1

WHAT CAN THE CONJUNCTION FALLACY TELL US ABOUT HUMAN REASONING?

In what follows, I will briefly summarize and discuss the main results obtained from more than three decades of studies on the conjunction fallacy (hereafter CF) and will argue that this striking and widely debated reasoning error is a robust phenomenon that can systematically affect laypeople's as much as experts' probabilistic inferences, with potentially relevant real-life consequences. I will then introduce what is, in my view, the best explanation for the CF and indicate how it allows the reconciliation of some classic probabilistic reasoning errors with the outstanding reasoning performances that humans have been shown capable of. Finally, I will tackle the open issue of the greater accuracy and reliability of evidential impact assessments over those of posterior probability and outline how further research on this topic might also contribute to the development of effective human-like computing.

1.1 The conjunction fallacy

When presented with the following scenarios (Tversky and Kahneman, 1983), the great majority (80-90%) of participants ranked the conjunctions ("Linda is a bank teller and is active in the feminist movement" and "Mr. F. has had one or more heart attacks and he is over 55 years old") as more probable than their less-representative constituents ("Linda is a bank teller" and "Mr. F. has had one or more heart attacks"):

> **Linda scenario**
> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. $[e]$
> Please rank the following statements by their probability
>
> - Linda is a teacher in elementary school.
> - Linda works in a bookstore and takes Yoga classes.
> - Linda is active in the feminist movement. $[h_2]$
> - Linda is a psychiatric social worker.
> - Linda is a member of the League of Women Voters.
> - Linda is a bank teller. $[h_1]$
> - Linda is an insurance salesperson.
> - Linda is a bank teller and is active in the feminist movement. $[h_1 \wedge h_2]$
>
> (The order of the response options was randomized.)

> **Mr. F scenario**
> A health survey was conducted in a representative sample of adult males in British Columbia of all ages and occupations. Mr. F. was included in the sample. He was selected by chance from the list of participants.
> Which of the following statements is more probable? (check one)
>
> - Mr. F. has had one or more heart attacks. $[h_1]$
> - Mr. F. has had one or more heart attacks and he is over 55 years old. $[h_1 \wedge h_2]$
>
> (The order of the response options was randomized.)

Similar results have been documented not only with a variety of hypothetical scenarios, but also in many real-life domains, both for laypeople and experts who had been asked for probability judgments in their own fields of specialization (e.g., Frederick and Libby (1986), Ho and Keller (1994), Tversky and Kahneman (1983), Adam and Reyna (2005), Garb (2006), Crupi *et al.* (2018)).

1.2 Fallacy or no fallacy?

From the very first description, judgments such as those described above have been considered violations of "the simplest and the most basic qualitative law of probability" (i.e., the *conjunction rule*, Tversky and Kahneman 1983, p.293; but already mentioned in Tversky and Kahneman 1982, p.90). This emphasis is easy to understand, since the

ordinal comparison between the probability of a conjunction and the probability of one if its conjuncts does not pose a great challenge to cognitive resources and can rest on elementary class-inclusion relationships, without requiring the mastery of formal logic or probability theory. The CF became then a key topic in the fervent debate on human rationality, and a remarkable body of empirical studies inspired by the pragmatics of communication tried to control for whether participants' responses were manifestations of a genuine reasoning error or were, rather, generated by interpretations of the CF stimuli that deprived them of their normative relevance. The evidence provided during this long and heated debate is too extensive to be fully discussed here, however, to give the reader the flavor of it, I will describe the three main candidate misunderstandings and how they have been controlled for.

The first potential misunderstanding concerns participants' interpretation of the verbal descriptions concerning the *isolated conjunct* h_1. Various researchers (e.g., Adler 1984; Dulany and Hilton 1991) have pointed out that the comparison between the relative probability of a set with its superset is anomalous, and widely shared principles of cooperative communication (Grice, 1975) might lead participants to interpret h_1 as $h_1 \wedge \neg h_2$. If this were true, participants' answers could not be evaluated as irrational, even less so than in typical CF scenarios $P(h_1 \wedge \neg h_2) < P(h_1 \wedge h_2)$. Several experimental techniques have been put forward to block such a conversational implicature. Among these, rephrasing of the single conjunct (e.g., "Linda is a bank teller *whether or not she is active in the feminist movement*", emphasis added), controlling for the interpretation of h_1 after the CF task, and above all, changing the set of options offered to participants by explicitly including among the response options the conjunction $h_1 \wedge \neg h_2$ (along with h_1 and conjunction $h_1 \wedge h_2$). The idea, in this case, is that it does not make sense to interpret the conjunct h_1 as $h_1 \wedge \neg h_2$ if the latter option is already available, since, from a conversational point of view, it would be uncooperative to repeat one of the options in a different form. When this technique was applied (as in Tentori *et al.* 2004; Wedell and Moro 2008), the rate of CF was lower than first reported in the original CF scenarios but remained prevalent. Such a pattern makes clear that the misunderstanding of the single conjunct should indeed be avoided in order to distinguish proper and improper fallacy answers, but also that it cannot be considered the primary reason for the occurrence of the CF.

According to a second line of thought (e.g., Gigerenzer 1996; Fiedler 1988), the linguistic misunderstanding between the experimenter and participants concerns the term *probable*. The conjunction rule is not violated, of course, if participants interpret this word not in its technical sense as assigned by modern probability theory, but rather as *plausible, believable, or imaginable* – all legitimate meanings, according to well-respected dictionaries. The vagueness of the term probable in everyday language can be overcome by asking participants to rate the hypotheses according to their "willingness to bet" on them (e.g., Tversky and Kahneman 1983) or, even more directly, by asking participants to bet real money on hypotheses that concern future events (e.g., Sides *et al.* 2002; Bonini *et al.* 2004). The implicit rationale is that participants would be aiming to maximize their winnings, and, in order to do so, they should bet on the most probable hypothesis in the intended mathematical meaning of the word. When this technique has been applied, a drop in the CF with respect the original scenarios

has been observed, but still, most participants committed it.

A final and major objection to the existence of a CF involves the interpretation of the connective *and*. This objection stems from an uncontroversial fact: the conjunction rule concerns the logical connective $\wedge$ while its experimental tests typically rely on a natural language sentential connective *and*, which, as opposed to the former, can reflect various set-theoretical operators and convey a wide range of *temporal* or *causal* relationships between the conjuncts. For example, the word *and* in the sentence "Tom invited friends and colleagues to his party" suggests a *union* rather than an *intersection* of sets, while the *and* in a sentence like "Sara will go to the party, and Mike will be extremely happy" clearly expresses more than the mere co-occurrence of two events. Moving from this premise, a number of authors have argued that responses commonly taken as manifestations of CF might in fact emerge from "reasonable pragmatic and semantic inferences" induced by the ambiguity of the *and* conjunction (Hertwig *et al.* 2008, but see also Gigerenzer 1996). The point is relevant because if the *and* were interpreted as suggesting a union rather than as an intersection operator, or if it were interpreted as indicating a conditional probability instead of the corresponding conjunctive probability (i.e., the probability that h_2 happened given that h_1 did), then there would of course be no fallacy (as already observed by Tversky and Kahneman themselves, 1983). Fortunately, there are various ways to prevent (ex-ante) a misinterpretation of the conjunction or to check (ex-post) whether such a misinterpretation did actually take place. With regard to the former, for example, Bonini *et al.* (2004) overtly point out the conjunctive meaning of *and* by including a reminder of the conjunctive meaning of *and* in the description of the bets that they offered to their participants (it read "both events must happen for you to win the money on this bet"). The control for the interpretation of *and* after the CF task can been accomplished using Venn diagrams (e.g., Tentori and Crupi 2012*b*) or questions that check whether participants hold that the *and* statement at issue implied the truth of both the conjuncts, and therefore, the corresponding $\wedge$ statement (e.g., Tentori *et al.* 2004). Yet again, when these various techniques have been applied, the CF remained prevalent and affected a great number of judgments. The CF has also been proven to be resistant to linguistic training aimed at improving participants' accuracy in distinguishing proper conjunctions from other meanings that may be conveyed by the word *and* (Crandall and Greenfield, 1986).

In summary, although the CF rates reported in the original scenarios (like Linda or Mr. F above) were somewhat inflated, none of the techniques that has been developed to prevent or control for the various sources of misinterpretations mentioned in the literature proved able ultimately to dissipate the effect (for a more comprehensive review on this topic and a similar conclusion, see Moro, 2009). Therefore, we can claim that the CF is a real cognitive bias that can, through careful phrasing of stimuli and with a suitable scenario, easily affect more than 50% of judgments (for a clarification of what makes a good CF scenario, see the following section). The first major point of this chapter is that reasoning errors do not always originate from computational difficulties, inexperience or carelessness. Probabilistic reasoning can exhibit systematic departures from relevant standards of rationality when very simple tasks are at issue and logically correct answers are rewarded, and even in statistically sophisticated individuals.

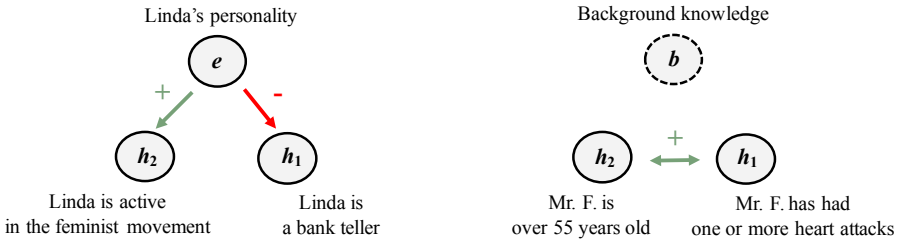

Fig. 1.1 Diagrams representing the Linda (left) and Mr. F (right) scenarios. According to Tversky & Kahneman (1983), in the former case there exists some psychologically salient connection between evidence e and the added conjunct h_2, while in the latter case what is crucial is the relation between the two conjuncts h_1 and h_2.

1.3 Explaining the fallacy

Once the CF is established as a real fallacy, it becomes interesting to explain why people are prone to such an elementary error in reasoning about chance. A good starting point is to observe that only a limited number of comparisons between the probability of a conjunction and that of one of its conjuncts results in a CF. A survey of the literature reveals that the added conjuncts typically employed in successful CF scenarios are both *highly probable* and *positively supported*, as specified by Bayesian theories of confirmation[1] (Carnap, 1962; Crupi and Tentori, 2016; Earman, 1992). Let's consider, for example, the Linda scenario introduced above: the hypothesis that "Linda is active in the feminist movement" (h_2) is probable in light of Linda's description (e), but also positively supported (or inductively confirmed) by this evidence; similarly, in the Mr. F scenario, the hypothesis that "Mr. F is over 55 years old" (h_2) is probable in light of the other hypothesis "Mr. F. has had one or more heart attacks" (h_1) but also positively supported by it. (See Figure 1.1.) The association between posterior probability and confirmation in CF scenarios is not surprising, given that these two variables are also often positively correlated in real life, that is high [low] probability hypotheses are typically confirmed [disconfirmed] by available evidence. However, the posterior probability of a hypothesis and the support for it can be dissociated, so one may wonder which of these two variables is the one crucial for the CF to occur. Tentori *et al.* (2013) designed four experiments to disentangle the perceived probability and confirmation of the added conjuncts in order to contrast them in a CF task, as illustrated by the following scenarios:

Violinist scenario
O. has a degree in violin performance. [e]
Which of the following hypotheses do you think is the most probable?

[1] A common way to formalize the notion of evidential impact (or confirmation) is to devise a function $C(h, e)$ assuming a positive value *iff* $P(h|e) > P(h)$, value zero *iff* $P(h|e) = P(h)$, and a negative value *iff* $P(h|e) < P(h)$. A variety of such functions have been proposed, for example $\log P(e|h)/P(e|\neg h)$ (Good 1984; but see also Fitelson 1999; Crupi *et al.* 2007; Tentori *et al.* 2007).

- O. is an expert mountaineer. [h_1]
- O. is an expert mountaineer and gives music lessons. [$h_1 \wedge h_2$]
- O. is an expert mountaineer and owns an umbrella. [$h_1 \wedge h_3$]

(The order of the response options was randomized.)

Swiss person scenario
Which of the following hypotheses do you think is the most probable?

- M. is Swiss. [h_1]
- M. is Swiss and can ski. [$h_1 \wedge h_2$]
- M. is Swiss and has a driving license. [$h_1 \wedge h_3$]

(The order of the response options was randomized.)

Assuming h_1= "O. is an expert mountaineer" as a (largely irrelevant) piece of background evidence, the majority of participants judged e= "O. has a degree in violin performance" as supporting h_2= "O. gives music lessons" more than h_3= "O. owns an umbrella", that is, they judged $C(h_2, e|h_1) > C(h_3, e|h_1)$. However, they also judged h_2 to be less probable than h_3 (in light of e and h_1), that is, $P(h_2|e \wedge h_1) < P(h_3|e \wedge h_1)$, since almost everybody (even expert mountaineers who give music lessons) owns an umbrella. A similar dissociation was obtained without providing explicit evidence e. For example, the majority of participants judged h_1= "M. is Swiss" as supporting h_2= "M. can ski" more than h_3= "M. has a driving license", that is, they judged $C(h_2, h_1) > C(h_3, h_1)$. At the same time, they judged the overall probability of h_2 given h_1 to be lower than that of h_3, that is, $P(h_2|h_1) < P(h_3|h_1)$. Once the perceived probability of the added conjunct (higher for h_3) and the perceived support for it (stronger for h_2) were dissociated, participants were presented with a CF task in which they had to select the most likely among $h_1, h_1 \wedge h_2$, and $h_1 \wedge h_3$. As a result, Tentori *et al.* (2013) found that a large majority of the fallacious responses targeted $h_1 \wedge h_2$ rather than $h_1 \wedge h_3$ (83% vs. 17% and 79% vs. 21%, respectively for the two above scenarios), a pattern that supported the role of inductive confirmation for the added conjunct rather than its probability as a major determinant of the CF (see also Crupi *et al.* 2008; Tentori and Crupi 2012*a*). This outcome is incompatible with most of the alternative explanations of the CF, from those that ascribe it to *non-normative combination rules* for calculating the conjunctive probability from the probabilities of the two conjuncts (for example, *weighted average*, Fantino *et al.* 1997, *configural weighted average*, Nilsson *et al.* 2009, and *signed summation*, Yates and Carlson 1986), to various *models of rationality rescue*, which consider the CF a consequence of participants' normative patterns of reasoning (for example, *random error variation*, Costello 2009, or *source reliability*, Bovens and Hartmann 2003). Indeed, although very different from each other, all these proposals predict that CF rates would have risen as the perceived probability of the added conjunct increased.

Tversky and Kahneman's (1983) original explanation of the CF by means of the *representativeness* heuristic deserves separate discussion. According to this account, a conjunction (e.g., "Linda is a bank teller and is active in the feminist movement") can appear more probable than one of its constituents ("Linda is a bank teller") because it

is more *representative* of the evidence provided (Linda's description) than the latter. Such a reading of the CF is flexible enough to accommodate a number of findings. Critics (e.g., Gigerenzer 1996) have countered, however, that the notion of representativeness is too vague and imprecisely characterized to serve as a full explanation. It falls short in accounting for the underlying cognitive processes (what drives the representativeness assessment?) and the antecedent conditions that could elicit or suppress it (when should a CF be expected to occur and to what extent?). In reply to these critiques, Tenenbaum and Griffiths (2001) provided a formal account of representativeness in the context of Bayesian inference by quantifying how much evidence e is representative of hypothesis h in terms of the $\log P(e|h)/P(e|\neg h)$. Such a proposal is completely in line with the confirmation-theoretic account of the CF introduced above, since it focuses on the Bayesian support for the hypotheses at issue. Indeed, the logarithm of the likelihood ratio is not only one of the most popular confirmation measures (Fitelson, 1999) but has also been shown to be one of the two measures that best captures people's intuitive judgments of impact (Tentori *et al.*, 2007; Crupi *et al.*, 2007). In this sense, the confirmation explanation of the CF can be seen as a formalization and generalization of the original representativeness account and, as such, could be extended to other phenomena that have been traced to this heuristic, from other reasoning errors (e.g., the base rate fallacy, Kahneman and Tversky 1973), to information retrieval (Gennaioli and Shleifer, 2009), and even to stereotype formation (Bordalo *et al.*, 2016).

1.4 The preeminence of impact assessment over probability judgments

The explanation of the CF provided in the previous section suggests that common probability errors can be determined by a preponderance of evidential reasoning over probabilistic reasoning. In this regard, it is worth noting that people's judgments of evidential impact have been reported to be accurate, both when applied to the evaluation of abstract arguments concerning, for example, urns and balls of different colors (Tentori *et al.*, 2007), and in everyday tasks that require participants to quantify the impact of uncertain evidence (Mastropasqua *et al.*, 2010) or the value of evidence with regard to competing hypotheses (Crupi *et al.*, 2009; Rusconi *et al.*, 2014). These results are consistent with those from the category-based induction literature, according to which adults, and even children as young as 5, when evaluating argument strength, follow popular principles of evidential impact, such as the *similarity* between premises and conclusion and the *diversity* of premises (Heit and Hahn, 2001; Lo *et al.*, 2002; Lopez *et al.*, 1992; Osherson *et al.*, 1990; Zhong *et al.*, 2014). The spontaneous, and often implicit, appreciation of evidential impact has been shown, often under other names, to play a fundamental role in a variety of other higher- and lower-level cognitive processes, including causal induction (Cheng and Novick, 1990; Cheng, 1997), conditional reasoning (Douven and Verbrugge, 2012; Krzyzanowska *et al.*, 2017), learning (Danks, 2003), language processing (Bhatia, 2017; Bullinaria and Levy, 2007; Nadalini *et al.*, 2018; Paperno *et al.*, 2014), and even perception (Mangiarulo *et al.*, 2019).

In light of the aforementioned results, one may wonder whether the updating of the probability of the hypothesis on new evidence and the estimation of the impact

of the new evidence on the credibility of the hypothesis are equally reliable cognitive assessments. Tentori *et al.* (2016) tried to answer this question by directly comparing impact and probability judgments on the very same arguments. More specifically, they asked 200 undergraduates (100 females and 100 males) drawn from various UCL departments to fill in a survey with dozens 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 conditional 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). Fifty-six arguments were than generated 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 hypotheses were selected so as to have (together with the two pieces of evidence) all possible combinations of high/low posterior probability and positive/neutral/negative impact, that is, an identical number of arguments with high ($> .5$) and low ($< .5$) posterior probability of the hypotheses, and, for each of these two classes, the same number of arguments with high, neutral, and low impact. A new sample of participants belonging to the same population (i.e., UCL undergraduates) 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 the 56 arguments generated and were asked to judge, for each of them, the probability of the hypothesis in the light of the evidence provided and the impact of the evidence provided on the credibility of the hypothesis. The results showed that, compared to probability judgments, impact judgments were more consistent over time and more accurate. Impact judgments also predicted the direction of errors in probability judgments.

The conclusions of the studies above converge in suggesting that human inductive reasoning relies more on the estimation of evidential impact than of posterior probability. They also offer a novel approach to bridge the so-called *reality-laboratory gap*, i.e., the alleged clash between the body of experimental work in the heuristics and biases tradition (Tversky and Kahneman, 1974) that has deeply challenged the assumption of people's rationality, and the claims of various evolutionary psychologists, who have argued that it is implausible that humans would have evolved with no "instinct for probability" and, hence, would be "blind to chance" (Pinker, 1997). According to the latter view, reasoning experiments have been designed to "trick our probability calculators," and when people are given "information in a format that meshes with the way they naturally think about probability, they can be remarkably accurate" (Pinker, 1997). Tentori *et al.* (2016) propose a third perspective that reconciles these two views, in that, in dealing with everyday uncertainty, people may appear more rational than in experimental psychology laboratories because they can derive posterior probability from impact. In most situations, indeed, these two kinds of assessments often yield similar results, i.e., when evidence has a strong positive [negative] impact on a hypothesis, then the probability of the latter in the light of the former is rather high [low]. For example, think of a physician who has to diagnose a patient's disease. Usually,

when the available evidence (e.g., symptoms, clinical signs, results of laboratory tests) strongly supports [opposes] the diagnosis of a certain disease, then the probability that the patient has that disease is high [low]. Because this association is so common in real life, one may use impact as a proxy for posterior probability without making critical errors. As shown in the previous section, however, posterior probability and evidential impact can be dissociated, a typical occurrence in classical experimental reasoning tasks, not only those in which the CF has been observed, but also those that showed other well-known fallacies, such as base-rate neglect (in which scenarios the target hypothesis retains a low probability because of its extremely low prior, even in light of supporting evidence). When such a dissociation between probability and confirmation takes place, people cannot derive correct posterior probability judgments from impact and appear to be particularly exposed to biased probability reasoning, whose direction and magnitude seems to depend precisely on perceived impact.

1.5 What suggestions for effective human-like computing?

The greater consistency over time of evidential impact over posterior probability and, above all, its greater accuracy, is of course an empirical finding of interest *per se*, but how can it inform human-like computing?

One of main targets of human-like computing is to improve data mining and machine learning explainability, that is, the process by which intelligent systems explain their outputs to humans, so to generate a shared understanding and, ultimately, increase trust. However, the concept of explanation, despite being a traditional topic in the philosophy of science and a central notion in human reasoning, lacks a unique definition and consensus on the features that would make an explanation "good" or at least "satisfactory". Classical Bayesian confirmation theory makes no explicit reference to explanation, however, the strength of an explanation (i.e., the degree of *explanatory power* of a candidate explanans h relative to its explanandum $e, E(e, h)$) can be expressed in a like manner to the quantification of confirmation, that is, as a function of probability values involving evidence e (for example, a certain symptom) and hypothesis h (for example, a disease that can cause the observed symptom). The connection between explanation and confirmation is not new, and the general idea is that, *ceteris paribus*, the greater the statistical relevance between evidence e and hypothesis h, the greater the strength with which h can explain e (a condition named *positive relevance* by Schupbach and Sprenger 2011; for some well-known probabilistic measures of explanatory power, see Good 1960 and Schupbach and Sprenger 2011).

To appreciate the relationship between confirmation and explanation, it might be of help to refer to the following CF scenario, which has been recently presented to 82 experienced internists (Crupi *et al.*, 2018).

> **Anemia scenario**
> A 50-year-old man from northern Italy has chronic anemia. Currently, the only additional information available comes from a blood exam: hemoglobin 10 g/dL and normal values of leukocytes and platelets. Mean corpuscular volume (MCV) is also in normal range. (Such values are essentially unchanged from a previous test two months back.) [e]

Please consider the following clinical conditions and rank them from the most to the least probable (ties are allowed).

- thalassemia trait [h_1]
- no thalassemia trait and alcoholism [$\neg h_1 \wedge h_2$]
- thalassemia trait and alcoholism [$h_1 \wedge h_2$]
- thalassemia trait and no alcoholism $h_1 \wedge \neg h_2$]
- alcoholism [h_2]

(The order of the response options was randomized.)

The conjunction $h_1 \wedge h_2$ ("thalassemia and alcoholism") was evaluated by 68% of internists as more likely than h_1 ("thalassemia"), by 60% of internists as more likely than h_2 ("alcoholism"), and by 49% of internists as more likely than h_1 as well as more likely than h_2 (i.e, around half of participants committed a *double* CF). These results show once again that experts can make defective probability judgments, which nevertheless rely on a sound intuitive assessment of relations of evidential impact (between diagnostic conditions and clinical signs). Indeed, in the Anemia scenario, each of the two conjuncts is disconfirmed by the available evidence: thalassemia (h_1) because it typically produces low MCV, while alcoholism (h_2) because it typically produces high MCV. However, the conjunction $h_1 \wedge h_2$ is supported by the clinical evidence e (that is $P(h_1 \wedge h_2|e) > P(h_1 \wedge h_2)$) because thalassemia and alcoholism together can *explain* MCV being at normal levels overall. (See Figure 1.2.)

The implications of these results for the debate over explainability are, at least, twofold. First, they offer a new perspective on the factors driving the explanation quality (the so-called "explanatory virtues"). In particular, they challenge the mainstream view (e.g., Lombrozo 2007; Miller 2019) that simpler explanations are judged better and more likely to be true. Indeed, although a conjunction of causes undoubtedly represents a more complex explanation, from both a syntactic and a semantic point of view, than each of the two causes mentioned in the individual conjuncts, most participants ranked the former explanation over the latter two. Therefore, the common

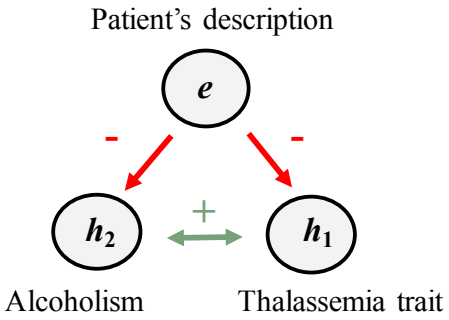

Fig. 1.2 Diagram representing the Anemia scenario. The patient's description disconfirms each of the two conjuncts h_1 and h_2 occurring alone, however, it confirms the conjunction $h_1 \wedge h_2$ (and the two conjuncts confirm each other in light of the evidence provided).

assumption in the data mining and machine learning literature that users always find simpler models easier to understand and more convincing to believe is not necessarily well-founded. For a similar conclusion that questions the predominant "simplicity bias" when the plausibility of rule-based models is at issue (intended as the likeliness that a user accept the model as an explanation for a prediction), see also Fürnkranz *et al.* (2019).

Second, the results of the Anemia scenario suggest that the perceived plausibility of a list of causal explanations does not depend on their probability given the available evidence as much as on their being supported by the available evidence. This is in line with Miller's (2019) claim that "the probability of the explanation being true" is not that important for having a good explanation. Moreover, it embraces one of the main tenets of the inference to the best explanation model (see Lipton 2014), which is the idea that inferences are often guided by explanatory considerations, yet it also suggests an opposing tendency: causal explanations need to rest on strong confirmation relations to be perceived as convincing.[2]

In this regard, it is worth noting that the constructs of confirmation and explanation share a number of interesting properties. To mention one, they both typically involve asymmetric relations: apart from some rare exceptions, if h explains e, the "backward" inference from e to h does not appear equally explanatory; similarly most popular Bayesian confirmation models do not classify inversely symmetric confirmatory arguments as equally strong (i.e., $C(h, e) \neq C(e, h)$, for more details on this, see Eells and Fitelson 2002; Crupi *et al.* 2007). However, the strength of the explanans-explanandum relation between h and e should not be equated to the strength of the impact of e on h. A formal argument supporting this statement is provided by Crupi (2012); for the purposes of this chapter, it is enough to observe that, when there is a positive probabilistic relevance of e to h, the conjunction of e with a piece of evidence x that is probabilistically independent from both e and h, as well as from $e \wedge h$, leaves the degree of confirmation on h unaffected while weakening the explanatory power of h, that is, $C(h, e) = C(h, e \wedge x)$ but $E(e, h) > E(e \wedge x, h)$.

Crupi and Tentori (2012) presented a treatment of the relation between confirmation and explanation, according to which for any e_1, e_2, h_1 and h_2 such that e_1 confirms h_1 (i.e., for which $P(h_1|e_1) > P(h_1)$) and e_2 confirms h_2 (i.e., $P(h_2|e_2) > P(h_2)$) then $C(h_1, e_1) \gtreqless C(h_2, e_2)$ *iff* $E(e_1, \neg h_1) \lesseqgtr E(e_2, \neg h_2)$. Such a principle postulates an inverse (ordinal) correlation between the degree of positive confirmation that a successful explanatory hypothesis h receives from the occurrence of explanandum e and the degree to which e fails to be explained by $\neg h$. In other words, an explanatory hypothesis h is confirmed by evidence e to the extent that the latter appears inexplicable (i.e., a sort of "miracle"), assuming the falsity of the former.

Future empirical studies might quantify more precisely the role of evidential reasoning in the understandability and acceptability of explanations by examining to what

[2]Note that this conclusion concerns only the *necessity* of relevant impact relation(s) for a causal explanation to be perceived as convincing, while it does not mean in any sense that confirmation should be intended as a *sufficient* condition for an explanation to occur. In fact, various statistical relations (for example the association between shapes and colors in a set of figures) allow inductive inferences that do not have an explanatory nature (and for those it seems weird even to talk about a "cause" in the strict sense).

extent equally probable explanations that are supported to various degrees by relevant evidence are perceived to be more or less understandable and convincing. Such an experimental manipulation might be of interest also with respect to the purpose of exploring how explanations are generated. In particular, one of the strongest claims in Miller's (2019) review on explanation in artificial intelligence is that people are "cognitively wired to process contrastive explanations", in the sense that they do not explain the causes for an event *per se* but only relative to some other counterfactual event that did not occur (i.e., an explanation is always of the form "why event e_1 rather than e_2?"). Contrastive explananda are, without a doubt, important for defining the specific *nature* of what has to be explained. However, Crupi and Tentori's (2012) proposal argues that convincing causal explanations make use of an additional type of contrast, the contrast that arises between candidate explanantia. According to this, the value of an explanation would be determined not only by how well it accounts for evidence but also by its "advantage" in doing so over available alternatives. Note that such an idea has a straightforward prediction to offer: when an event equally admits multiple competing explanations, despite the fact that they each account for the event, none would gain any particular credit.

1.6 Conclusion

Cognitive scientists have long been interested in people's systematic violations of basic principles of probability theory, not only because these occurrences detail specific limitations of human reasoning but also because, by elucidating the underlying cognitive processes through which people make judgments, they may offer insight into how to improve the quality of thinking (see on this point the "negative" and "positive" agendas of the heuristics and biases program). In continuity with this tradition, the current chapter reviewed some of the main findings generated in more than 30 years of studies on the CF and to discuss how they can be extended beyond cognitive science by informing human-like computing. To sum up, the results of CF experiments make clear that, when it comes to human reasoning, difficulties do not depend on computational overload or poor statistical numeracy alone. Moreover, the understanding of why laypeople and experts alike commit such an elementary error provides useful information on how inductive inferences are made and shows that the very same cognitive processes that are responsible for errors in one instance allow for gains and accurate performances in another. Finally, what we know about the weaknesses and strengths of probabilistic and evidential reasoning can be used for supporting other thinking processes, and more specifically for developing practical suggestions on how to make explanations understandable and convincing. Future empirical and modelling studies might delve into these proposals and tell us if they are effective in enhancing the reasoning capacity of machines and allowing them to better communicate with humans.

References

Adam, M.B. and Reyna, V.F. (2005). Coherence and correspondence criteria for rationality: Experts' estimation of risks of sexually transmitted infections. *Journal of Behavioral Decision Making*, **18**(3), 169–186.

Adler, J.E. (1984). Abstraction is uncooperative. *Journal for the Theory of Social Behaviour*, **14**(2), 165–181.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, **124**(1), 1–20.

Bonini, N., Tentori, K., and Osherson, D. (2004). A different conjunction fallacy. *Mind and Language*, **19**(2), 199–210.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, **131**(4), 1753–1794.

Bovens, L. and Hartmann, S. (2003). *Bayesian Epistemology*. Oxford University Press, Oxford.

Bullinaria, J.A. and Levy, J.P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, **39**(3), 510–526.

Carnap, R. (1950/1962). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, **104**(2), 367–405.

Cheng, P.W. and Novick, L.R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, **58**(4), 545–567.

Costello, F.J. (2009). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, **22**(3), 213–234.

Crandall, C. and Greenfield, B. (1986). Understanding the conjunction fallacy: A conjunction of effect. *Social Cognition*, **4**(4), 408–419.

Crupi, V. (2012). An argument for not equating confirmation and explanatory power. *The Reasoner*, **6**(3), 39–40. Erratum: The Reasoner, 6, 68.

Crupi, V., Elia, F., Aprà, F., and Tentori, K. (2018, Jun). Double conjunction fallacies in physicians' probability judgment. *Medical Decision Making*, **38**(6), 756–760.

Crupi, V., Fitelson, B., and Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, **14**(2), 182–199.

Crupi, V. and Tentori, K. (2012). A second look at the logic of explanatory power (with two novel representation theorems. *Philosophy of Science*, **79**(3), 365–385.

Crupi, V. and Tentori, K. (2016). Confirmation theory. In *Oxford Handbook of Philosophy and Probability* (ed. A. Hájek and C. Hitchcock), p. 650–665. Oxford University Press.

Crupi, V., Tentori, K., and Gonzalez, M. (2007). On bayesian measures of evidential

support: Theoretical and empirical issues. *Philosophy of Science*, **74**(2), 229–252.
Crupi, V., Tentori, K., and Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, **116**(4), 971–985.
Danks, D. (2003). Equilibria of the rescorla–wagner model. *Journal of Mathematical Psychology*, **47**(2), 109–121.
Douven, I. and Verbrugge, S. (2012). Indicatives, concessives, and evidential support. *Thinking & Reasoning*, **18**(4), 480–499.
Dulany, D.E. and Hilton, D.J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, **9**(1), 85–110.
Earman, J. (1992). *Bayes or Bust?* MIT Press, Cambridge.
Eells, E. and Fitelson, B. (2002). Symmetries and asymmetries in evidential support. *Philosophical Studies*, **107**(2), 129–142.
Fantino, E., Kulik, J., Stolarz-Fantino, S., and Wright, W. (1997). The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin & Review*, **4**(1), 96–101.
Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, **50**(2), 123–129.
Fitelson, B. (1999). The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, **66**, S362–S378.
Frederick, D.M. and Libby, R. (1986). Expertise and auditors judgments of conjunctive events. *Journal of Accounting Research*, **24**(2), 270–290.
Fürnkranz, J., Kliegr, T., and Paulheim, H. (2019). On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, 1–46.
Garb, H.N. (2006). The conjunction effect and clinical judgment. *Journal of Social and Clinical Psychology*, **25**(9), 1048–1056.
Gennaioli, N. and Shleifer, A. (2009). What comes to mind. *The Quarterly Journal of Economics*, **125**(4), 1399–1433.
Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to kahneman and tversky. *Psychological Review*, **103**(3), 592–596.
Good, I.J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society, Series B (Methodological)*, **22**, 319–331.
Good, I.J. (1984). C197. the best explicatum for weight of evidence. *Journal of Statistical Computation and Simulation*, **19**(4), 294–299.
Grice, H.P. (1975, Dec). Logic and conversation. *Speech Acts*.
Heit, E. and Hahn, U. (2001). Diversity-based reasoning in children. *Cognitive Psychology*, **43**(4), 243–273.
Hertwig, R., Benz, B., and Krauss, S. (2008). The conjunction fallacy and the many meanings of and. *Cognition*, **108**(3), 740–753.
Ho, J.L. and Keller, L.R. (1994). The effect of inference order and experience-related knowledge on diagnostic conjunction probabilities. *Organizational Behavior and Human Decision Processes*, **59**(1), 51–74.
Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, **80**(4), 237–251.
Krzyzanowska, K., Collins, P.J., and Hahn, U. (2017). Between a conditional's an-

tecedent and its consequent: Discourse coherence vs. probabilistic relevance. *Cognition*, **164**, 199–205.

Lipton, P. (2014). *Inference to the Best Explanation, 2nd Edition.* London and New York: Routledge.

Lo, Y., Sides, A., Rozelle, J., and Osherson, D. (2002). Evidential diversity and premise probability in young childrens inductive judgment. *Cognitive Science*, **26**(2), 181–206.

Lombrozo, T (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, **55**(3), 232–257.

Lopez, A., Gelman, S.A., Gutheil, G., and Smith, E.E. (1992). The development of category-based induction. *Child Development*, **63**(5), 1070.

Mangiarulo, M., Pighin, S., Polonio, L., and Tentori, K. (2019). The effect of evidential impact on perceptual probabilistic judgments. Under review.

Mastropasqua, T., Crupi, V., and Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence. *Memory & Cognition*, **38**(7), 941–950.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, **267**, 1–38.

Nadalini, A., Marelli, M., Bottini, R., and Crepaldi, D. (2018). Local associations and semantic ties in overt and masked semantic priming. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, p. 283–287.

Nilsson, H., Winman, A., Juslin, P., and Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General*, **138**(4), 517–534.

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

Paperno, D., Marelli, M., Tentori, K., and Baroni, M. (2014). Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood. *Cognitive Psychology*, **74**, 66–83.

Pinker, S. (1997). *How the Mind Works.* Norton, New York.

Rusconi, P., Marelli, M., D'Addario, M., Russo, S., and Cherubini, P. (2014). Evidence evaluation: Measure z corresponds to human utility judgments better than measure l and optimal-experimental-design models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **40**(3), 703–723.

Schupbach, J.N. and Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science*, **78**(1), 105–127.

Sides, A., Osherson, D., Bonini, N., and Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, **30**(2), 191–198.

Tenenbaum, J.B. and Griffiths, T.L. (2001). The rational basis of representativeness. In *Proceedings of 23rd Annual Conference of the Cognitive Science Society* (ed. J. Moore and K. Stenning), p. 1036–1041.

Tentori, K., Bonini, N., and Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, **28**(3), 467–477.

Tentori, K., Chater, N., and Crupi, V. (2016). Judging the probability of hypotheses versus the impact of evidence: Which form of inductive inference is more accurate

and time-consistent? *Cognitive Science*, **40**(3), 758–778.

Tentori, K. and Crupi, V. (2012*a*). How the conjunction fallacy is tied to probabilistic confirmation: Some remarks on schupbach (2009). *Synthese*, **184**(1), 3–12.

Tentori, K. and Crupi, V. (2012*b*). On the conjunction fallacy and the meaning of and, yet again: A reply to hertwig, benz, and krauss (2008). *Cognition*, **122**(2), 123–134.

Tentori, K., Crupi, V., Bonini, N., and Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, **103**(1), 107–119.

Tentori, K., Crupi, V., and Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, **142**(1), 235–255.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**(4157), 1124–1131.

Tversky, A. and Kahneman, D. (1982). Judgments of and by representativeness. In *Judgment under uncertainty: Heuristics and biases* (ed. D. Kahneman, P. Slovic, and A. Tversky), pp. 84–98. Cambridge University Press.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, **90**(4), 293–315.

Wedell, D.H. and Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, **107**(1), 105–136.

Yates, J.F. and Carlson, B.W (1986). Conjunction errors: Evidence for multiple judgment procedures, including "signed summation". *Organizational Behavior and Human Decision Processes*, **37**(2), 230–253.

Zhong, L., Lee, M.S., Huang, Y., and Mo, L. (2014). Diversity effect in category-based inductive reasoning of young children: Evidence from two methods. *Psychological Reports*, **114**(1), 198–215.