doubt
2.2 Reduction in uncertainty
2.3 Divergence
3. Inconsistency with EVI
4. Problem of Measure Sensitivity
5. Reduction in Expected Inaccuracy
6. Resolution of the Problem of Measure Sensitivity
6.1 Alternative measures of inaccuracy
6.2 Resolution by strict propriety
6.3 Range of applications
7. Global Scoring Rules
8. Conclusion
Appendix A
Appendix B

1. Introduction

Mutual information (MI) is a powerful tool in many areas of research beyond communication theory to which Shannon ([1948]) introduced it originally.[1] MI is formally straightforward. Let X

[1] See (Cover and Thomas [2006], Ch. 1) for an overview of how information theory is used in different areas of research.

= $\{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ be partitions (of propositions).[2] Let P be a probability function. MI(X; Y) is defined as follows:[3]

$$\mathrm{MI}(X;Y) =_{\mathrm{def}} \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i \wedge y_j) \log \frac{P(x_i \wedge y_j)}{P(x_i)P(y_j)} \qquad (1)$$

It is easy to see that MI(X; Y) = MI(Y; X). Hence the term 'mutual' in the expression 'mutual information'.

There are two common interpretations of MI. One of them comes from the idea that information reduces doubt. According to this interpretation, the amount of information the proposition y provides on the proposition x is the amount of reduction in doubt about x due to y. MI, which is defined over the partitions X and Y, is then taken to be the expected amount (the weighted average over $X \times Y$) of information that a member of Y provides on a member of X.[4] The other common interpretation of MI comes from the idea that information reduces uncertainty about X as to which of its members is true. According to this interpretation, the amount of information the proposition y provides on the partition X is the amount of reduction in uncertainty about X due to y. MI is then taken to be the expected amount (the weighted average over Y) of information that a member of Y provides on X.[5,6]

Common as they are, we aim to show that the two interpretations are inadequate for many applications of MI. Consider:

> **Epistemic Value of Information (EVI)**: The greater is the amount of information we acquire, the better is our epistemic position, other things being equal.

---

[2] Mutual information is sometimes formulated in terms of *variables*. See (Cover and Thomas [2006]). But since variable talk can always be translated into partition talk (though infinite partitions are needed in cases involving continuous, as opposed to discrete, variables), and since variable talk seems rather forced in many contexts, we prefer to formulate mutual information in terms of partitions.

[3] We are assuming, as is standard in information theory, that the log base is 2. But nothing essential for our purposes hinges on this assumption.

[4] Fano ([1961], Ch. 2) explains MI along these lines (though he does not speak in terms of reduction in doubt).

[5] Cover and Thomas ([2006], Ch. 2) explain MI in terms of reduction in uncertainty.

[6] The terms 'doubt' and 'uncertainty' are standard but somewhat arbitrary. For example, some may say that y reduces 'doubt' as to which member of X is true. The important point is that on the first interpretation information is provided by a proposition y about a proposition x, whereas on the second interpretation information is provided by a proposition y about a partition X.

EVI is assumed in many applications of MI. The two common interpretations of MI, though, are inconsistent with EVI. This prompts us to seek an alternative interpretation of MI that is appropriate for those applications.

A third extant interpretation of MI that is adopted in some contexts comes from the idea that information changes the distribution of probabilities. According to this interpretation, the amount of information the proposition y provides on the partition X is the amount of divergence between the original probability distribution over X and the updated (in light of y) probability distribution over X. Unlike the second interpretation by reduction in uncertainty, any change in the probability distribution over X counts as a gain in information even if there is no reduction in uncertainty. MI, which is defined over the partitions X and Y, is then taken to be the expected amount (the weighted average over Y) of information a member of Y provides on X.[7]

The third interpretation fares better than the first two in that it is consistent with EVI. However, it is faced with the problem of measure sensitivity. MI is formally the expected amount of Kullback-Leibler Divergence (D_{KL}), but there are many other formal measures of divergence by which MI is not the expected amount of divergence.[8] If the amount of information is simply the amount of divergence between the prior and posterior probability distributions, then we can use any of those alternative measures. As a result, the answers given by D_{KL} and MI to questions of information are not definitive. We will show that some important principles of information theory provable on MI fail to hold on some alternative measures of expected divergence. Such principles are then measure sensitive and cannot be regarded as uncontestable principles of information, according to the third interpretation.[9]

We propose a fourth interpretation of MI by reduction in expected inaccuracy, where inaccuracy is measured by a strictly proper monotonic scoring rule. This interpretation is consistent with EVI because the greater is the amount of reduction in expected inaccuracy, the better is our epistemic position, other things being equal. Moreover, it is impervious to the problem of measure sensitivity because D_{KL}, of which MI is the weighted average, is itself the weighted average of inaccuracy as measured by the logarithmic scoring rule (SR_L), and SR_L is

---

[7] We noted above that Cover and Thomas ([2006], Ch. 2) interpret MI in terms of reduction in uncertainty. There we had in mind their main (or most common) interpretation of MI. It is worth noting, though, that Cover and Thomas ([2006], Ch. 2) sometimes interpret MI in terms of divergence. Perhaps they hold that there are no significant differences between the two interpretations. Our view, to be explained and defended below, is that, on the contrary, there are some significant differences between the two interpretations.

[8] We could also raise this point against the first two interpretations of MI; there are alternative measures of reduction in doubt by which MI is not the expected amount of reduction in doubt, and there are alternative measures of reduction in uncertainty by which MI is not the expected amount of reduction in uncertainty. We need not discuss this point, however, because the two interpretations are inconsistent with EVI.

[9] This is similar to the problem of measure sensitivity in Bayesian confirmation theory. See (Brossel [2013]; Fitelson [1999]) for helpful discussion.

the only scoring rule that is strictly proper and monotonic. This means that the answers to questions of information given by D_{KL} and MI are definitive whenever the fourth interpretation is appropriate. We will show that this interpretation is appropriate in a wide range of applications with epistemic implications because it is appropriate to measure inaccuracy by a strictly proper monotonic scoring rule in those applications.[10]

2. Formal Analyses of the Three Interpretations

This section examines the three interpretations of MI by reduction in doubt (2.1), by reduction in uncertainty (2.2), and by divergence (2.3). It is shown that they arise, respectively, from three different ways of parsing MI(X; Y) formally.

2.1 Reduction in doubt

The first interpretation of MI relates information to doubt: information reduces doubt. This leads naturally to the suggestion that the amount of information the proposition y provides on the proposition x is the amount of reduction in doubt about x due to y. The reduction is then measured by pointwise mutual information (PMI) as follows:[11]

$$\mathrm{PMI}(x;y) =_{\mathrm{def}} \log\frac{P(x \wedge y)}{P(x)P(y)} = \log\frac{P(x \mid y)}{P(x)} \tag{2}$$

Since PMI measures increase in the probability of x due to y, and increase in the probability is reduction in doubt, PMI can be considered a measure of the information that y provides on x as determined by reduction in doubt about x due to y. MI itself is the weighted average (over $X \times Y$) of PMI:

$$\begin{aligned}\mathrm{MI}(X;Y) &=_{\mathrm{def}} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i \wedge y_j)\log\frac{P(x_i \wedge y_j)}{P(x_i)P(y_j)} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i \wedge y_j)\mathrm{PMI}(x_i;y_j)\end{aligned} \tag{3}$$

---

[10] It is interesting that MI can be naturally interpreted in (at least) four different ways: by reduction in doubt, by reduction in uncertainty, by divergence, and by reduction in expected inaccuracy. We leave it for future investigation whether there is some conceptual reason for this.

[11] PMI(x; y) is formally equivalent to the log-ratio measure of confirmation (Milne [1996]). It is also formally equivalent to Schupbach's measure of coherence (Schupbach [2011]) in the special case where the set of propositions has exactly two members. The latter (Schupbach's measure of coherence) is in turn ordinally equivalent to Shogenji's measure of coherence (Shogenji [1999], [2001]) in the special case where the set of propositions has exactly two members.

So, according to this interpretation, MI actually measures the expected amount of information, while it is PMI defined over an ordered pair of propositions which measures the amount of information that one proposition (y) provides on another proposition (x). Taking PMI instead of MI to be a measure of information makes good sense because we receive information from a proposition, and not from a partition of propositions.

2.2 Reduction in uncertainty

The second interpretation of MI relates information to uncertainty: information reduces uncertainty as to which member of the partition is true. So, the amount of information is the amount of reduction in uncertainty. The most widely used measure of uncertainty is the entropy (H) defined over a partition as follows:

$$\mathrm{H}(X) =_{\mathrm{def}} -\sum_{i=1}^{n} P(x_i)\log P(x_i) \tag{4}$$

The uncertainty (entropy) is minimal at $\mathrm{H}(X) = 0$ when one member of the partition X receives the probability one, while the rest receive the probability zero. The uncertainty (entropy) increases as the probability distribution becomes less lopsided, and it reaches the maximum value at $\mathrm{H}(X) = \log n$ when all members of X receive the same probability $1/n$. Upon learning the truth of the proposition y, the uncertainty of X is measured by the conditional entropy $\mathrm{H}(X \mid y)$ defined as follows:

$$\mathrm{H}(X \mid y) =_{\mathrm{def}} -\sum_{i=1}^{n} P(x_i \mid y)\log P(x_i \mid y) \tag{5}$$

So, we can measure reduction in uncertainty about X due to y by $\mathrm{RH}(X; y)$ as follows:

$$\begin{aligned}\mathrm{RH}(X; y) &=_{\mathrm{def}} \mathrm{H}(X) - \mathrm{H}(X \mid y)\\ &= -\sum_{i=1}^{n} P(x_i)\log P(x_i) - -\sum_{i=1}^{n} P(x_i \mid y)\log P(x_i \mid y)\\ &= \sum_{i=1}^{n} P(x_i \mid y)\log P(x_i \mid y) - \sum_{i=1}^{n} P(x_i)\log P(x_i)\end{aligned} \tag{6}$$

MI itself is the weighted average (over Y) of RH:

$$\begin{aligned}\mathrm{MI}(X;Y) &=_{\mathrm{def}} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i \wedge y_j)\log\frac{P(x_i \wedge y_j)}{P(x_i)P(y_j)}\\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i \wedge y_j)\log\frac{P(x_i \mid y_j)}{P(x_i)}\\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i \wedge y_j)[\log P(x_i \mid y_j) - \log P(x_i)]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{m}\sum_{i=1}^{n} P(x_i \mid y_j)P(y_j)\log P(x_i \mid y_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} P(y_j \mid x_i)P(x_i)\log P(x_i) \\
&= \sum_{j=1}^{m} P(y_j)\sum_{i=1}^{n} P(x_i \mid y_j)\log P(x_i \mid y_j) - \sum_{i=1}^{n} P(x_i)\log P(x_i)\sum_{j=1}^{m} P(y_j \mid x_i) \\
&= \sum_{j=1}^{m} P(y_j)\sum_{i=1}^{n} P(x_i \mid y_j)\log P(x_i \mid y_j) - \sum_{i=1}^{n} P(x_i)\log P(x_i) \\
&= \sum_{j=1}^{m} P(y_j)\sum_{i=1}^{n} P(x_i \mid y_j)\log P(x_i \mid y_j) - \sum_{j=1}^{m} P(y_j)\sum_{i=1}^{n} P(x_i)\log P(x_i) \\
&= \sum_{j=1}^{m} P(y_j)\left[-\sum_{i=1}^{n} P(x_i)\log P(x_i) - -\sum_{i=1}^{n} P(x_i \mid y_j)\log P(x_i \mid y_j)\right] \\
&= \sum_{j=1}^{m} P(y_j)[\mathrm{H}(X) - \mathrm{H}(X \mid y_j)] \\
&= \sum_{j=1}^{m} P(y_j)\mathrm{RH}(X; y_j) \qquad (7)
\end{aligned}
$$

So, according to this interpretation, MI measures the expected amount of information, while it is RH defined over a partition and a proposition which measures the amount of information the proposition (y) provides on the partition (X).

2.3 Divergence

The third interpretation of MI relates information to divergence: information changes the probability distribution. So, the amount of information the proposition y provides on the partition X is measured by the amount of divergence between the two (the prior and the posterior) probability distributions over X due to y. One of the widely used measures of divergence between two probability distributions is Kullback-Leibler Divergence ($\mathrm{D_{KL}}$) defined over two probability distributions P and Q over the partition X as follows:

$$\mathrm{D_{KL}}(P \| Q) =_{\mathrm{def}} \sum_{i=1}^{n} P(x_i)\log\frac{P(x_i)}{Q(x_i)} \qquad (8)$$

When the two probability distributions $P(x_i)$ and $Q(x_i)$ are the posterior and the prior probability distributions, $P(x_i \mid y)$ and $P(x_i)$, respectively, $\mathrm{D_{KL}}(P \| Q)$ can be written as follows:

$$\mathrm{D_{KL}}(X; y) = \sum_{i=1}^{n} P(x_i \mid y)\log\frac{P(x_i \mid y)}{P(x_i)} \qquad (9)$$

MI is then the weighted average (over Y) of $\mathrm{D_{KL}}(X; y)$:

$$\mathrm{MI}(X; Y) =_{\mathrm{def}} \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i \wedge y_j)\log\frac{P(x_i \wedge y_j)}{P(x_i)P(y_j)}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i \mid y_j) P(y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)}$$

$$= \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i \mid y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)}$$

$$= \sum_{j=1}^{m} P(y_j) \mathrm{D}_{\mathrm{KL}}(X; y_j) \qquad (10)$$

So, according to this interpretation, MI measures the expected amount of divergence, while it is D_{KL} defined over two probability distributions $P(x_i \mid y)$ and $P(x_i)$ over X which measures the amount of information the proposition (y) provides on the partition (X).

3. Inconsistency with EVI

Consider the following general setting. You are about to conduct an experiment to test hypotheses $x_1, \ldots, x_n$ on a certain subject. The hypotheses are jointly exhaustive and pairwise incompatible, so that $X = \{x_1, \ldots, x_n\}$ is a partition. Ideally the experiment will settle the issue once and for all by eliminating all hypotheses except one, but that is highly unlikely with a single experiment. The best realistic scenario is that the experiment raises the probability of one hypothesis close to one while lowering the probabilities of the other hypotheses close to zero. It is conceivable that the experiment will actually make the probabilities of all hypotheses far from one. That depends on the outcome of the experiment, which we take to be the partition $Y = \{y_1, \ldots, y_m\}$. Different outcomes of the experiment can affect the probability distribution over X differently, and we compare the epistemic value of the outcome on one hand, and the amount of information as measured by reduction in doubt (PMI) and by reduction in uncertainty (RH) on the other.[12] The purpose is to show that the first two interpretations of MI are inconsistent with EVI.

In some cases there is no conflict between the two interpretations and EVI. Suppose, for example, $X = \{x_1, \ldots, x_5\}$ and the initial probability distribution is $P(x_1) = 0.6$ and $P(x_2) = P(x_3) = P(x_4) = P(x_5) = 0.1$. Suppose further that the experimental outcome y_1 would not change the probability distribution over X at all, that is, $P(x_1 \mid y_1) = 0.6$ while $P(x_2 \mid y_1) = P(x_3 \mid y_1) = P(x_4 \mid y_1) = P(x_5 \mid y_1) = 0.1$. Meanwhile, the outcome y_2 would raise the probability of x_1 to $P(x_1 \mid y_2) = 0.9$ while lowering the probabilities of the other hypotheses to $P(x_2 \mid y_2) = P(x_3 \mid y_2) = P(x_4 \mid y_2) = P(x_5 \mid y_2) = 0.025$. The second outcome has considerable epistemic value. It is certainly more valuable, other things being equal, than the first outcome that does not change the probability

[12] See (Crupi and Tentori [2014]) for a discussion of different ways one might measure the epistemic utility of an experiment.

distribution over X at all.[13] The two interpretations of MI are not in conflict with this because $\mathrm{PMI}(x_1; y_1) = \mathrm{RH}(X; y_1) = 0$, while both $\mathrm{PMI}(x_1; y_2)$ and $\mathrm{RH}(X; y_2)$ are positive.

There are, however, many cases where the two interpretations of MI run counter to EVI. Suppose the experimental outcome y_3 would reveal that x_1 is no more probable than its competition, so that $P(x_1 \mid y_3) = P(x_2 \mid y_3) = P(x_3 \mid y_3) = P(x_4 \mid y_3) = P(x_5 \mid y_3) = 0.2$. In light of this outcome we would no longer regard x_1 as the leading hypothesis on the subject. The outcome y_3 would have considerable epistemic value. It would certainly be more valuable than the outcome y_1 that would not change the probability distribution at all. The trouble with the two interpretations is that both $\mathrm{PMI}(x_1; y_3)$ and $\mathrm{RH}(X; y_3)$ are negative. $\mathrm{PMI}(x_1; y_3)$ is negative because $P(x_1 \mid y_3) < P(x_1)$ and hence x_1 is more doubtful in light of y_3 than before. $\mathrm{RH}(X; y_3)$ is negative because y_3 raises the uncertainty of X to the maximum by rendering all its members equi-probable. This means that if we use either PMI or RH to measure the amount of information, then the outcome y_3 that would undercut the leading hypothesis is epistemically less valuable than the outcome y_1 that would not change the probability distribution over X at all.[14] This is untenable. Other things being equal, no sensible journal editor, for example, would find y_3 (a paper featuring y_3) less worthy of publication than y_1 (a paper featuring y_1).

It is also worth noting that epistemic value does not accrue solely from change in the 'pattern' of the probability distribution—getting more lopsided, closer to even, etc. A finding that keeps the pattern of the probability distribution exactly the same can be highly informative and epistemically valuable. Suppose the outcome y_4 would reverse the probabilities of x_1 and x_2 while the rest of the probability distribution would remain the same, so that $P(x_2 \mid y_4) = 0.6$ while $P(x_1 \mid y_4) = P(x_3 \mid y_4) = P(x_4 \mid y_4) = P(x_5 \mid y_4) = 0.1$. This means that x_1 is replaced by x_2 as the leading hypothesis on the subject. Clearly, the outcome y_4 is of great epistemic value, though the pattern of the probability distribution remains the same: one of the five hypotheses receives the probability 0.6 while the rest receive the probability 0.1 each. Meanwhile, the degree of uncertainty depends only on the pattern of the probability distribution. As a result, $\mathrm{RH}(X; y_4)$ is zero: y_4 neither increases nor decreases the degree of uncertainty about X. Since a finding like y_4 that replaces the leading hypothesis is of great epistemic value, we cannot measure the epistemic value of a finding by reduction in uncertainty.

To clarify our point, we do not deny that there are applications of MI in which EVI is absent or unimportant. The two interpretations of MI by reduction in doubt and by reduction in uncertainty may be appropriate in those applications. There are, however, many applications of

---

[13] The first outcome may still be epistemically significant to some extent. In some cases where the evidence does not change the probability distribution over X, the evidence makes the probability distribution more stable in the face of potential future data (Joyce [2005]). However, it is part of our assumption ('other things being equal') that the different experimental outcomes do not affect the stability of the probability distribution differently.

[14] PMI yields the right result with respect to x_2, x_3, x_4, and x_5 in that each of $\mathrm{PMI}(x_2; y_3)$, $\mathrm{PMI}(x_3; y_3)$, $\mathrm{PMI}(x_4; y_3)$, and $\mathrm{PMI}(x_5; y_3)$ is positive whereas each of $\mathrm{PMI}(x_2; y_1)$, $\mathrm{PMI}(x_3; y_1)$, $\mathrm{PMI}(x_4; y_1)$, and $\mathrm{PMI}(x_5; y_1)$ equals 0. Our point is that PMI yields the wrong result with respect to x_1.

MI—in communication theory, learning theory, economics, etc.—where EVI is important. Since the two interpretations of MI are inconsistent with EVI, they are inappropriate in those applications.

The point that reduction in uncertainty is inadequate as a measure of epistemic value is not new. For example, Evans and Over ([1996]) raise it against Oaksford and Chater's analysis of the Wason selection task (Oaksford and Chater [1994]). The task is commonly regarded as an indication of a shortcoming in human reasoning because a majority of participants select an observation (an experiment) that is apparently useless for answering the given question. Oaksford and Chater argue that the apparently useless selection is actually rational because the expected information gain from the observation is high under plausible assumptions. The trouble is that Oaksford and Chater interpret information gain in terms of reduction in uncertainty. Evans and Over raise the point (with a case similar to those used above) that we cannot measure the epistemic utility of an observation by reduction in uncertainty.

Interestingly, Oaksford and Chater ([1996]) reply to Evans and Over by dropping the interpretation of MI by reduction in uncertainty in favor of the interpretation of MI by divergence. This interpretation is the focus of the next section.

4. Problem of Measure Sensitivity

The discussion in Section 3 indicates that any finding that changes the probability distribution is informative and epistemically valuable even if it casts doubt on the leading hypothesis and increases uncertainty. Once we recognize this point, it makes sense to measure the amount of information by divergence between the old and the updated probability distributions, but the divergence interpretation has its own problem.

There are many formal measures of divergence that are adequate in the sense of meeting the basic constraint that the degree of divergence between P and Q should be zero (the minimum degree) in cases where P and Q are identical to each other and should be positive in all other cases. D_{KL} meets this constraint, but so do many alternative measures of divergence.[15] Consider, for example, Rectilinear Divergence (D_{RL}) and Squared Euclidean Divergence (D_{SE}):

$$D_{RL}(P \| Q) =_{def} \sum_{i=1}^{n} \left| P(x_i) - Q(x_i) \right| \tag{11}$$

$$D_{SE}(P \| Q) =_{def} \sum_{i=1}^{n} \left[P(x_i) - Q(x_i) \right]^2 \tag{12}$$

[15] See (Cha [2007]) for a comprehensive survey of divergence measures. We are using the term 'divergence' broadly so that the class of divergence measures includes the class of so-called 'distance/similarity' measures.

The two measures, D_{RL} and D_{SE}, are not ordinally equivalent to each other and neither of them is ordinally equivalent to D_{KL}. Yet each of them, as with D_{KL}, meets the basic constraint that the degree of divergence between *P* and *Q* should be zero in cases where *P* and *Q* are identical to each other and should be positive in all other cases. Indeed both of them are sensible measures of divergence, and have been in actual use for measuring divergence.

This is problematic for the divergence interpretation of MI because we can construct alternative formulas for mutual information as expected divergence from alternative measures of divergence. First, let *P* be $P(x_i \mid y)$ and *Q* be $P(x_i)$, and rewrite $D_{RL}(P \parallel Q)$ and $D_{SE}(P \parallel Q)$ accordingly:

$$D_{RL}(X; y) =_{def} \sum_{i=1}^{n} \left| P(x_i \mid y) - P(x_i) \right| \tag{13}$$

$$D_{SE}(X; y) =_{def} \sum_{i=1}^{n} \left[P(x_i \mid y_j) - P(x_i) \right]^2 \tag{14}$$

Next, take the weighted averages (over *Y*) of $D_{RL}(X; y)$ and $D_{SE}(X; y)$:

$$D^*_{RL}(X; Y) =_{def} \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} \left| P(x_i \mid y_j) - P(x_i) \right| \tag{15}$$

$$D^*_{SE}(X; Y) =_{def} \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} \left[P(x_i \mid y_j) - P(x_i) \right]^2 \tag{16}$$

The two measures of expected divergence, D^*_{RL} and D^*_{SE}, are not ordinally equivalent to each other, and more importantly, neither of them is ordinally equivalent to MI. (See Appendix A.1 for proof.) Since D_{RL} and D_{SE} are sensible measures of divergence, D^*_{RL} and D^*_{SE} are sensible measures of expected divergence. If information is to be measured by divergence, as the third interpretation of MI suggests, then D^*_{RL} and D^*_{SE} are sensible alternatives to MI.

This is problematic because the availability of sensible alternative measures makes the answers to many questions of information 'measure sensitive'. Suppose we are comparing the epistemic values of two experiments, E_1 and E_2, by MI as understood on the divergence interpretation. It is possible that E_1 has a greater epistemic value than E_2 by measure MI, but not by measure D^*_{RL} or by measure D^*_{SE}. Consequently, the answer given by MI cannot be considered definitive.

The problem of measure sensitivity arises not only for evaluation in particular cases but also for general principles of information. Many important principles of information theory provable on MI fail to hold on D^*_{RL} or D^*_{SE}. Two such principles are:

> **Symmetry (S):** Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ be partitions. Then $MI(X; Y) = MI(Y; X)$.

Data-Processing Inequality (DPI): Let $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_m\}$, and $Z = \{z_1, \ldots, z_l\}$ be partitions, and suppose Y screens off Z from X in that $P(x_i \mid y_j \wedge z_k) = P(x_i \mid y_j)$ for any $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $k = 1, \ldots, l$. Then (a) $\mathrm{MI}(X; Y) \geq \mathrm{MI}(X; Z)$ and (b) $\mathrm{MI}(Y; Z) \geq \mathrm{MI}(X; Z)$.

These principles fail to hold if we replace MI with $\mathrm{D}^*_{\mathrm{SE}}$. (See Appendices A.2 and A.3 for proof.) DPI is especially important in that it is used in many areas of research beyond information theory, for example, philosophy of science (see Barrett and Sober [1992]; Sober [2008]; Sober and Barrett [1992]; Sober and Steel [2002], [2011], [2014]).[16]

There are, in general, two possible responses to the problem of measure sensitivity. One is to eliminate all but one measure by additional constraints. In the present case this would mean that we seek plausible additional constraints on an adequate measure of divergence beyond the basic one, and show that all but D_{KL} fail to meet these constraints. This is not a promising approach in the case of D_{KL}. If anything, D_{RL} and D_{SE} look more natural than D_{KL} as measures of divergence. For example, both D_{RL} and D_{SE} are symmetric while D_{KL} is not.[17] In other words, it is possible by measure D_{KL} that P diverges from Q more than Q diverges from P. This is peculiar for a measure of divergence. Also, D_{KL} is not 'uniform' in that not all differences between $P(x_i)$ and $Q(x_i)$ increase the degree to which P diverges from Q. Let $X = \{x_1, x_2\}$ be a partition, and $P(x_1) = 0.8$ while $Q(x_1) = 0.2$. Then, $\mathrm{D}_{\mathrm{KL}}(P \parallel Q)$ is given by:

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(P \parallel Q) &= P(x_1)\log\frac{P(x_1)}{Q(x_1)} + P(x_2)\log\frac{P(x_2)}{Q(x_2)} \\ &= (0.8)\log\frac{0.8}{0.2} + (0.2)\log\frac{0.2}{0.8} \end{aligned} \tag{17}$$

Since $\log x$ is negative for $0 < x < 1$, the second addend in (17) is negative. This means that by D_{KL} the difference between $P(x_2)$ and $Q(x_2)$ decreases the degree to which P diverges from Q. This is not a feature we naturally expect from a measure of divergence, for the difference between $P(x_2)$ and $Q(x_2)$ is a respect in which P diverges from Q.

In raising these points, we are not questioning the use of D_{KL} as a measure of divergence. D_{KL} still meets the basic constraint. $\mathrm{D}_{\mathrm{KL}}(P \parallel Q)$ in (17) is positive overall despite the negative

[16] See, however, (Roche and Shogenji [2014]) for the result that a counterpart of DPI in confirmation theory ('dwindling confirmation') is not measure sensitive under a plausible constraint on an adequate measure of confirmation.

[17] As noted earlier, MI, which is the expected D_{KL} divergence, is symmetric, but D_{KL} itself is not symmetric. It is also worth noting that originally Kullback and Leibler ([1951]) do not put forward $\mathrm{D}_{\mathrm{KL}}(P \parallel Q)$ as a divergence measure. They put it forward as an information measure. Their divergence measure is $\mathrm{D}_{\mathrm{KL}}(P \parallel Q) + \mathrm{D}_{\mathrm{KL}}(Q \parallel P)$. The latter, which is due to Jeffreys ([1946], [1948]), is symmetric.

second addend because this addend is more than offset by the positive first addend with a greater weight $P(x_1)$. Besides, the expected value of D_{KL}, which is MI, has some nice features—such as symmetry, DPI, and additivity—that make D_{KL} attractive.[18] Our point here is that some features of D_{KL} make it hard to argue that D_{KL} is the only adequate measure of divergence, and that because of this the answers to questions of information given by MI cannot be considered definitive.

We turn now to our preferred interpretation of MI.

5. Reduction in Expected Inaccuracy

This section describes a fourth interpretation of MI by reduction in expected inaccuracy. We begin with the notion of inaccuracy. If we assign any probability other than one to a true proposition, the assignment is inaccurate. The degree of inaccuracy is inversely related to the probability assigned because the higher is the probability, the closer it is to the most accurate value, which is one. Obviously, we can eliminate all inaccuracy by correctly guessing which member of the partition is true and assigning it the probability one. But it is unwise to simply guess which member of the partition is true. For example, when the partition has three or more members and they are equally probable, it is more likely that we guess incorrectly and end up assigning the probability one to a false member and the probability zero to the true member. The best we can do on the basis of available evidence is to minimize the *expected* inaccuracy. The fourth interpretation of MI proposes that we measure the amount of information the proposition y provides on the partition X by reduction in the expected inaccuracy of the probability distribution over X due to y. MI, which is defined over the partitions X and Y, is then taken to be the expected amount (the weighted average over Y) of information a member of Y provides on X. In other words, MI measures the expected reduction in the expected inaccuracy of the probability distribution. In the remainder of this section we present a formal analysis of MI in support of the fourth interpretation.

We begin with a measure of inaccuracy. Let P be a probability distribution over X. The degree of inaccuracy of P depends on which member of X is true. A scoring rule $SR(P; i)$ is a function that determines the degree of inaccuracy of P when the i-th member of X is true. There are many scoring rules (see Murphy and Winkler [1984]; Winkler [1967], [1969], [1971], [1994]; Winkler and Murphy [1968]) but MI is grounded in one particular rule called the logarithmic scoring rule (SR_L):

$$\mathrm{SR}_{\mathrm{L}}(P;i) =_{\mathrm{def}} \log\frac{1}{P(x_i)} = -\log P(x_i) \tag{18}$$

[18] MI is additive in the sense that $\mathrm{MI}(X;\ Y \times Z) = \mathrm{MI}(X\ ;\ Z) + \mathrm{MI}(X;\ Y \mid Z)$.

SR_L is a decreasing function of $P(x_i)$, which makes good sense. When the probability $P(x_i)$ that is assigned to the true member x_i of X is higher, the degree of inaccuracy is lower. On the logarithmic scoring rule (SR_L) the expected inaccuracy (EI_L) of the probability distribution over X is as follows:

$$\mathrm{EI}_{\mathrm{L}}(X) =_{\mathrm{def}} \sum_{i=1}^{n} P(x_i)\mathrm{SR}_{\mathrm{L}}(P;i) = -\sum_{i=1}^{n} P(x_i)\log P(x_i) \tag{19}$$

When we update the probability distribution over X upon learning y, the expected inaccuracy is also updated:

$$\mathrm{EI}_{\mathrm{L}}(X \mid y) = -\sum_{i=1}^{n} P(x_i \mid y)\log P(x_i \mid y) \tag{20}$$

$EI_L(X)$ and $EI_L(X \mid y)$ are identical, respectively, to $H(X)$ and $H(X \mid y)$ discussed earlier in Subsection 2.2, but we now regard them as measures of expected inaccuracy.

Reinterpretation would be pointless if we measured reduction in expected inaccuracy by $EI_L(X \mid y) - EI_L(X)$, as we measured reduction in uncertainty by $RH(X; y) = H(X \mid y) - H(X)$ in Subsection 2.2. We pointed out in Section 3 that a finding with great epistemic value often increases uncertainty, and the negative value of $RH(X; y)$ in such a case is inconsistent with EVI which is assumed in many applications of MI. Since $EI_L(X \mid y) - EI_L(X)$ is mathematically no different from $RH(X; y) = H(X \mid y) - H(X)$, it takes a negative value when a finding with great epistemic value increases uncertainty, which is inconsistent with EVI. However, the fourth interpretation does not use $EI_L(X \mid y) - EI_L(X)$ to measure reduction in expected inaccuracy. That is the difference in substance from the uncertainty interpretation.

The reason for not using $EI_L(X \mid y) - EI_L(X)$ is that once y is given, $EI_L(X)$ is no longer an appropriate measure of the expected inaccuracy of the old probability distribution. The appropriate measure of the expected inaccuracy of the old probability distribution is given by:

$$\mathrm{EI}_{\mathrm{L}}(X; y) = -\sum_{i=1}^{n} P(x_i \mid y)\log P(x_i) \tag{21}$$

If x_i turns out to be true, the degree of inaccuracy is $-\log P(x_i)$ instead of $-\log P(x_i \mid y)$ because $EI_L(X; y)$ measures the expected inaccuracy of the old probability distribution. However, the weight assigned to $-\log P(x_i)$ for calculating the weighted average is $P(x_i \mid y)$ instead of $P(x_i)$. This is because we have already updated the probability to $P(x_i \mid y)$ in light of y. $EI_L(X; y)$ measures the expected inaccuracy of the old distribution in retrospect after learning y.

We measure reduction in expected inaccuracy (REI_L) due to y by comparing the expected inaccuracy $EI_L(X; y)$ of the old distribution as evaluated in retrospect after learning y, and the expected inaccuracy $EI_L(X \mid y)$ of the new probability distribution:

$$\mathrm{REI}_{\mathrm{L}}(X; y) =_{\mathrm{def}} \mathrm{EI}_{\mathrm{L}}(X; y) - \mathrm{EI}_{\mathrm{L}}(X \mid y)$$

$$= -\sum_{i=1}^{n} P(x_i \mid y) \log P(x_i) - -\sum_{i=1}^{n} P(x_i \mid y) \log P(x_i \mid y)$$
$$= \sum_{i=1}^{n} P(x_i \mid y) \log P(x_i \mid y) - \sum_{i=1}^{n} P(x_i \mid y) \log P(x_i) \quad (22)$$

Note that $\mathrm{REI}_{\mathrm{L}}(X; y)$ is formally equivalent to D_{KL}:

$$\mathrm{REI}_{\mathrm{L}}(X; y) = \sum_{i=1}^{n} P(x_i \mid y) \log P(x_i \mid y) - \sum_{i=1}^{n} P(x_i \mid y) \log P(x_i)$$
$$= \sum_{i=1}^{n} P(x_i \mid y)[\log P(x_i \mid y) - \log P(x_i)]$$
$$= \sum_{i=1}^{n} P(x_i \mid y) \log \frac{P(x_i \mid y)}{P(x_i)}$$
$$= \mathrm{D}_{\mathrm{KL}}(X; y) \quad (23)$$

Recall that D_{KL} is a measure of divergence that takes the value zero when the two distributions are identical, and is positive in all other cases. So is $\mathrm{REI}_{\mathrm{L}}$. There is no reduction in the expected inaccuracy when y does not change the probability distribution over X, but there is reduction in the expected inaccuracy if y changes the probability distribution over X regardless of the direction of change. So, updating the probability distribution in light of a new finding always reduces the expected inaccuracy and is epistemically valuable.

We noted in Section 4 that D_{KL} has some peculiar features. First, D_{KL} is not symmetric; it is possible by measure D_{KL} that P diverges from Q more than Q diverges from P. Second, D_{KL} is not uniform; not all differences between $P(x_i)$ and $Q(x_i)$ increase the degree to which P diverges from Q on D_{KL}. $\mathrm{REI}_{\mathrm{L}}(X; y)$ is formally equivalent to D_{KL} and thus has those same features, but here they are not peculiar. First, there is no obvious reason why a measure of reduction in expected inaccuracy (as opposed to a measure of divergence) should be symmetric. Second, when $P(x_i \mid y) < P(x_i)$, y increases inaccuracy if x_i is true. It makes sense, then, that $P(x_i \mid y) \log \frac{P(x_i \mid y)}{P(x_i)}$ is negative when $P(x_i \mid y) < P(x_i)$ and x_i is true. Not all differences between $P(x_i \mid y)$ and $P(x_i)$ serve to increase the reduction in expected inaccuracy.

To complete the fourth interpretation by reduction in expected inaccuracy, MI is the weighted average (over Y) of the reduction in the expected inaccuracy of the probability distribution over the partition X due to a member of the partition Y:

$$\mathrm{MI}(X; Y) =_{\mathrm{def}} \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i \wedge y_j) \log \frac{P(x_i \wedge y_j)}{P(x_i) P(y_j)}$$
$$= \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i \mid y_j) \log \frac{P(x_i \mid y_j)}{P(x_i)}$$
$$= \sum_{j=1}^{m} P(y_j) \sum_{i=1}^{n} P(x_i \mid y_j)[\log P(x_i \mid y_j) - \log P(x_i)]$$

$$= \sum_{j=1}^{m} P(y_j)[\sum_{i=1}^{n} P(x_i \mid y_j) \log P(x_i \mid y_j) - \sum_{i=1}^{n} P(x_i \mid y_j) \log P(x_i)]$$
$$= \sum_{j=1}^{m} P(y_j)[\mathrm{EI}_{\mathrm{L}}(X; y_j) - \mathrm{EI}_{\mathrm{L}}(X \mid y_j)]$$
$$= \sum_{j=1}^{m} P(y_j)\mathrm{REI}_{\mathrm{L}}(X; y_j) \qquad (24)$$

So, MI measures the expected amount of information, while it is $\mathrm{REI}_{\mathrm{L}}$ defined over a proposition and a partition which measures the amount of information the proposition *y* provides on the partition *X*, where the amount of information is measured by the reduction in expected inaccuracy.[19]

6. Resolution of the Problem of Measure Sensitivity

Recall our point in Section 4 that the interpretation of MI by divergence gives rise to the problem of measure sensitivity: there are many sensible measures of divergence other than D_{KL}, and if we start from a different measure, we obtain a different measure of expected divergence than MI. As a result, the answers to questions of information (as divergence) given by MI cannot be considered definitive. We proposed the fourth interpretation of MI in response to this problem: D_{KL} measures more specifically reduction in expected inaccuracy. Our response is not complete, though, for there are many measures of inaccuracy other than the logarithmic scoring rule (SR_{L}), from which D_{KL} is derived, and if we start from a different measure, we obtain a different measure of expected information—expected reduction in expected inaccuracy—than MI. In this section we complete our response by specifying the sense of inaccuracy that is uniquely captured by SR_{L}.

6.1 Alternative measures of inaccuracy

Scoring rules are measures of inaccuracy for a probability distribution *P* over a partition of propositions. One common feature of scoring rules is that they are exclusively 'truth directed' in the sense that they measure the inaccuracy of *P* solely by the truth and falsity of the propositions in the partition. So, once we identify the true member x_i of the partition, we can calculate the degree of inaccuracy $SR(P; i)$ for any *P*. Given their exclusive truth-directedness, a simple and sensible approach is to measure the inaccuracy of *P* by two probabilities: $P(x_i)$ assigned to the true member, and $1 - P(x_i)$ assigned to the rest of the partition. Since $1 - P(x_i)$ is a function of $P(x_i)$, this means that the inaccuracy of *P* is determined by $P(x_i)$ alone. Scoring rules of this kind

[19] See (Crupi and Tentori [2014]); Grunwald and Dawid [2004]; Winkler and Murphy [1968]) among others for observations of the connection between scoring rules and measures of information.

are called 'local scoring rules'.[20] Since inaccuracy decreases as $P(x_i)$ gets closer to one, an adequate local scoring rule should be a monotonic function of $P(x_i)$:

> **Monotonicity (M)**: Let *SR* be a scoring rule. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a partition and P and Q be probability distributions over X. Suppose x_i is true and $P(x_i) > Q(x_i)$. Then $SR(P; i) < SR(Q; i)$.

This requirement is equivalent to the requirement that an adequate scoring rule should be local, on the assumption that $SR(P; i)$ is exclusively truth-directed and thus is a decreasing function of $P(x_i)$ and an increasing function of $P(x_j)$ for any $j \neq i$ (see Fallis [2007] for proof).

The logarithmic scoring rule $\mathrm{SR_L}(P; i) = -\log P(x_i)$, from which $\mathrm{D_{KL}}$ and MI are constructed, meets M. There are, however, many other functions of $P(x_i)$ such as $\mathrm{SR_C}$, $\mathrm{SR_{SC}}$, and $\mathrm{SR_R}$ below that meet M:[21]

$$\mathrm{SR_C}(P;i) =_{\mathrm{def}} 1 - P(x_i) \tag{25}$$

$$\mathrm{SR_{SC}}(P;i) =_{\mathrm{def}} \left[1 - P(x_i)\right]^2 \tag{26}$$

$$\mathrm{SR_R}(P;i) =_{\mathrm{def}} \frac{1}{P(x_i)} \tag{27}$$

If we start from any of these alternative measures of inaccuracy, we obtain a measure of reduction in expected inaccuracy different than $\mathrm{D_{KL}}(X; y)$.

We can construct measure $REI(X; y)$ of reduction in expected inaccuracy from any scoring rule $SR(P; i)$ in the same way we derived $\mathrm{D_{KL}}(X; y)$ from $\mathrm{SR_L}$ in Section 5. First, after updating the probability distribution from P to P_y, the inaccuracy of the updated probability distribution P_y is $SR(P_y; i)$, where $P_y(x) = P(x \mid y)$. The expected inaccuracy $EI(X \mid y)$ of P_y is then the weighted average of $SR(P_y; i)$, where the weights are the probabilities $P(x_i \mid y)$:

$$EI(X \mid y) = \sum_{i=1}^{n} P(x_i \mid y) SR(P_y; i) \tag{28}$$

Meanwhile, the expected inaccuracy $EI(X; y)$ of the old probability distribution P (as evaluated in retrospect after learning y) is the weighted average of $SR(P; i)$, where the weights are the updated probabilities $P(x_i \mid y)$:

---

[20] There are also 'global scoring rules' on which inaccuracy is not determined by $P(x_i)$ alone. We will discuss global scoring rules in the next section.

[21] The subscripts in '$\mathrm{SR_C}$', '$\mathrm{SR_{SC}}$', and '$\mathrm{SR_R}$' stand for 'Complement', 'Squared Complement', and 'Reciprocal'.

$$EI(X; y) = \sum_{i=1}^{n} P(x_i \mid y) SR(P; i) \tag{29}$$

Putting (28) and (29) together, we can measure the reduction in the expected inaccuracy of the probability distribution due to *y* as follows:

$$\begin{aligned} REI(X; y) &= \sum_{i=1}^{n} P(x_i \mid y) SR(P; i) - \sum_{i=1}^{n} P(x_i \mid y) SR(P_y; i) \\ &= \sum_{i=1}^{n} P(x_i \mid y)[SR(P; i) - SR(P_y; i)] \end{aligned} \tag{30}$$

Since we are measuring information by reduction in the expected inaccuracy and a scoring rule is a measure of inaccuracy, we can plug in any scoring rule *SR* into (30) to obtain a measure of information. We obtained D_{KL} in Section 5 by starting from the logarithmic scoring rule SR_L.

Finally, the expected amount of information $REI(X; Y)$ that a member of *Y* provides on *X* is the weighted average of $REI(X; y)$ over *Y*, just as $MI(X; Y)$ is the weighted average of $D_{KL}(X; y)$ over *Y*. Some of these alternative measures of expected information $REI(X; Y)$ constructed from different decreasing functions of $P(x_i)$ are clearly inadequate. For example, $REI_R(X; Y)$ constructed from $SR_R(P; i) = 1/P(x_i)$ is a constant function: $REI_R(X; Y) = 0$ for any *X* and *Y*, and for any probability distribution defined over *X* and *Y*. So, $REI_R(X; Y)$ should be rejected since a constant function does not measure anything. There are, however, many alternative measures that are not so obviously inadequate, including $REI_C(X; Y)$ and $REI_{SC}(X; Y)$ constructed from $SR_C(P; i) = 1 - P(x_i)$ and $SR_{SC}(P; i) = [1 - P(x_i)]^2$, respectively. Our challenge is to specify the distinctive sense of inaccuracy captured by SR_L that points to the type of applications in which the answers given by D_{KL} and MI are definitive.

6.2 Resolution by strict propriety

An answer to our challenge is found in the extant literature on scoring rules. It is known that the logarithmic scoring rule SR_L is the only local scoring rule that meets the following constraint (see Bernardo 1970 for proof).[22]

> **Strict Propriety (SP)**: Let *SR* be a scoring rule. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a partition. Let *P* and *Q* be probability distributions over *X*. Then $\sum_{i=1}^{n} P(x_i) SR(P; i) \leq \sum_{i=1}^{n} P(x_i) SR(Q; i)$ where the two are equal to each other if and only if $P(x_i) = Q(x_i)$ for each *i*.

SR_L is thus the only strictly proper monotonic scoring rule. Consequently, the answers to questions of information (as reduction in expected inaccuracy) given by D_{KL} and MI are definitive when the application calls for a strictly proper monotonic scoring rule as a measure of

[22] There is a technical requirement that a scoring rule is a 'smooth' function; all scoring rules we are considering are smooth functions.

inaccuracy. In the remainder of this subsection we examine the role of SP in the measurement of inaccuracy.

SP is widely considered a condition of adequacy for a scoring rule, and there is a compelling reason for that: a strictly proper scoring rule gives forecasters an incentive to announce, honestly, the probability distribution they believe in. The importance of the incentive is easy to see in cases where a scoring rule is in violation of SP. Take $\mathrm{SR_C}$ for example. Let $X = \{x_1, x_2\}$ be a partition, and suppose your personal probabilities are $P(x_1) = 0.8$ and $P(x_2) = 0.2$. If you honestly announce your personal probabilities as your forecast, then the expected inaccuracy of the announced probabilities, as measured by $\mathrm{SR_C}$, is:

$$\begin{aligned}\sum_{i=1}^{2} P(x_i)\mathrm{SR_C}(P;i) &= P(x_1)\big(1-P(x_1)\big)+P(x_2)\big(1-P(x_2)\big)\\ &= 0.8(1-0.8)+0.2(1-0.2)\\ &= 0.32\end{aligned} \tag{31}$$

However, by announcing the probabilities $Q(x_1) = 1$ and $Q(x_2) = 0$, instead of $P(x_1) = 0.8$ and $P(x_2) = 0.2$, you can reduce the expected inaccuracy, as measured by $\mathrm{SR_C}$:

$$\begin{aligned}\sum_{i=1}^{2} P(x_i)\mathrm{SR_C}(Q;i) &= P(x_1)\big(1-Q(x_1)\big)+P(x_2)\big(1-Q(x_2)\big)\\ &= 0.8(1-1)+0.2(1-0)\\ &= 0.2\end{aligned} \tag{32}$$

As a result, $\mathrm{SR_C}$ encourages forecasters to announce, dishonestly, the probability distribution Q instead of the personal probability distribution P. This occurs because $\mathrm{SR_C}$ is in violation of SP: there is some probability distribution Q whose expected inaccuracy $\sum_{i=1}^{n} P(x_i)SR_C(Q;i)$ is smaller than $\sum_{i=1}^{n} P(x_i)SR_C(P;i)$.

In contrast, when a scoring rule meets SP, the expected inaccuracy of one's own probability distribution P (where P gives the weights for averaging) is smaller than the expected inaccuracy of any other probability distribution Q (where P still gives the weights for averaging because P is one's own probability distribution). This gives forecasters an incentive to be always honest and announce the probability distribution they believe in.

An incentive to honesty is a compelling reason for adopting a strictly proper scoring rule in many contexts, but there is also an epistemic reason for adopting SP that is directly relevant to our present concern.[23] Recall the observation that prompted our search for a new interpretation of MI. The doubt and uncertainty interpretations of MI are inconsistent with EVI which is assumed in many applications of MI, because any finding that changes the probability distribution is

[23] We thank an anonymous referee of this journal for pointing this out. See (Joyce [2009]) for further discussion of SP and for further references.

informative and epistemically valuable even if it does not reduce doubt or uncertainty. It turns out that a scoring rule that is in violation of SP is inconsistent with EVI for the same reason. Suppose finding y changes the probability distribution over X from P to P_y. This means that y is informative and epistemically valuable. So, if we are to measure the amount of information by reduction in expected inaccuracy, then reduction in expected inaccuracy $REI(X; y)$ should be positive whenever P_y is different from P. However, that is not true if the scoring rule is not strictly proper. If the scoring rule is not strictly proper, there are cases in which the expected inaccuracy of the updated distribution $EI(X \mid y) = \sum_{i=1}^{n} P(x_i \mid y) SR(P_y; i)$ is not smaller than the expected inaccuracy of the old distribution $EI(X; y) = \sum_{i=1}^{n} P(x_i \mid y) SR(P; i)$. In other words, if the scoring rule is not strictly proper, some finding that changes the probability distribution is judged uninformative and devoid of epistemic value. So, in applications in which EVI is assumed, we should use a strictly proper scoring rule in constructing a measure of information.

6.3 Range of applications

Our resolution of the problem of measure sensitivity also reveals that the use of MI is inappropriate in some applications. Scoring rules are exclusively truth directed in the sense that they measure the inaccuracy of P solely by the truth and falsity of the propositions in the partition. The use of a scoring rule is therefore inappropriate if we are interested in a kind of inaccuracy that goes beyond the truth and falsity of the propositions.

Here is one example. Let $X = \{x_1, x_2, x_3, x_4\}$ be a partition of propositions, where x_1, x_2, and x_3 are that JFK died in 1963, 1962, 1961, respectively, while x_4 is the 'catch-all' hypothesis that JFK died in some other year. Suppose $P(x_3) = 0.7$ and $P(x_1) = P(x_2) = P(x_4) = 0.1$ initially, but the probability distribution is updated in light of y to $P(x_2 \mid y) = 0.7$ and $P(x_1 \mid y) = P(x_3 \mid y) = P(x_4 \mid y) = 0.1$. In other words, the distribution remains the same except that x_2 now receives the only high probability 0.7 instead of x_3. This change makes no difference for the purpose of an exclusively truth-directed evaluation of inaccuracy because x_2 and x_3 are both false. However, there is a sense in which (given that JFK actually died in 1963) the updated distribution is less inaccurate than the initial distribution because the year 1962 that is falsely asserted in proposition x_2 with the updated probability of $P(x_2 \mid y) = 0.7$ is closer to the true year of JFK's death than is the year 1961 that is falsely asserted in proposition x_3 with the initial probability of $P(x_3) = 0.7$. If the different proximities to the truth among the false members of the partition are important, then it is not appropriate to use a scoring rule because a scoring rule is an exclusively truth-directed measure of inaccuracy.

This means that even in applications where EVI is important and thus it is appropriate to measure information by reduction in expected inaccuracy, we should still avoid using SR_L—or any scoring rule—if the kind of inaccuracy of interest goes beyond the truth and falsity of the propositions. Consequently, the use of D_{KL} and MI is also inappropriate since they are constructed from a scoring rule that only takes into account the truth and falsity of the propositions.

7. Global Scoring Rules

In this section we take up a class of scoring rules that are not monotonic. Let us assume that EVI is important, and that the kind of inaccuracy of interest does not go beyond the truth and falsity of the propositions over which the probabilities are distributed. This means that it is appropriate to use a scoring rule, and the scoring rule should be strictly proper. SR_L is the only scoring rule that is both strictly proper and monotonic but it is not the only scoring rule that is strictly proper. Some scoring rules are strictly proper but not monotonic. Here we have in mind 'global scoring rules'. Although monotonicity is a sensible condition, there may be some applications where the use of a global scoring rule is not problematic and thus SR_L—and D_{KL} and MI constructed from SR_L—is not the only measure that is appropriate. We now examine the extent of such cases.

To understand the idea of global scoring rules it is helpful to compare them with their respective local counterparts. Take $SR_{SC}(P; i) = [1 - P(x_i)]^2$ from Section 6.1 which is a local scoring rule. Since it is a local scoring rule, the degree of inaccuracy $SR_{SC}(P; i)$ decreases as the probability of the true member $P(x_i)$ gets closer to one. According to its global counterpart $SR_B(P; i)$ the degree of inaccuracy decreases not only as $P(x_i)$ gets closer to one, but also as the probability of any false member $P(x_j)$ gets closer to zero, as follows:[24]

$$\mathrm{SR_B}(P;i) =_{\mathrm{def}} [1-P(x_i)]^2 + \sum_{j\neq i}[0-P(x_j)]^2 \tag{33}$$

We can apply the same idea to construct a global scoring rule from any local scoring rule, but SR_B is notable because it is strictly proper.[25]

Since SR_B is non-monotonic, it is possible that $P(x_i) < Q(x_i)$ and yet $SR_B(P; i) < SR_B(Q; i)$. This is easy to see by an example. Let $X = \{x_1, x_2, x_3\}$ be a partition, and x_1 be the true member. Suppose $P(x_1) = 0.38$ and $P(x_2) = P(x_3) = 0.31$, while $Q(x_1) = 0.4$, $Q(x_2) = 0.05$, and $Q(x_3) = 0.55$. Although $Q(x_1) = 0.4$ is closer to one than is $P(x_1) = 0.38$, the degree of inaccuracy for Q is greater than the degree of inaccuracy for P:

$$\begin{aligned}\mathrm{SR_B}(P;1) &= [1-P(x_1)]^2 + \sum_{j\neq 1}[0-P(x_j)]^2 \\ &= [1-0.38]^2 + [0-0.31]^2 + [0-0.31]^2 \\ &= 0.5766\end{aligned} \tag{34}$$

$$\begin{aligned}\mathrm{SR_B}(Q;1) &= [1-Q(x_1)]^2 + \sum_{j\neq 1}[0-Q(x_j)]^2 \\ &= [1-0.4]^2 + [0-0.05]^2 + [0-0.55]^2\end{aligned}$$

[24] The subscript in 'SR_B' stands for 'Brier' since an equivalent measure was introduced by Brier ([1950]). Common names for SR_B are 'the Brier scoring rule' and 'the quadratic scoring rule'.

[25] There are other global scoring rules, such as the spherical rule, that are also strictly proper, but SR_B is the most prominent among them.

$$= 0.665 \qquad (35)$$

We want to underscore the point here that this rather surprising result is not due to the proximity of the propositional contents discussed earlier. Once we identify the true member x_i of the partition, we can calculate $SR_B(P; i)$ and $SR_B(Q; i)$ regardless of the proximities of the propositional contents among x_1, x_2, and x_3.[26] Global scoring rules are exclusively truth directed just as local scoring rules are.

Since SR_B is non-monotonic, it is not ordinally equivalent to SR_L which is monotonic. Further, the measure of reduction in expected inaccuracy $REI_B(X; y)$ constructed from SR_B by the recipe in Section 6.1 is not ordinally equivalent to $D_{KL}(X; y)$; and its weighted average $REI_B(X; Y)$ over Y is not ordinally equivalent to $MI(X; Y)$, either. For example, $REI_B(X; Y)$ is not symmetric and DPI does not hold on $REI_B(X; Y)$. (See Appendix B.1 and B.2 for proof.) It is therefore important to be aware of cases where we can use a global scoring rule and thus the answers to questions of information given by $D_{KL}(X; y)$ and $MI(X; Y)$ are not definitive.

Such cases are limited in our view, but it may look otherwise by the following line of reasoning. A local scoring rule is appropriate for measuring the inaccuracy of a probability assigned to a single proposition, but for measuring the inaccuracy of a probability distribution over a partition, we must use a global scoring rule to take into account the entire probability distribution. It looks like a local scoring rule conflates two kinds of inaccuracy—the inaccuracy of a probability assigned to a single proposition and the inaccuracy of a probability distribution over a partition.[27]

There is, of course, no such conflation. As we saw earlier, the idea of a local scoring rule is to measure the inaccuracy of a probability distribution P by $P(x_i)$ which is the probability assigned to the true member of the partition. It does not ignore the rest of the partition, however, because $1 - P(x_i)$ which is the probability assigned to the rest of the partition is a function of $P(x_i)$. The distinctive feature of a global scoring rule is not that it takes into account the rest of the partition, but its sensitivity to the way $1 - P(x_i)$ is distributed among the false members. As we can see from the equations (34) and (35) above, SR_B is non-monotonic because even if $P(x_i) < Q(x_i)$ and thus $1 - P(x_i) > 1 - Q(x_i)$, it is still possible that $SR_B(P; i) < SR_B(Q; i)$ because of the way $1 - P(x_i)$ and $1 - Q(x_i)$ are distributed, respectively, among the false members of the partition.

---

[26] Note also that in the JFK case above, which we used for illustrating different proximities among the propositional contents, there would be no reduction in expected inaccuracy if we used SR_B for measuring inaccuracy.

[27] The expression 'local scoring rules' is sometimes used in reference to rules that measure the inaccuracy of a probability assigned to a single proposition (see, for example, Leitgeb and Pettigrew [2010]). This alternative terminology is consistent with the view that it is a mistake to measure the inaccuracy of a probability distribution over a partition solely by the probability assigned to the true member of the partition.

Once we are clear about the distinctive feature of a global scoring rule, we see that the use of a global scoring rule such as SR_B is not problematic in special cases where the partition is binary. This is not because monotonicity is unimportant for binary partitions, but because even a global scoring rule is monotonic for binary partitions. Let $X = \{x_1, x_2\}$ be a partition, and x_1 be the true member. Since the entire probability distribution, $P(x_1)$ and $P(x_2) = 1 - P(x_1)$, is determined by $P(x_1)$, the inaccuracy of P over X only depends on $P(x_1)$ by any scoring rule. For example, $SR_B(P; 1)$ is the following function of $P(x_1)$:

$$\begin{aligned} SR_B(P;1) &= [1 - P(x_1)]^2 + [0 - P(x_2)]^2 \\ &= [1 - P(x_1)]^2 + [0 - [1 - P(x_1)]]^2 \\ &= [1 - P(x_1)]^2 \times 2 \end{aligned} \tag{36}$$

Since SR_B is a global scoring rule, it is possible in general that $P(x_i) < Q(x_i)$ and yet $SR_B(P; i) < SR_B(Q; i)$, but there is no such possibility in special cases where the partition is binary.

So, SR_L is not the only scoring rule appropriate for measuring the inaccuracy of a probability distribution in special cases where the partition is binary. As a result, the answers given to questions of information by D_{KL} and MI are not definitive in those special cases even if EVI is important and thus information is to be measured by reduction in expected inaccuracy. Of course, if we use a global rule such as SR_B as our measure of inaccuracy, we may have difficulties in the future when we want to expand the application beyond binary cases, but that may not be an important consideration in some contexts.

It may be suggested that the use of a global scoring rule beyond binary cases is not problematic in cases where monotonicity is not essential, or that the use of a global scoring rule may even be required in some cases where the distribution of probabilities among the false members is important. We do not rule out those possibilities. We are doubtful, though, that there are many such cases.[28] We take it to be the default position that P is less accurate than Q if $P(x_i)$ for the true proposition is higher than $Q(x_i)$, and thus $1 - P(x_i)$ for the rest of the partition is lower than $1 - Q(x_i)$. SR_L is the only appropriate scoring rule beyond binary cases in the absence of some special reason.

[28] The way $1 - P(x_i)$ is distributed among the false members can make P misleading (Fallis [2007]). Suppose two distributions P and Q assign the same probability to the true member, but P assigns most of the remaining probability to one particular false member thereby making it the leading hypothesis (the member of X with the highest probability), whereas Q distributes the remaining probability evenly among the false members so that the true member is the leading hypothesis. We grant that P is misleading while Q is not, but we need not attribute the difference to different degrees of inaccuracy. Even if two probability distributions are inaccurate to the same degree, it might be that one of them is misleading while the other is not.

8. Conclusion

We examined four interpretations of MI in order to justify its use in applications where EVI (Epistemic Value of Information) is assumed. Two common interpretations of MI by reduction in doubt and reduction in uncertainty turned out to be inconsistent with EVI. The third interpretation of MI by divergence is consistent with EVI, but is faced with the problem of measure sensitivity: there are many sensible measures of divergence, and so the answers to questions of information (as divergence) given by MI are not definitive. We proposed a fourth interpretation of MI by reduction in expected inaccuracy to resolve the problem. More specifically, it was shown that the answers to questions of information (as reduction in expected inaccuracy) given by MI are definitive when inaccuracy is exclusively truth directed and a strictly proper monotonic scoring rule is appropriate. Our resolution of the problem of measure sensitivity revealed that MI is not appropriate in applications where inaccuracy is not exclusively truth directed. It is a question for further inquiry whether there is an alternative measure of information that is appropriate in such applications.

Acknowledgments

We wish to thank Vincenzo Crupi and two anonymous referees for helpful comments. We also wish to thank Elliott Sober for helpful discussion.

William Roche
Department of Philosophy
Texas Christian University
Fort Worth, TX, USA
w.roche@tcu.edu

Tomoji Shogenji
Department of Philosophy
Rhode Island College
Providence, RI, USA
tshogenji@ric.edu

Appendix A

A.1 Ordinal inequivalence

Let $X = \{x, \neg x\}$, $Y = \{y, \neg y\}$, and $Z = \{z, \neg z\}$ be partitions. Consider the following probability distribution:

x	y	z	P
T	T	T	$\frac{15}{74}$
T	T	F	$\frac{1}{1188}$
T	F	T	$\frac{1}{22}$
T	F	F	$\frac{1}{999}$
F	T	T	$\frac{17235}{76516}$
F	T	F	$\frac{994}{13959}$
F	F	T	$\frac{13}{47}$
F	F	F	$\frac{91391}{516483}$

It can be readily verified that on this distribution:

$$\mathrm{D}^*_{\mathrm{RL}}(Y;X) = \mathrm{D}^*_{\mathrm{RL}}(X;Y) > \mathrm{D}^*_{\mathrm{RL}}(Z;X)$$
$$\mathrm{D}^*_{\mathrm{SE}}(Y;X) > \mathrm{D}^*_{\mathrm{SE}}(X;Y) > \mathrm{D}^*_{\mathrm{SE}}(Z;X)$$
$$\mathrm{MI}(Z;X) > \mathrm{MI}(Y;X) = \mathrm{MI}(X;Y)$$

So, $\mathrm{D}^*_{\mathrm{RL}}$ and $\mathrm{D}^*_{\mathrm{SE}}$ are not ordinally equivalent to each other, and neither of them is ordinally equivalent to MI. QED

A.2 S (Symmetry)

First, we have:

$$\begin{aligned}\mathrm{D}^*_{\mathrm{RL}}(X;Y) &= \sum\nolimits_{j=1}^{m} P(y_j) \sum\nolimits_{i=1}^{n} \left| P(x_i \mid y_j) - P(x_i) \right| \\ &= \sum\nolimits_{j=1}^{m} \sum\nolimits_{i=1}^{n} \left| P(x_i \wedge y) - P(x_i)P(y_j) \right| \\ &= \sum\nolimits_{i=1}^{n} P(x_i) \sum\nolimits_{j=1}^{m} \left| P(y_j \mid x_i) - P(y_j) \right| \\ &= \mathrm{D}^*_{\mathrm{RL}}(Y;X)\end{aligned}$$

Meanwhile, since $\mathrm{D}^*_{\mathrm{SE}}(Y; X) > \mathrm{D}^*_{\mathrm{SE}}(X; Y)$ on the probability distribution given above in A.1, it follows that $\mathrm{D}^*_{\mathrm{SE}}$ is not symmetric. Thus, S carries over to $\mathrm{D}^*_{\mathrm{RL}}$ but not to $\mathrm{D}^*_{\mathrm{SE}}$. QED

A.3 DPI (Data-Processing Inequality)

Let $X = \{x_1, x_2\}$, $Y = \{y_1, y_2, y_3\}$, and $Z = \{z_1, z_2\}$ be partitions. Consider the following probability distribution:

x_1	y_1	y_2	y_3	z_1	P
T	T	F	F	T	$\frac{297}{800}$
T	T	F	F	F	$\frac{99}{500}$
T	F	T	F	T	$\frac{49}{79200}$
T	F	T	F	F	$\frac{3}{2000}$
T	F	F	T	T	$\frac{1}{1584}$
T	F	F	T	F	$\frac{3}{2000}$
F	T	F	F	T	$\frac{3}{800}$
F	T	F	F	F	$\frac{1}{500}$
F	F	T	F	T	$\frac{49}{800}$
F	F	T	F	F	$\frac{297}{2000}$
F	F	F	T	T	$\frac{1}{16}$
F	F	F	T	F	$\frac{297}{2000}$

It can be readily verified that on this probability distribution:

$$P(x_1 \mid y_1 \wedge z_1) = \tfrac{99}{100} = P(x_1 \mid y_1) \qquad P(x_1 \mid y_1 \wedge z_2) = \tfrac{99}{100} = P(x_1 \mid y_1)$$
$$P(x_1 \mid y_2 \wedge z_1) = \tfrac{1}{100} = P(x_1 \mid y_2) \qquad P(x_1 \mid y_2 \wedge z_2) = \tfrac{1}{100} = P(x_1 \mid y_2)$$
$$P(x_1 \mid y_3 \wedge z_1) = \tfrac{1}{100} = P(x_1 \mid y_3) \qquad P(x_1 \mid y_3 \wedge z_2) = \tfrac{1}{100} = P(x_1 \mid y_3)$$
$$P(x_2 \mid y_1 \wedge z_1) = \tfrac{1}{100} = P(x_2 \mid y_1) \qquad P(x_2 \mid y_1 \wedge z_2) = \tfrac{1}{100} = P(x_2 \mid y_1)$$
$$P(x_2 \mid y_2 \wedge z_1) = \tfrac{99}{100} = P(x_2 \mid y_2) \qquad P(x_2 \mid y_2 \wedge z_2) = \tfrac{99}{100} = P(x_2 \mid y_2)$$
$$P(x_2 \mid y_3 \wedge z_1) = \tfrac{99}{100} = P(x_2 \mid y_3) \qquad P(x_2 \mid y_3 \wedge z_2) = \tfrac{99}{100} = P(x_2 \mid y_3)$$

$$0.046 \approx \mathrm{D}^*_{\mathrm{SE}}(Y;Z) < \mathrm{D}^*_{\mathrm{SE}}(X;Z) \approx 0.059$$

It follows that (b) in DPI does not carry over to $\mathrm{D}^*_{\mathrm{SE}}$. So DPI does not carry over to $\mathrm{D}^*_{\mathrm{SE}}$. QED

Appendix B

B.1 S (Symmetry)

$\mathrm{REI}_B(Y; X) > \mathrm{REI}_B(X; Y)$ on the probability distribution given above in A.1. Thus S fails to carry over to REI_B. QED

B.2 DPI (Data-Processing Inequality)

It can be readily verified that on the probability distribution given above in A.3:

$$0.046 \approx \mathrm{REI}_{\mathrm{B}}(Y;Z) < \mathrm{REI}_{\mathrm{B}}(X;Z) \approx 0.059$$

It follows that (b) in DPI does not hold if we replace MI by $\mathrm{REI}_{\mathrm{B}}$. So, DPI fails to carry over to $\mathrm{REI}_{\mathrm{B}}$. QED

References

Barrett, M., and Sober, E. [1992]: 'Is Entropy Relevant to the Asymmetry between Retrodiction and Prediction?', *British Journal for the Philosophy of Science*, **43**, pp. 141-60.

Bernardo, J. [1970]: 'Expected Information as Expected Utility', *Annals of Statistics*, **7**, pp. 686-90.

Brier, G. W. [1950]: 'Verification of Forecasts Expressed in terms of Probability', *Monthly Weather Review*, **78**, pp. 1-3.

Brossel, P. [2013]: 'The Problem of Measure Sensitivity Redux', *Philosophy of Science*, **80**, pp. 378-97.

Cha, S. [2007]: 'Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions', *International Journal of Mathematical Models and Methods in Applied Sciences*, **4**, pp. 300-7.

Cover, T., and Thomas, J. [2006)]: *Elements of Information Theory* (2nd ed.), Hoboken: John Wiley & Sons.

Crupi, V., and Tentori, K. [2014]: 'State of the Field: Measuring Information and Confirmation', *Studies in History and Philosophy of Science*, **47**, pp. 81-90.

Evans, J., and Over, D. [1996]: 'Rationality in the Selection Task: Epistemic Utility versus Uncertainty Reduction', *Psychological Review*, **103**, pp. 356-63.

Fallis, D. [2007]: 'Attitudes toward Epistemic Risk and the Value of Experiments', *Studia Logica*, **86**, pp. 215-46.

Fano, R. [1961]: *Transmission of Information: A Statistical Theory of Communications*, Cambridge, MA: MIT Press.

Fitelson, B. [1999]: 'The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity', *Philosophy of Science*, **66**, pp. S362-78.

Grunwald, P., and Dawid, A. [2004]: 'Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory', *Annals of Statistics*, **32**, pp. 1367-433.

Jeffreys, H. [1946]: 'An Invariant form for the Prior Probability in Estimation Problems', *Proceedings of the Royal Society of London (Series A)*, **186**, pp. 453-61.

Jeffreys, H. [1948]: *Theory of Probability* (2nd ed.), Oxford: Oxford University Press.

Joyce, J. [2005]: 'How Probabilities Reflect Evidence', *Philosophical Perspectives*, **19**, pp. 153-78.

Joyce, J. [2009]: 'Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief', in F. Huber and C. Schmidt-Petri (*eds*.), *Degrees of Belief*, Dordrecht: Springer, pp. 263-97.

Kullback, S., and Leibler, R. [1951]: 'On Information and Sufficiency', *Annals of Mathematical Statistics*, **22**, pp. 79-86.

Leitgeb, H., and Pettigrew, R. [2010]: 'An Objective Justification of Bayesianism I: Measuring Inaccuracy', *Philosophy of Science*, **77**, pp. 201-35.

Milne, P. [1996]: 'log[P(h/eb)/P(h/b)] is the One True Measure of Confirmation', *Philosophy of Science*, **63**, pp. 21-6.

Murphy, A., and Winkler, R. [1984]: 'Probability Forecasting in Meteorology', *Journal of the American Statistical Association*, **79**, pp. 489-500.

Oaksford, M., and Chater, N. [1994]: 'A Rational Analysis of the Selection Task as Optimal Data Selection', *Psychological Review*, **101**, pp. 608-31.

Oaksford, M., and Chater, N. [1996]: 'Rational Explanation of the Selection Task', *Psychological Review*, **103**, pp. 381-91.

Roche, W., and Shogenji, T. [2014]: 'Dwindling Confirmation', *Philosophy of Science*, **81**, pp. 114-37.

Schupbach, J. [2011]: 'New Hope for Shogenji's Coherence Measure', *British Journal for the Philosophy of Science*, **62**, pp. 125-42.

Shannon, C. [1948]: 'A Mathematical Theory of Communication', *Bell System Technical Journal*, **227**, pp. 379-423 and pp. 623-56.

Shogenji, T. [1999]: 'Is Coherence Truth Conducive?', *Analysis*, **59**, pp. 338-45.

Shogenji, T. [2001]: 'Reply to Akiba on the Probabilistic Measure of Coherence', *Analysis*, **61**, pp. 147-50.

Sober, E. [2008]: *Evidence and Evolution: The Logic behind the Science*, Cambridge: Cambridge University Press.

Sober, E., and Barrett, M. [1992]: 'Conjunctive Forks and Temporally Asymmetric Inference', *Australasian Journal of Philosophy*, **70**, pp. 1-23.

Sober, E., and Steel, M. [2002]: 'Testing the Hypothesis of Common Ancestry', *Journal of Theoretical Biology*, **218**, pp. 395-408.

Sober, E., and Steel, M. [2011]: 'Entropy Increase and Information Loss in Markov Models of Evolution', *Biology & Philosophy*, **26**, pp. 223-50.

Sober, E., and Steel, M. [2014]: 'Time and Knowability in Evolutionary Processes', *Philosophy of Science*, **81**, pp. 558-79.

Winkler, R. [1967]: 'The Quantification of Judgment: Some Methodological Suggestions', *Journal of the American Statistical Association*, **62**, pp. 1105-20.

Winkler, R. [1969]: 'Scoring Rules and the Evaluation of Probability Assessors', *Journal of the American Statistical Association*, **64**, pp. 1073-8.

Winkler, R. [1971]: 'Probabilistic Prediction: Some Experimental Results', *Journal of the American Statistical Association*, **66**, pp. 675-85.

Winkler, R. [1994]: 'Evaluating Probabilities: Asymmetric Scoring Rules', *Management Science*, **40**, pp. 1395-405.

Winkler, R., and Murphy, A. [1968]: '"Good" probability assessors', *Journal of Applied Meteorology*, **7**, pp. 751-8.