COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 40 (2016) 758–778
Copyright © 2015 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12259

Judging the Probability of Hypotheses Versus the Impact of Evidence: Which Form of Inductive Inference Is More Accurate and Time-Consistent?

Katya Tentori,[a] Nick Chater,[b] Vincenzo Crupi[c,d]

[a]*Center for Mind/Brain Sciences, University of Trento*
[b]*Behavioural Science Group, Warwick Business School*
[c]*Center for Logic, Language, and Cognition, University of Turin*
[d]*Munich Center for Mathematical Philosophy, Ludwig Maximilian University*

Received 3 June 2014; received in revised form 20 December 2014; accepted 6 January 2015

Abstract

Inductive reasoning requires exploiting links between evidence and hypotheses. This can be done focusing either on the posterior probability of the hypothesis when updated on the new evidence or on the impact of the new evidence on the credibility of the hypothesis. But are these two cognitive representations equally reliable? This study investigates this question by comparing probability and impact judgments on the same experimental materials. The results indicate that impact judgments are more consistent in time and more accurate than probability judgments. Impact judgments also predict the direction of errors in probability judgments. These findings suggest that human inductive reasoning relies more on estimating evidential impact than on posterior probability.

Keywords: Inductive reasoning; Probabilistic reasoning; Inference; Impact; Confirmation judgments; Confirmation measures

1. Introduction

1.1. *Posterior probability and evidential impact*

Humans' spectacular ability to draw inferences from limited information underpins perception, categorization, prediction, diagnostic reasoning, and scientific discovery. Such inferences are *inductive* because they venture beyond the information given to draw conclusions that are probable given the available evidence but are not logically implied by it.

Correspondence should be sent to Katya Tentori, CIMeC, Corso Bettini 31, 38068, Rovereto (TN), Italy E-mail: katya.tentori@unitn.it

Any inductive inference concerns the relation between two elements: the hypothesis of interest (*h*) and the available evidence (*e*). Different emphasis can be given to each of these elements, focusing on one and leaving the other in the background. For example, in the inductive argument:

"X is a male student" (*e*), therefore "X owns a videogame console" (*h*)

we can focus on hypothesis *h* and wonder how much we believe it in the light of evidence *e* or we can consider evidence *e* and wonder how much impact it has on hypothesis *h*. Within a Bayesian framework, these two questions map onto two distinct notions:

(a) the *posterior probability, Pr*(*h*|*e*), of a hypothesis as updated on the new evidence, that is, the overall degree of belief in *h* given *e*;
(b) the *impact* (or degree of *confirmation*[1]), *Imp*(*h*,*e*), of new evidence on the credibility of a hypothesis, that is, whether or not (and how much) *e* strengthen/weakens the belief in *h*.

Although conceptually related, posterior probability and impact can be dissociated. For example, hypotheses with a high prior probability (e.g., *h*: "Next July, it will rain at least once in London") retain their high probability even in the light of irrelevant evidence (e.g., *e*: "the BBC has just launched a new cookery program"). In such a situation, *Pr*(*h*|*e*) is high, whereas *Imp*(*h*,*e*) is nil. On the other hand, hypotheses with an extremely low prior probability (e.g., *h*: "Nick will win the next national lottery"), might remain rather improbable even in the light of a considerable body of evidence in their favor (e.g., *e*: "Nick has just bought one tenth of the tickets of the next national lottery"). Posterior probability and evidential impact are therefore distinct notions (see below for their formal description), which are both needed to properly describe inductive arguments. Human reasoners have been shown to be able to distinguish between these two quantities (see, e.g., the results in Tentori, Crupi, Bonini, & Osherson, 2007; Tentori, Crupi, & Russo, 2013). It is, then, interesting to see whether they are equally good at estimating them. Answering this question is the primary goal of our study. In what follows, we will briefly review the literature on judgments of posterior probability and evidential impact. We will then outline an experiment comparing accuracy and time-consistency of probability versus evidential impact judgments. Finally, we will discuss the implications of the results obtained and provide possible directions for future research.

1.2. The assessment of posterior probability

The experimental study of inductive reasoning has focused mainly on the probability of hypotheses. In the psychological literature, investigations on this topic are often grouped under the label *probabilistic reasoning* and have fostered a wide-ranging debate on human rationality. Indeed, people's intuitive probability judgments systematically depart from normative benchmarks: the axioms of probability theory and their consequences, such as Bayes' theorem. With regard to belief revision in the light of new information, the documented errors include *conservatism* (Edwards, 1968), which is the tendency to stick with prior probabilities, as well as *base-rate neglect* (Tversky & Kahneman, 1982), which is, in a sense, the opposite tendency to underweight (or even ignore) prior probabilities. These findings indicate that, as new evidence becomes available, people update their subjective probabilities about hypotheses in the right direction but to an extent that significantly deviates from the amount prescribed by Bayes' theorem. Belief updating has also been shown to be affected by *order effects* (Hogarth & Einhorn, 1992), whereby judgments concerning the final probability of a hypothesis in the light of multiple pieces of evidence are not independent from the order of presentation of the evidence. Yet again, these judgments appear to be normatively defective in different directions: More importance can be given to earlier (*primacy* effect) or later (*recency* effect) pieces of evidence in a sequence, depending on factors such as the tasks' characteristics or the complexity of the stimuli. Probability judgments have also been reported to depart from elementary principles of *class inclusion* under specific circumstances (see section 1.4 and Tentori et al., 2013; for more on such circumstances). Prominent examples are the *conjunction fallacy* (Tentori, Bonini, & Osherson, 2004; Tversky & Kahneman, 1983), in which a conjunctive statement is assessed to be more likely than one of its conjuncts, and the *disjunction fallacy* (Bar-Hillel & Neter, 1993), in which a disjunctive statement is assessed to be less likely than one of its disjuncts. These phenomena have proved rather robust (see, e.g., Bar-Hillel, 1980; Crupi & Girotto, 2014; Gilovich, Griffin, & Kahneman, 2002; Tentori & Crupi, 2012b), revealing that explicit judgments of probability may indeed fall short of rational standards in systematic ways.

A more optimistic picture of human probabilistic reasoning has recently emerged from studies which tested various forms of Bayesian modeling of inference, ranging from *causal learning* and *categorization*, to *prediction* and *argumentation* (for overviews, see Chater, Oaksford, Hahn, & Heit, 2010; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). For example, Lucas and Griffiths (2010) showed that people's judgments about the functional form of different causal relationships are based on covariation data, category information, and verbal cues, in a way that is consistent with a hierarchical Bayesian model. Kemp and Tenenbaum (2009) modeled *property induction* in different (e.g., biological, spatial knowledge, etc.) contexts as a Bayesian inference. Griffiths and Tenenbaum (2011) found that, when predicting the duration or extent of phenomena from their current state, participants integrated prior knowledge and observed data in a way that is more consistent with their Bayesian model than some simple heuristics. Téglas et al. (2011) showed that a Bayesian model accounts for observed patterns of expectation and surprise in 12-month-old infants, assuming elementary and plausible abstract principles of object motion. Finally, drawing from classical catalogues of putative logical fallacies, Hahn and Oaksford (2007) provided theoretical and empirical evidence that Bayesian probability can account for how arguments are employed and assessed in various everyday settings.

Even if their actual meaning has been sometimes disputed (see, e.g., Chater et al., 2011; Jones & Love, 2011), these results suggest that human probabilistic reasoning is, when dealing with real-world problems, surprisingly impressive. Indeed, they are also consonant with the prevalence of Bayesian models of lower-level processes in perception and motor control, language, and even brain function (e.g., Chater & Manning, 2006; Doya, Ishii, Pouget, & Rao, 2007; Knill & Richards, 1996; Körding & Wolpert, 2006). We will consider one way in which to reduce the tension between the Bayesian viewpoint and the systematic biases in probabilistic reasoning tasks mentioned above. Such an attempt focuses on the kind of stimuli that participants have been presented with in these different research traditions in probabilistic reasoning: scenarios in which probability and impact values tend to co-vary versus scenarios in which probability and impact are typically dissociated.

1.3. The assessment of evidential impact

The assessment of evidential impact has received much less attention than posterior probability. A notable exception is given by experiments on *categorical induction* (Heit, 2000; Medin, Coley, Storms, & Hayes, 2003), which have considered the perceived impact from evidence to hypotheses involving familiar categories (e.g., "cats," "mammals," etc.) and so-called blank predicates (e.g., "have an ulnar artery"). Overall, participants' judgments seem to be aligned with popular principles of evidential impact, primarily based on relationships between the categories involved (Osherson, Smith, Wilkie, López, & Shafir, 1990; Sloman, 1993). For example, the *diversity principle* states that hypotheses are better supported by varied than by uniform evidence. This principle appropriately predicts participants' judgments that "rabbits use norepinephrine as a neurotransmitter" (h) is better supported by the evidence that lions and giraffes use norepinephrine as a neurotransmitter (e_1 and e_2) than by the evidence that lions and tigers do (e_1 and e_3), because lions are usually considered less similar to giraffes than to tigers. Although the diversity principle can be useful at the *descriptive* level (i.e., capturing a general tendency in people's judgments), Lo, Sides, Rozelle, and Osherson (2002) convincingly argued that it does not have the *normative* status that psychologists often attribute to it. In fact, there are many arguments in which the principle is inappropriate or is over-ridden. For example, although it seems undeniable that housecats resemble tigers more than they resemble fieldmice, it appears perfectly defensible to judge the conclusion "all mammals often carry the parasite Floxum" (h) as better supported by the evidence that housecats and tigers often carry the parasite Floxum (e_1 and e_2) than by the evidence that housecats and fieldmice do (e_1 and e_3) on the grounds of a possible predator–prey relation.

Lo et al. (2002) replaced the diversity principle with a rule based purely on probabilities defined over hypotheses and evidence. Rules of this kind are known in the epistemology literature as *Bayesian confirmation measures* and give numerical expression to the impact of evidence e on hypothesis h as a function of some combination of probability values defined over e and h. These models differ with regard to what specific probability values have to be considered, and how they should be combined, but share the following qualitative properties of impact (Carnap, 1962):

$$Imp(h,e) = \begin{cases} > 0 \, iff \; Pr(h|e) > Pr(h) \\ = 0 \, iff \; Pr(h|e) = Pr(h) \\ < 0 \, iff \; Pr(h|e) < Pr(h) \end{cases}$$

This means that, for all these models, evidence e has a *positive* [*negative*] impact on hypothesis h if and only if the posterior probability of the hypothesis h in the light of evidence e is *higher* [*lower*] than the prior probability of h (for some examples of positive vs. negative impact, see section 1.1).

There are several ways to quantify impact (for recent reviews, see Brössel, 2013; Crupi, Tentori, & Gonzalez, 2007; Festa, 2012; Glass, 2013; Roche & Shogenji, 2014). We decided to employ three different models (that will be presented in section 2.1.1, along with the motivation for their selection from those available in the literature.). However, in addition to the qualitative definition provided above, all Bayesian models of impact share two important properties. First, they provide precise normative benchmarks against which to evaluate the accuracy of impact judgments. Second, they can be used for any kind of inductive arguments, and not only categorical ones. For these reasons, we decide to use them in our study.

In spite of their popularity in epistemology, Bayesian models of impact are rarely studied in psychological research. When they are, however, participants consistently have proved accurate in estimating evidential impact, both with categorical (Lo et al., 2002) and non-categorical arguments concerning artificial material (e.g., urns and balls of different colors, Tentori, Crupi, Bonini, et al., 2007) as well as with real-world predicates (e.g., "to be a male," "to own a motorbike worth 10,000 Euros," Mastropasqua, Crupi, & Tentori, 2010). Accurate impact judgments were also obtained when the uncertainty of evidence was manipulated, either explicitly (directly providing numerical information concerning the probability of the evidence) or implicitly (employing ambiguous pictures as evidence). Note that the latter tasks are particularly difficult because they require the degree of uncertainty of the evidence to be integrated into the assessment of impact. Impact judgments seemed to be particularly accurate when participants were most confident about the relevant probability distribution (e.g., in urns and balls scenarios where the number of balls of each kind in each urn is explicitly given). Otherwise (e.g., in most real-life scenarios in which the relevant probability values are less sharply defined), impact judgments, although in the right direction, tended to be systematically more moderate than they should be (Tentori, Crupi & Osherson, 2007, 2010).

1.4. Comparing the assessment of posterior probability and impact

The results of the studies cited in the previous paragraphs are not, of course, directly comparable, given that they employed different participants, procedures, and stimuli. They do raise, however, the possibility that people might be better at judging impact than posterior probability. Data reported in Tentori, Crupi, Bonini, et al. (2007), and Crupi et al. (2007) support this conjecture: In an urn setting, normative Bayesian confirmation measures were better predictors of elicited impact judgments when degrees of impact were computed by the true statistical probabilities rather than those subjectively estimated by the participants, the latter having been found to be prone to well-known biases (in particular, conservative posteriors, see section 1.2. above). This result suggests that, psychologically, impact may be more fundamental than probability. Finally, convergent evidence arises also from the observation that impact affects the occurrence and prevalence of probabilistic errors, as in the conjunction fallacy (see section 1.2) (Crupi, Fitelson, & Tentori, 2008; Tentori & Crupi, 2012a; Tentori et al., 2013). In particular, in Tentori et al. (2013), participants were presented with three statements of the form h_1, $h_1 \wedge h_2$, and $h_1 \wedge h_3$. Hypotheses h_2 and h_3 were selected in such a way that h_2 ranked higher than the h_3 in assessments of impact, but lower in judged probability. An example of such a scenario is the following:

O. has a degree in violin performance. [e]

Which of the following hypotheses do you think is the most probable?
O. is an expert mountaineer [h_1, correct option]
O. is an expert mountaineer and gives music lessons [$h_1 \wedge h_2$, conjunction fallacy]
O. is an expert mountaineer and owns an umbrella [$h_1 \wedge h_3$, conjunction fallacy]

Fallacious responses systematically targeted $h_1 \wedge h_2$ more than $h_1 \wedge h_3$, showing that the impact of the evidence on the added conjunct (and not the probability of the added conjunct in the light of the evidence) is the key determinant of the conjunction fallacy. This result also reveals that the assessment of impact can be implicit and relevant even in tasks in which only the probability of hypotheses should be at issue.

The purpose of this study is directly to compare the reliability of impact versus probability judgments, by testing their accuracy and consistency on the same stimuli. While consistency over time might be a fairly uncontroversial notion, one might still wonder what it means precisely for these two kinds of judgments to be or not to be accurate. This point seems to require clarification before we proceed, because important differences exist as to how the normative status of these models is usually advocated.[2] Key theoretical results are known to show that (given appropriate auxiliary assumptions) departures from probability principles would imply sure monetary losses (see Osherson, 1995; Vineberg, 2011) or otherwise avoidable epistemic costs (see D'Agostino & Sinigaglia, 2010; Leitgeb & Pettigrew, 2010a,b; Predd et al., 2009). Indeed, the fact that a variety of normative arguments lead to the same set of probabilistic principles is a powerful motivation for the uniqueness of probability as a measure of degree of belief. To the best of our knowledge, no comparable result exists so far for evidential impact. Available arguments in support of different impact measures as normatively compelling have been largely based on the theoretical or intuitive appeal of specific formal properties. Importantly, our talk of accurate judgments solely concerns agreement with the chosen benchmark models (i.e., standard probability vs. impact measures, respectively; see section 2.1.1 below) in the experimental setting, and not the normative debates mentioned above. A tentative discussion of the wider cognitive role of probability versus impact judgments will be postponed to the concluding section.

Regarding our experimental comparison of assessments of impact and probability, we aimed at a compelling test by using real-world arguments, namely a more demanding setup for impact judgments as shown by previous findings (see section 1.3). If assessments of impact are both more accurate and more consistent than corresponding probability judgments even here, then the former rather than the latter is likely to be the very basis for sound inductive reasoning.

2. The experimental study

2.1. Stimuli

2.1.1. Preliminary phase

In a preliminary phase, we asked a convenience sample of 200 undergraduates drawn from various UCL departments (100 females and 100 males) to fill in a survey with a series of personal questions, such as the following:

> *Do you have a driving license? Do you own (at least) one videogame console? Can you ski? Do you support any football team? Do you like cigars? Do you like shopping? Do you have freckles?*

Response frequencies were used to derive objective probabilities (e.g., the probability that a UCL student has a driving license given that s/he is female/male) and corresponding impact values (e.g., the impact of the evidence that a UCL student is female/male on the hypothesis that s/he has a driving license). Impact values were computed according to the following measures:

$$Imp_R(h,e) = \frac{Pr(h|e) - Pr(h)}{Pr(h|e) + Pr(h)} \quad \text{(Keynes, 1921; Horwich, 1982);}$$

$$Imp_L(h,e) = \frac{Pr(e|h) - Pr(e|\neg h)}{Pr(e|h) + Pr(e|\neg h)} \quad \text{(Kemeny \& Oppenheim, 1952; Good, 1984);}$$

$$Imp_Z(h,e) = \begin{cases} \frac{Pr(h|e) - Pr(h)}{Pr(\neg h)} \text{ iff } Pr(h|e) \geq Pr(h) \\ \frac{Pr(h|e) - Pr(h)}{Pr(h)} \text{ iff } Pr(h|e) < Pr(h) \end{cases} \quad \text{(Crupi et al., 2007).}$$

These models—based on the *probability ratio* (*R*), *likelihood ratio* (*L*), and *relative distance* (*Z*), respectively—satisfy the qualitative definition of impact presented in section 1.3, but they differ quantitatively and indeed ordinally. This means that *R*, *L*, and *Z* always agree in classifying the impact of evidence *e* on hypothesis *h* as positive versus negative, but they can differ in the quantification of such impact. Since our study aims quantitatively to compare probability and impact judgments, employing three distinct normative models of impact allows us to test the generalizability of our results. We selected these specific three models among those available on the basis of the following appealing metric features, that are widely employed in the literature. All three measures range over the bounded interval $[-1,1]$, which turns out to be convenient for an empirical study. Imp_R also complies with so-called *law of likelihood*, stating that evidence e has a stronger impact on h_1 than on h_2 if and only if e is more likely under the former than under the latter hypothesis (see, e.g., Crupi, Chater, & Tentori, 2013; Milne, 1996). Imp_L and Imp_Z, on the other hand, satisfy other popular principles such as the symmetry condition $Imp(h,e) =-Imp(\text{not-}h,e)$ (see, e.g., Crupi et al., 2007; Eells & Fitelson, 2002; and references therein). A distinctive feature of Imp_L, advocated by Good (1984), is that it is fully determined once that pair of likelihood values $P(e|h)$ and $P(e|\text{not-}h)$ is given, regardless of the value of the prior, $P(h)$. The specific appeal of measure Imp_Z lies, instead, in the generalization of the basic logical relations of *entailment* and *refutation*, as shown in Crupi and Tentori (2013, 2014a,b). Finally, note that, although Bayesian models of impact express impact as a function of some combination of probability values (a property named *formality* in Tentori, Crupi, and Osherson, 2007, 2010), this appears contingent upon historical circumstances (the notion of probability was formalized much earlier than impact) and does not necessarily imply that probability is the more psychologically fundamental notion. On the contrary, the results of the experimental tests presented in section 1.2 suggest that impact judgments can be provided directly, without making the relevant probability values explicit.

2.1.2. Constructing the experimental arguments

Once estimates of the objective probabilities and corresponding impact values were computed, we generated 56 arguments by combining two complementary pieces of evidence ("X is a male [female] student") with 28 different hypotheses (e.g., "X has a driving license," "X likes cigars," etc.).

The decision to use only two pieces of evidence was motivated by the objective of keeping the base rate of the evidence constant and equal to .5, in order to have a complete mapping between probability and impact.[3] Evidence about gender serves this end appropriately, because e and not-e are naturally perceived as equally probable and both can be expressed in the affirmative mode (e.g., "female" rather than "not male"). Moreover, gender is a simple attribute that participants can easily connect with the properties appearing in the hypotheses of interest. Note also that evidence e (e.g., "X is a male") and not-e ("X is a female") either are neutral with regard to a hypothesis h or have an opposite (i.e., positive vs. negative) impact on it. Therefore, using both these pieces of evidence guarantees an identical number of arguments with *positive* and *negative* impact.

Although high [low] posterior probability and positive [negative] impact tend to be positively correlated, they can be dissociated, as illustrated in section 1.1. To minimize possible biases in the stimuli, we selected the 28 hypotheses to generate (together with our two pieces of evidence) all possible combinations of high/low posterior probability

Table 1
Classification of the 56 arguments employed, according to the (higher vs. lower than .5) probability of the hypothesis and the (positive vs. neutral vs. negative) impact of the evidence

		$Imp(h, e)$			
		>0	= 0	<0	
$Pr(h\mid e)$	>.5	N = 18 aver. $Pr(h\mid e)$ = .74	N = 4 aver. $Pr(h\mid e)$ = .92	N = 6 aver. $Pr(h\mid e)$ = .62	N = 28 aver. $Pr(h\mid e)$ = .74
	<.5	N = 6 aver. $Pr(h\mid e)$ = .33	N = 4 aver. $Pr(h\mid e)$ = .16	N = 18 aver. $Pr(h\mid e)$ = .30	N = 28 aver. $Pr(h\mid e)$ = .29
		N = 24 aver. $Pr(h\mid e)$ = .64	N = 8 aver. $Pr(h\mid e)$ = .54	N = 24 aver. $Pr(h\mid e)$ = .38	N = 56 aver. $Pr(h\mid e)$ = .51

Notes. In each cell in the table, the upper value shows the number of arguments, *N*. Immediately below the number of augments is the average objective probability of the hypotheses in that cell.

and positive/negative impact. More specifically, we generated an identical number of arguments with high (>.5) and low (<.5) posterior probability of the hypotheses. The 28 arguments with high posterior probability included 18 arguments with positive impact, four arguments with neutral impact, and six arguments with negative impact, whereas the 28 arguments with low posterior probability included six arguments with positive impact, four arguments with neutral impact, and 18 arguments with negative impact (for a schematic overview, see Table 1; for the complete list of the arguments in each class, see Appendix A). Note that, as all Bayesian models of impact agree on the qualitative definition of impact (see section 1.3), this classification does not depend on the specific model adopted.

2.2. Participants

A new convenience sample of 35 UCL undergraduates drawn from various UCL departments (M_{age} = 22.43 years; 21 females) was recruited for the experiment. Each student received £10 for her/his participation.

2.3. Procedure

Participants came to the laboratory twice, with an interval of 7–10 days. The two sessions were identical and, on both occasions, participants were presented with each of the 56 arguments selected and asked to judge:

- the *probability of the hypothesis* in the light of the evidence provided;
- the *impact of the evidence* provided on the credibility of the hypothesis.

To control for possible carry-over effects, the order of arguments as well as the order of probability and impact questions were balanced across participants. We also wanted to minimize possible effects of the response format. To this aim, for both probability and impact judgments, we employed two different scales: a *discrete* 100-point scale (ranging from 0 to 100 for probability judgments and from −50 to +50 for impact judgments) and a *continuous* scale (ranging from "certainly true" to "certainly false" for probability judgments and from "maximally weakens" to "maximally strengthens" for impact judgments). So, while in the former case, participants provided a number, in the latter, they marked a position on a line spaced evenly from left to right (for a detailed outline of the scales and the questions used, see Appendix B). To avoid any confusion between the two scales, we randomly assigned participants to two groups: 19 (group 1) were presented with a discrete probability scale and a continuous impact scale, 16 (group 2) with a continuous probability scale and a discrete impact scale.

2.4. Results

Let us denote by $Pr_{\_survey}$ the probability values obtained from the survey reported above, and with $ImpR_{\_survey}$, $ImpL_{\_survey}$, $ImpZ_{\_survey}$ the impact values computed by plugging the relevant survey probabilities into the three impact models discussed in the previous paragraph. Participants' judgments in the two experimental sessions will be denoted with Pr_{1_judged} and Pr_{2_judged}, for probability, and Imp_{1_judged} and Imp_{2_judged}, for impact.

2.4.1. Time-consistency

As a general consistency measure, we correlated each participant's 56 judgments in the first session with her/his corresponding 56 judgments in the second session, that is, Pr_{1_judged} with Pr_{2_judged}, and Imp_{1_judged} with Imp_{2_judged}. The average correlations across all participants ($N = 35$) are $r = .86$ for probability and $r = .91$ for impact, and crucially the difference between these correlations is significant by a paired t-test ($t(34) = -3.722$, $p < .01$).[4] For both probability and impact judgments, there are no differences between the two (continuous vs. discrete scale) groups by an independent t-test ($t(33) = 0.842$, n.s., for probability and $t(33) = -0.765$, n.s., for impact). Therefore, although both judgments are rather consistent over time, impact judgments are significantly more consistent than probability judgments, and this does not depend on the scale used to collect the judgments.[5]

2.4.2. Accuracy

As a first test of accuracy, we correlated each participant's 56 judgments in the first session with the corresponding 56 values obtained from the survey. That is, we correlated the Pr_{1_judged} with $Pr_{\_survey}$, and similarly we correlated Imp_{1_judged} with each of $ImpR_{\_survey}$, $ImpL_{\_survey}$, and $ImpZ_{\_survey}$. The average correlations across all participants ($N = 35$) are $r = .59$ for probability and $r = .77, .82, .80$ for impact, as quantified with $Imp_{R,L,Z}$, respectively. The difference in the correlations between impact and probability is significant by a paired t-test ($t(34) = 8.055, 11.144, 10.155$, for $Imp_{R,L,Z}$, respectively, all $p < .01$). Therefore, impact judgments are more correlated than probability judgments with the corresponding values derived from the survey to estimate the objective probabilities, regardless of the measure employed to quantify objective impact.

Correlations for impact judgments do not differ by an independent t-test ($t(33) = 0.717$, 0.378, 0.635 for $Imp_{R,L,Z}$, respectively, all n.s.) in the two (continuous vs. discrete scale) groups. In contrast, probability judgments correlate more with the corresponding objective values when expressed on a discrete rather than continuous scale ($r = .67$ for group 1 vs. $r = .49$ for group 2) by an independent t-test ($t(33) = 6.197$, $p < .01$). However, even if we focus exclusively on group 1, the average correlation for probability judgments ($r = .67$) is still significantly lower than those for impact ($r = .77$, .83, .80 as quantified with $Imp_{R,L,Z}$, respectively) by a paired t-test ($t(18) = -4.245$, -7.144, -6.204, for $Imp_{R,L,Z}$, respectively, all $p < .01$). Therefore, impact judgments are more correlated than probability judgments with the corresponding values obtained from the survey, even if we selectively consider the group of participants whose probability judgments are most correlated with the corresponding objective values.

We then analyzed the degree to which the judgments of the experimental participants' agree with the values derived from our prior survey. To begin with, we computed, for each participant, the average absolute error for both impact and probability judgments across all 56 arguments, that is, average $|Pr_{1\_survey} - Pr_{1\_judged}|$, $|Imp_{R\_survey} - Imp_{1\_\ judged}|$, $|Imp_{L\_survey} - Imp_{1\_judged}|$, and $|Imp_{Z\_survey} - Imp_{1\_judged}|$. The average absolute error (normalized on a 100 point scale) across all participants ($N = 35$) is 20.08 for probability versus 13.12, 10.67, 10.97 for impact, as quantified with $Imp_{R,L,Z}$, respectively. The difference in error between probability and impact is significant by a paired t-test ($t(34) = 9.733$, 16.791, 17.135 for $Imp_{R,L,Z}$, respectively, all $p < .01$). Therefore, the errors in probability judgments are almost twice as large as those in impact judgments, regardless of the measure employed to quantify impact.

As with the correlations, we compared the average error for the two (continuous vs. discrete scale) groups of participants. The average absolute error (normalized on a 100 point scale) in probability judgments is 18.79 for group 1 and 21.62 for group 2, which are significantly different by an independent t-test ($t(33) = -2.778$, $p < .01$), indicating again, that, for this type of stimuli, probability judgments are more accurate when expressed on a discrete scale. The average absolute errors (normalized on a 100 point scale) in impact judgments are 11.67, 9.59, 9.85 in group 1 and 14.83, 11.95, 12.30 in group 2, for impact, as quantified with $Imp_{R,L,Z}$, respectively. Their difference is significant by an independent t-test ($t(33) = -2.345$, -2.648, -2.785, for $Imp_{R,L,Z}$, respectively, all $p < .05$). This suggests that impact judgments are closer to the corresponding objective values when expressed on a continuous scale. Note, however, that even if we focus exclusively on judgments provided on a discrete scale (i.e., probability judgments in group 1 and impact judgments in group 2), the average error in probability judgments (18.79) is still significantly larger than the average errors in impact judgments (14.83, 11.95, 12.30, for impact, as quantified with $Imp_{R,L,Z}$, respectively) by an independent t-test ($t(33) = 3.050$, 6.421, 6.123, for $Imp_{R,L,Z}$, respectively, all $p < .01$. Therefore, the average error is larger in probability than impact judgments, even if we selectively compare the groups of participants with the most accurate probability judgments and the least accurate impact judgments.

Finally, one might wonder if errors in probability judgments depend on the corresponding impact values. In particular, previous experiments on the conjunction fallacy (Tentori & Crupi, 2012a; Tentori et al., 2013) suggest that people tend to overestimate [underestimate] the probability of hypotheses which are confirmed [disconfirmed] by the available evidence. The arguments employed in our study are not ideal for testing this possibility because (much as in ordinary real-life) a positive correlation exists between impact and probability ($r = .53, .57, .67$, for impact, as quantified with Imp_{R},$_{L}$,$_{Z}$, respectively), and the average probability of the confirmed hypotheses is higher than that of disconfirmed ones (.64 vs. .38, see Table 1). As a consequence, the probability estimates for the two classes could be affected by ceiling and floor effects and cannot be directly compared. To circumvent these difficulties, we quantified the degree of agreement between probability errors and the direction of impact judgments. In particular, we considered, for each participant, all the arguments (of the 56) whose impact s/he had judged as different from zero. Then, we computed the absolute difference between the corresponding objective and subjective probability, that is $|Pr_{1\_obj} - Pr_{1\_subj}|$, and assigned it a positive sign whenever the participant had judged the impact as positive [negative] and had overestimated [underestimated] the objective probability, a negative sign otherwise. Finally, we averaged all these differences taken with their sign. Such an index is positive if the participant overestimated the probability of the confirmed hypotheses and underestimated the probability of the disconfirmed hypotheses more than s/he overestimated the probability of disconfirmed hypotheses or underestimated the probability of confirmed hypotheses. Zero represents the situation in which the participant made no errors at all or the same amount of error in line with versus against what is predicted by impact. For the great majority (80%) of participants, the index was positive. The sample mean across all 35 participants is +6.1, which is statistically different from the assumed null value of 0 (one-sample t-test, $t(34) = 6.087$, $p < .01$). Thus, errors in probability judgments are influenced by corresponding impact values in the predicted direction: When impact is positive [negative], participants tended to overestimate [underestimate] the corresponding posterior probability (by 6%, on average).

To summarize, impact judgments are significantly more time-consistent and more accurate than probability judgments. Impact judgments also predict the direction of the errors in probability judgments. These results do not depend on the specific measure employed to quantify impact or on the specific scale used to collect the judgments.

3. Discussion

The results of this study corroborate, with a different procedure and new materials, a previous result of ours (Tentori, Crupi, & Osherson, 2007, 2010): Although in the Bayesian tradition impact is formally expressed as a function of probability, its cognitive assessment seems to be a primitive type of judgment. In addition, the current study allows us to conclude that, under comparable experimental conditions, impact judgments are more accurate and consistent than probability judgments also in a real-life setting, whose statistical structure has to be subjectively estimated by the participants.

The greater time-consistency and accuracy of impact judgments has been found by employing different kinds of arguments, which differ for the (high vs. low) probability of the hypothesis and the (positive vs. negative) impact of the evidence, as well as two (continuous vs. discrete) scales to collect the judgments. The stability of these results across the scale used is particularly striking, because our participants were likely to be somewhat familiar with (at least some) explicit probability questions but had almost certainly never previously explicitly rated evidential impact.

Might our findings depend on the specific content of the arguments? Realistic material as that employed in this study has been proven (see section 1.3) to be associated to less accurate impact judgments. This is not necessarily the case for probability judgments, which often proved more markedly suboptimal with abstract materials. It is therefore plausible that the difference that we found in time-consistency and accuracy between impact and probability judgments could be even stronger with abstract arguments. On the other hand, gender stereotypes are usually well formed and might have helped our participants in guessing the direction of impact at least with reference to some hypotheses. However, it is on quantitative (and not only qualitative) grounds that judgments of impact proved to be more accurate and consistent than judgments of probability. More generally, some stereotypes could themselves be generated and maintained by a prevalence of impact over posterior probability. For example, stereotypes such as “males like cigars” and “males do not like shopping” could arise even if actually a minority (38%, in our sample) of males like cigars and the majority of them (61%, in our sample) like shopping simply because even fewer [more] females like cigars [shopping].

There has been much prior discussion of the puzzling phenomenon that people seem to reason effectively about a highly uncertain world, even though their explicit probability judgments, especially in laboratory conditions, appear to be unstable and inconsistent (e.g., Evans & Over, 1996; Oaksford & Chater, 2007). According to a standard line of argument inspired by the pragmatics of communication, such a discrepancy might result from participants’ misinterpretation of the stimuli and/or experimental task. Although many techniques have been developed to control for these alleged sources of misinterpretation, typically the biases in probability judgments are not dissipated (for a review concerning the conjunction fallacy, e.g., see Moro, 2009). Our results suggest a novel approach to bridge the reality-laboratory gap. In dealing with everyday uncertainty, we suggest, people appear rational because they rely more on detecting relations of impact than on computing values of posterior probability. In most real-life circumstances, this would not constitute a problem, because these two kinds of assessments often yield similar results. That is, when evidence has a strong positive [negative] impact on a hypothesis, then the probability of the latter in the light of the former is usually rather high [low]. However, as said above, the two variables can be dissociated, as it occurs, for example, in a typical conjunction fallacy scenario (see Tentori et al., 2013). When this happens, we argue, people are particularly exposed to biased probability judgments, whose direction and magnitude depend on relevant impact relations. Note, also, that such a view allows a possible way to reconcile the results of traditional versus more recent studies of human probabilistic reasoning mentioned in section 1.2. In fact, the experimental scenarios usually considered in the Bayesian modeling studies (in which good performances are typically reported) try to simulate statistical inferences in real-world settings and thus differ from the classical scenarios employed in the heuristics and biases tradition in that there is not any strong dissociation between impact and probability.

Why are people more sensitive to impact than posterior probability? The question is far from trivial, because probability judgments are acknowledged to be essential in various activities, such as the prediction of future events, decision between risky prospects, categorization, etc. The usefulness of impact judgments has been much less discussed in the literature. However, the assessment of impact is crucial for many tasks and, in particular, those in which the value of information or the soundness of arguments has to be considered. For example, in medical practice, impact judgments may help establish the most useful clinical evidence to acquire to test diagnostic hypotheses (e.g., Crupi & Tentori, 2014b; Klayman & Ha, 1987; Nelson, 2005). Considering impact relations assists hypothesis generation and, in general, learning. Another important area in which impact judgments are presumably involved is communication and persuasion. In fact, a shared assumption in linguistics and pragmatics (see Frank & Goodman, 2012) is that speakers attempt to be informative and convincing, while the listeners use inference to recover speakers' intended referents. Being skilled at identifying impact could be therefore crucial for winning arguments and influencing others' opinions.[6]

It is also interesting to notice that impact captures the relation between two variables, whereas probability is an absolute judgment. Impact is therefore more suitable than probability to track more or less direct causal dependencies and, as a consequence, might achieve a higher degree of stability in response to changing background knowledge. Imagine asking yourself, for example, the probability that a person living in a certain place (e.g., Italy, Sudan, etc.) has a tuberculosis infection (TBI) given that she is coughing up blood. To provide a precise answer seems quite hard because, even if you are aware that TBI can be rather common in patients who cough up blood, to quantify the exact probability requires more information about the spread of TBI and alternative diseases compatible with that symptom in the population under consideration. However, when it comes to impact, things seem to be different. You know that, in most (if not all) circumstances, to cough up blood is valuable (albeit inconclusive) supporting evidence for the hypothesis of having TBI (in Italy as well as in Sudan, etc.). Therefore, even without being expert on the specific population under consideration, you can conclude that this piece of evidence has an appreciable positive impact on the hypothesis at issue. Thus, it is possible that, although impact and causality judgments are distinct, they may support each other: Detecting impact may be an effective way of discovering and modeling new causal relations, while the knowledge of causal dependencies may lead to reliable impact assessments. The study of the connections between assessment of impact and perception of causality could provide important insights, as has been apparent in formal work relating measures of confirmation and causal strength (Fitelson & Hitchcock, 2011). Note, finally, that stability may, indeed, be a criterion for choosing which measures of impact are more causally relevant, computationally useful, and, perhaps, cognitively plausible.

Future research should be able to deepen our understanding of why we are more able to judge impact over probability exploring experimentally some of the issues outlined above. This will, we hope, help shift discussion from often inconclusive debates concerning the extent of human rationality, to the question of how our mind works when making reliable inferences in an uncertain world.

Acknowledgments

Katya Tentori acknowledges support by the MIUR grant PRIN 2010RP5RNM 006 and the Marie Curie grant INDUCT 221350; Nick Chater was supported by the ERC grant 295917-RATIONALITY, the ESRC Network for Integrated Behavioural Science ES/K002201/1, and the Leverhulme Trust RP2012-V-022, Research Councils UK Grant EP/K039830/1; Vincenzo Crupi was supported by the DFG grant SPP 1516 CR 409/1-2 and the MIUR grant FIRB D11J12000470001.

Notes

1. In the epistemological literature, evidential support is often referred to as *confirmation*. However, here we prefer to call it *impact*, for a number of reasons. First, the technical meaning of *confirmation* departs from that of natural language, in which it often implies complete validation (thus, in normal usage, if we confirm that Bill came to the party, this implies that he definitely came to the party). Second, *confirmation* only conveys the idea of positive support while, of course, impact can be negative as well (*disconfirmation*). Third, in the psychological literature, the term *confirmation* has gained a negative connotation because of so-called *confirmation bias* (Nickerson, 1998).
2. We are grateful to an anonymous reviewer for raising this point.
3. According to *ImpR*, in particular, positive impact cannot be very high if $Pr(e)$ is itself very high.
4. All the results reported in this study have been replicated with non-parametric statistics (Wilcoxon signed-rank test for paired samples and Mann–Whitney *U*-test for independent samples).
5. Nor it seems to be related to the dispersion of participants' impact and probability judgments.
6. We thank Nick Beckstead for this suggestion.

References

Bar-Hillel, M. (1980). The base rate fallacy in probability judgments. *Acta Psvchologica*, *44*, 211–233.

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely it is: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology*, *65*, 1119–1131.

Brössel, P. (2013). The problem of measure sensitivity redux. *Philosophy of Science*, *80*, 378–397.

Carnap, R. (1950/62). *Logical foundations of probability*. Chicago: University of Chicago Press.

Chater, N., Goodman, N., Griffiths, T., Kemp, C., Oaksford, M., & Tenenbaum, J. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral & Brain Sciences*, *34*, 194–196.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*, 335–344.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 811–823.

Crupi, V., Chater, N., & Tentori, K. (2013). New axioms for probability and likelihood ratio measures. *The British Journal for the Philosophy of Science*, *64*, 189–204.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, *14*, 182–199.

Crupi, V., & Girotto, V. (2014). From *is* to *ought*, and back: How normative concerns foster progress in reasoning research. *Frontiers in Psychology*, *5*, 219.

Crupi, V., & Tentori, K. (2013). Confirmation as partial entailment: A representation theorem in inductive logic. *Journal of Applied Logic*, *11*, 364–372.

Crupi, V., & Tentori, K. (2014a). Erratum to “Confirmation as partial entailment.” *Journal of Applied Logic*, *12*, 230–231.

Crupi, V., & Tentori, K. (2014b). State of the field: Measuring information and confirmation. *Studies in the History and Philosophy of Science*, *47*, 81–90.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, *74*, 229–252.

D’Agostino, M., & Sinigaglia, C. (2010). Epistemic accuracy and subjective probability. In M. Suárez, M. Dorato, & M. Rèdei (Eds.), *Epistemology and methodology of science* (pp. 95–105). Berlin: Springer.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Eells, E., & Fitelson, B. (2002). Symmetries and asymmetries in evidential support. *Philosophical Studies*, *107*, 129–142.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.

Festa, R. (2012). For unto every one that hath shall be given. Matthew properties for incremental confirmation. *Synthese*, *184*, 89–100.

Fitelson, B., & Hitchcock, C. (2011). Probabilistic measures of causal strength. In P. McKay Illari, F. Russo & J. Williamson (Eds.), *Causality in the sciences* (pp. 600–627). New York: Oxford University Press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.

Glass, D. H. (2013). Confirmation measures of association rule interestingness. *Knowledge-Based Systems*, *44*, 65–77.

Good, I. J. (1984). The best explicatum for weight of evidence. *Journal of Statistical Computation and Simulation*, *19*, 294–299.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring the laws of thought. *Trends in Cognitive Sciences*, *14*, 357–364.

Griffiths, T. L., & Tenenbaum, J. B. (2011). Predicting the future as Bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *140*, 725–743.

Hahn, U. & Oaksford, M. (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies'. *Psychological Review*, *114*, 704–732.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*, 569–592.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55.

Horwich, P. (1982). *Probability and evidence*. Cambridge, UK: Cambridge University Press.

Jones, M. & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–188.

Kemeny, J., & Oppenheim, P. (1952). Degrees of factual support. *Philosophy of Science*, *19*, 307–324.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20–58.

Keynes, J. (1921). *A treatise on probability*. London: Macmillan.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information. *Psychological Review*, *94*, 211–228.

Knill, D. C., & Richards, W. (Eds.) (1996). *Perception as bayesian inference*. Cambridge, UK: Cambridge University Press.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*, 319–326.

Leitgeb, H., & Pettigrew, R. (2010a). An objective justification of Bayesianism I: Measuring inaccuracy. *Philosophy of Science*, *77*, 201–235.

Leitgeb, H., & Pettigrew, R. (2010b). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*, 236–272.

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science*, *26*, 181–206.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, *34*, 113–147.

Mastropasqua, T., Crupi, V., & Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence. *Memory & Cognition*, *38*, 941–950.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review*, *10*, 517–532.

Milne, P. (1996). Log[P($h|eb$)/P($h|b$)] is the one true measure of confirmation. *Philosophy of Science*, *63*, 21–26.

Moro, R. (2009). On the nature of the conjunction fallacy. *Synthese*, *171*, 1–24.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact and information gain. *Psychological Review*, *112*, 979–999.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford, UK: Oxford University Press.

Osherson, D. (1995). Probability judgment. In E. E. Smith, & D. Osherson (Eds.), *An invitation to cognitive science, 3: Thinking* (pp. 35–75). Cambridge, MA: MIT Press.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.

Predd, J. B., Seiringer, R., Lieb, E. J., Osherson, D., Poor, H. V., & Kulkarni, S. R. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, *55*, 4786–4792.

Roche, W., & Shogenji, T. (2014). Dwindling confirmation. *Philosophy of Science*, *81*, 114–137.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference *Science*, *332*, 1054–1059.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, *28*, 467–477.

Tentori, K., & Crupi, V. (2012a). How the conjunction fallacy is tied to probabilistic confirmation: Some remarks on Schupbach (2012). *Synthese*, *184*, 3–12.

Tentori, K., & Crupi, V. (2012b). On the conjunction fallacy and the meaning of and yet again: A reply to Hertwig, Benz, and Krauss (2008). *Cognition*, *122*, 123–134.

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*, 107–119.

Tentori, K., Crupi, V., & Osherson, D. (2007). Determinants of confirmation. *Psychonomic Bulletin & Review*, *14*, 877–883.

Tentori, K., Crupi, V., & Osherson, D. (2010). Second order probability affects hypothesis confirmation. *Psychonomic Bulletin & Review*, *17*, 129–134.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Confirmation versus probability. *Journal of Experimental Psychology: General*, *142*, 235–255.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). New York: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–3 l5.

Vineberg, S. (2011). Dutch book arguments. In E. N. Zalta (Eds.), *Stanford encyclopedia of philosophy*. Available at: http://plato.stanford.edu/archives/sum2011/entries/dutch-book. Accessed May 25, 2014.

Appendix A: Arguments employed in the experiment

	Pr(h|e) > .5 and Pr(h|¬e) > .5	**Pr(h|e) > .5 and Pr(h|¬e) < .5**	**Pr(h|e) < .5 and Pr(h|¬e) > .5**	**Pr(h|e)< .5 and Pr(h|¬e)< .5**
***Imp*(*h*,*e*) > 0**	e= X is a **male** h= X has a driving licence h= X owns (at least) one bike h= X can play volleyball e= X is a **female** h= X likes tea h= X likes carrots h= X likes shopping	e= X is a **male** h= X can play poker h= X supports a football team h= X likes beer h= X can play football h= X own (at least) a videogame console h = X can play basketball e= X is a **female** h= X likes ice-figure skating h= X likes candles h= X worked as a babysitter h= X own (at least) one cuddle toy h= X likes reading fashion magazines h= X can dance		e= X is a **male** h= X likes cigars h= X can surf h= X snores e= X is a **female** h= X owns (at least) one plant h= X has freckles h= X owns (at least) one weighting scale
***Imp*(*h*,*e*) = 0**	e= X is a **male** h= X owns (at least) a mp3 player h= X likes going to the cinema e= X is a **female** h= X owns (at least) a mp3 player h= X likes going to the cinema			e = X is a **male** h = X has his/her own website h = X has (at least) 3 siblings e= X is a **female** h = X has his/her own website h = X has (at least) 3 siblings
***Imp*(*h*,*e*) < 0**	e= X is a **female** h= X has a driving licence h= X owns (at least) one bike h= X can play volleyball e= X is a **male** h= X likes tea h= X likes carrots h= X likes shopping		e= X is a **female** h= X can play poker h= X supports a football team h= X likes beer h= X can play football h= X own (at least) a videogame console h= X can play basketball e= X is a **male** h= X likes ice-figure skating h= X likes candles h= X worked as a babysitter h= X own (at least) one cuddle toy h= X likes reading fashion magazines h= X can dance	e= X is a **female** h= X likes cigars h= X can surf h= X snores e= X is a **male** h= X owns(at least) one plant h= X has freckles h= X owns (at least) one weighting scale

Note: the crossed cells represent impossible combinations of probability and impact values.

Appendix B: Scales and questions employed in the experiment

General **INTRODUCTION** (identical for all the tasks):

Consider a group of 200 students, 100 males and 100 females, randomly selected at UCL.

PROBABILITY task (**discrete** scale):

How many of the 100 ***female*** students ***have a driving license***? _________

PROBABILITY task (**continuous** scale):

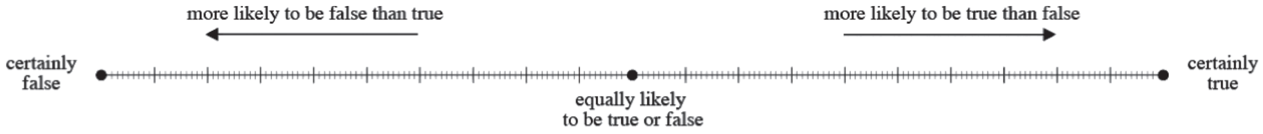
Imagine we draw at random one of these 200 students. Let's call this student A.

Consider the following hypothesis (possibly true or false) concerning A:
A has a driving license.

Now you are given a new piece of information (surely true) concerning A:
A is female.

In the light of this new piece of information (i.e., that *A is female*),
which is the probability of the hypothesis under consideration (i.e., that *A has a driving license*)?

In the light of the information that *A is* ***female***,
the hypothesis that *A* ***has a driving licence*** is


more likely to be false than true
more likely to be true than false
certainly false
equally likely to be true or false
certainly true

IMPACT task (**discrete** scale):

Imagine we draw at random one of these 200 students. Let’s call this student A.

Consider the following hypothesis (possibly true or false) concerning A:
A has a driving license.

Now you are given a new piece of information (surely true) concerning A:
A is female.

How does this new piece of information (i.e., that A is female) affect the hypothesis under consideration (i.e., that A has a driving license)?

Express your opinion indicating a number between
- 50 (“the information maximally weakens the hypothesis”) and + 50 (“the information maximally strengthens the hypothesis”).
Use 0 to indicate no impact at all (“the information does not weaken nor strengthen even a little the hypothesis”).

The impact of the information that A is ***female*** on the hypothesis that A ***has a driving licence*** is: _______

IMPACT task (**continuous** scale):

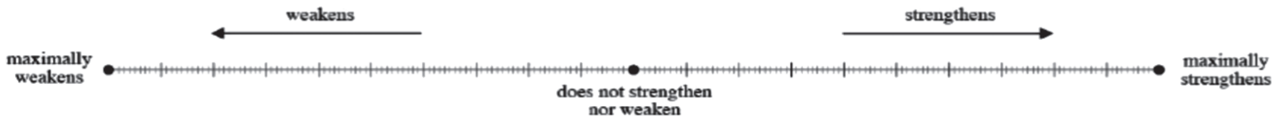
Imagine we draw at random one of these 200 students. Let’s call this student A.

Consider the following hypothesis (possibly true or false) concerning A:
A has a driving license.

Now you are given a new piece of information (surely true) concerning A:
A is female.

How does this new piece of information (i.e., that A is female) affect the hypothesis under consideration (i.e., that A has a driving license)?

The information that *A is **female***



the hypothesis that *A **has a driving licence***