Alan Author strikes again: more on confirming conjunctions of disconfirmed hypotheses

Jakob Koscholke

Abstract

The so-called *Alan Author Effect* is a surprising phenomenon in Bayesian Confirmation Theory. It occurs when a piece of evidence e confirms the conjunction of two hypotheses $h_1 \wedge h_2$ but at the same time *disconfirms* each hypothesis h_1 and h_2 individually. In this paper, I present a new and prima facie stronger version of this effect where additionally, the evidence e confirms the conjunction of the *negated* hypotheses $\neg h_1 \wedge \neg h_2$. I say 'prima facie' because it can be shown that this seemingly stronger effect and the original effect are actually *coextensional*. I use this insight to formulate a new sufficient (and also necessary) condition for the two equivalent effects. I also examine how likely the two effects are to occur with the help of Monte Carlo simulation methods.

Keywords: Bayesian Confirmation Theory, probability, conjunction, Monte Carlo simulation

1. Introduction

The so-called *Alan Author Effect* is a somewhat surprising phenomenon in Bayesian Confirmation Theory. Its name is due to Atkinson et al. (2009), but the phenomenon was already noticed by Douven (2007) and in fact anticipated much earlier by Salmon (1983) and Carnap (1950: 382). It occurs when a piece of evidence e confirms the conjunction of two hypotheses $h_1 \wedge h_2$ but at the same time *disconfirms* each hypothesis h_1 and h_2 individually. To get a better grip on this phenomenon, consider Alan Author's story:

> Alan Author has just made an important discovery. From his calculations it follows that recent evidence e supports the conjunction of two popular hypotheses, h_1 and h_2. With great gusto he sets himself to the writing of a research proposal in which he explains his idea and asks for time and money to work out all its far-reaching consequences. Alan Author's proposal is sent to Rachel Reviewer, who – to his dismay – writes a devastating report. Ms. Reviewer first recalls what is common knowledge within the scientific community, namely that e strongly disconfirms not only h_1, but also h_2 as well. Then she intimates that Alan Author is clearly not familiar with the relevant literature; for if he were, he would have realized that any calculation that results in confirming

Analysis Vol. XX | Number XX | XX 2024 | 1–11 doi: https://doi.org/10.1093/analys/anad103

> the conjunction of two disconfirmed hypotheses must contain a mistake. (Atkinson et al. 2009: 1)[1]

One might think that Rachel Reviewer is right here. How could a piece of evidence confirm a conjunction if it is known to disconfirm its two conjuncts? Or as Künne puts it in a slightly different context:

> How could there fail to be evidence for either of the conjuncts if there is evidence for the conjunction as a whole? (Künne 2003: 438)

But surprisingly, Rachel Reviewer is at fault: it is very well possible for a piece of evidence e to confirm a conjunction $h_1 \wedge h_2$ while disconfirming each of the two hypotheses h_1 and h_2 individually. That is, more formally, there are probability functions P and variables e, h_1 and h_2 such that the following inequality holds,

$$P(h_1 \wedge h_2|e) > P(h_1 \wedge h_2), \tag{1}$$

while at the same time the following two inequalities hold:

$$P(h_1|e) < P(h_1) \quad \text{and} \quad P(h_2|e) < P(h_2). \tag{2}$$

A proof that such probability functions exist is given by the distribution shown in Table 1, a so-called *stochastic truth table* (for details see Fitelson 2008). Let us follow Atkinson et al. in calling any probability function P under which (1) and (2) are jointly satisfied an instance of the *Alan Author Effect*.[2]

Table 1 Instance of the Alan Author Effect

e	h_1	h_2	P
0	0	0	1/16
0	0	1	2/16
0	1	0	2/16
0	1	1	2/16
1	0	0	2/16
1	0	1	2/16
1	1	0	2/16
1	1	1	3/16

1 Atkinson et al. (2009) not only discuss the Alan Author Effect, but also extend a formal result by Crupi et al. (2007a) that supports Sides et al.'s (2002) confirmation-theoretic analysis of Kahneman et al.'s (1982: 32-47) famous conjunction fallacy. More precisely, Atkinson et al. show that the condition $P(h_1 \wedge \neg h_2|e) = P(\neg h_1 \wedge h_2|e) = 0$ guarantees that, according to most Bayesian confirmation measures, the degree to which the evidence e confirms the conjunction of two hypotheses $h_1 \wedge h_2$ will be *at least as high* as the degree to which e confirms each of the two hypotheses h_1 and h_2 individually.

2 Less formally, instances of the Alan Author Effect where (1) and (2) are jointly satisfied show that Bayesian confirmation violates conjunction elimination, or equivalently that

Now, the mere possibility of this effect is perhaps already surprising enough for some readers. In this paper, however, I would like to present an even stronger, or at least prima facie stronger version of the Alan Author Effect. I will describe this previously unnoticed and perhaps even more surprising effect in §2 and show that it can in fact occur. In §3 I show that this new effect and the original Alan Author Effect are actually *coextensional*. And, after providing a new sufficient condition for the two equivalent effects in §4, in §5 I study how likely the two effects are to arise. I conclude in §6.

2. *Strengthening the effect*

Consider the following variation of Alan Author's story:

> Alan Author has made another important discovery. This time, it follows from his calculations that recent evidence e supports the conjunction of two other, similarly popular hypotheses $h_1 \wedge h_2$. He writes another research proposal which again is sent to Rachel Reviewer. Reviewer writes another devastating report. She recalls that it is not only common knowledge that e disconfirms h_1 and h_2 *individually*, and thus confirms the negated hypotheses $\neg h_1$ and $\neg h_2$ individually, but that it is also common knowledge that e actually confirms the *conjunction* of the *negated* hypotheses $\neg h_1 \wedge \neg h_2$. Again, she intimates that Alan Author is not familiar with the relevant literature. For if he were, so she argues, he would have realized that since the evidence e *disconfirms* h_1 and h_2 individually and additionally, e confirms the conjunction of the *opposite* of what the hypotheses state, i.e. $\neg h_1 \wedge \neg h_2$, it simply cannot be that e supports the conjunction of the two hypotheses $h_1 \wedge h_2$.

Here it seems that, in addition to the two Alan Author Effect conditions (1) and (2), the following inequality is also satisfied:

$$P(\neg h_1 \wedge \neg h_2 | e) > P(\neg h_1 \wedge \neg h_2). \tag{3}$$

And one might think that Rachel Reviewer must be right *this time*. How on earth could the conjunction $h_1 \wedge h_2$ be confirmed by a piece of evidence e

disconfirmation violates conjunction introduction. Notice that such violations are *not* possible if confirmation is understood as *high conditional probability*: it cannot be that $P(h_1 \wedge h_2|e) \geq \theta$ while $P(h_1|e) < \theta$ or $P(h_2|e) < \theta$, where θ is the threshold for sufficiently high conditional probability. This is a simple consequence of the Boole–Fréchet inequalities (see Fréchet 1935).

that is known to speak against h_1 and h_2 in *multiple respects*, namely individually and conjunctively? But Rachel Reviewer is at fault *again*.

Let us extend our previous terminology and call any probability function P that jointly satisfies (1), (2) and (3) an instance of the *Strong Alan Author Effect*. If one already finds the possibility of the original Alan Author Effect surprising, then one should be similarly surprised about the possibility of this new effect. For a proof that this effect is possible, that is, that there are probability functions P and variables e, h_1 and h_2 such that the conditions (1), (2) and (3) are jointly satisfied, inspect again the probability model shown in Table 1. Here we have the following probabilities:

- $P(h_1|e) = \dfrac{5}{9} < P(h_1) = 9/16,$
- $P(h_2|e) = \dfrac{5}{9} < P(h_2) = 9/16,$
- $P(\neg h_1 \wedge \neg h_2|e) = \dfrac{2/16}{9/16} > P(\neg h_1 \wedge \neg h_2) = 3/16,$
- $P(h_1 \wedge h_2|e) = \dfrac{3/16}{9/16} > P(h_1 \wedge h_2) = 5/16.$

So Alan Author did it again. His latest discovery teaches us that, even if a piece of evidence e disconfirms two hypotheses h_1 and h_2 individually and the conjunction of the negated hypotheses $\neg h_1 \wedge \neg h_2$, it is still possible that e confirms the conjunction $h_1 \wedge h_2$.

In the next section, we will see that this is not yet the whole story. There is an interesting and previously unnoticed logical relationship between the conditions characterizing the original and the Strong Alan Author Effect.

3. *The original is strong*

Recall our Alan Author conditions discussed so far:

$$P(h_1 \wedge h_2|e) > P(h_1 \wedge h_2), \tag{1}$$

$$P(h_1|e) < P(h_1) \text{ and } P(h_2|e) < P(h_2), \tag{2}$$

$$P(\neg h_1 \wedge \neg h_2|e) > P(\neg h_1 \wedge \neg h_2). \tag{3}$$

Conditions (1) and (2) are the original Alan Author conditions, while condition (3) is what sets apart the Strong Alan Author Effect from the original

Alan Author Effect. Thus, rather trivially, (3) is a *necessary* condition for the Strong Alan Author Effect: no *Strong* Alan Author Effect without the condition that actually makes it *strong*.

Now, one might think that if the Alan Author Effect is not strong, it can at least be an instance of the *original* Alan Author Effect. But surprisingly, this turns out to be false: condition (3) is also a necessary condition for the original Alan Author Effect. To see this, assume for *reductio* that there is a probability function P over e, h_1 and h_2 such that the two original Alan Author Effect conditions (1) and (2) are true while the Strong Alan Author Effect condition (3) is false. The falsity of (3) is equivalent to

$$P(h_1 \vee h_2 | e) \geq P(h_1 \vee h_2), \tag{4}$$

which by general additivity expands to

$$P(h_1|e) + P(h_2|e) - P(h_1 \wedge h_2|e) \geq P(h_1) + P(h_2) - P(h_1 \wedge h_2). \tag{5}$$

And, by simple algebra, this is equivalent to

$$P(h_1|e) - P(h_1) + P(h_2|e) - P(h_2) \geq P(h_1 \wedge h_2|e) - P(h_1 \wedge h_2), \tag{6}$$

which together with our assumption (1) and transitivity entails

$$P(h_1|e) - P(h_1) + P(h_2|e) - P(h_2) > 0. \tag{7}$$

But together with our assumption (2) that both summands are negative, this yields a contradiction, which completes our *reductio*. Thus, whenever conditions (1) and (2) are satisfied, condition (3) is satisfied too.

Less formally, this means that any instance of the Alan Author Effect is already an instance of the Strong Alan Author Effect. And since the other implication direction holds trivially, this means that the original Alan Author Effect and the Strong Alan Author Effect are actually *coextensional*: whenever one of them occurs, the other occurs as well and vice versa. In the next section, I examine under which circumstances these two equivalent effects occur.[3]

4. *Another sufficient condition*

Atkinson et al. (2009) not only gave the Alan Author Effect its unique name, but also provided a condition under which the effect occurs. Given the

3 The two Alan Author conditions (1) and (2) entail not only the Strong Alan Author condition (3), but also $P(h_1 \wedge \neg h_2|e) < P(h_1 \wedge \neg h_2)$ and $P(\neg h_1 \wedge h_2|e) < P(\neg h_1 \wedge h_2)$. This conjunction states that e disconfirms that one of the hypotheses h_1 and h_2 is true while the other is false, which makes the Alan Author Effect even more peculiar.

schematic probability distribution shown in Table 2, they showed that the following conjunction is sufficient for the original Alan Author conditions (1) and (2) and thus also for the Strong Alan Author condition (3):

$$\frac{1-2z}{6} < x < \min\left(\frac{1-2z}{5}, \frac{1-4z}{4}\right) \quad \text{and} \quad 0 \leq z \leq \frac{1}{8}. \tag{8}$$

Douven (2007) also provided a sufficient condition for the Alan Author Effect. For any natural number n, the schematic probability distribution shown in Table 2 will satisfy the two Alan Author conditions (1), (2) and thus the Strong Alan Author condition (3).

Table 2 Schematic distributions from Atkinson et al. 2009 and Douven 2007

e	h_1	h_2	Atkinson et al.	Douven
0	0	0	$1-(5x+2z)$	0
0	0	1	x	$1/3\,(n+2)$
0	1	0	x	$1/3\,(n+2)$
0	1	1	x	$n/3\,(n+2)$
1	0	0	x	$3/3\,(n+2)$
1	0	1	z	0
1	1	0	z	0
1	1	1	x	$(2n+1)/3\,(n+2)$

Atkinson et al.'s and Douven's conditions are of course very useful for *generating* probability functions that instantiate the Alan Author Effect. But they are somewhat difficult to *interpret*: the schematic probability distributions in Table 2 provide rather complicated answers to the question under which circumstances the Alan Author Effect occurs. I would therefore like to provide a sufficient condition for the effect that is easier to interpret.

Let me start by saying that the Strong Alan Author condition (3) itself cannot serve as such a condition: it is simply not enough that the evidence e supports $\neg h_1 \wedge \neg h_2$. But if it does so *to a certain degree*, then the Alan Author Effect arises. In order to state more precisely to which degree e must support $\neg h_1 \wedge \neg h_2$, I will make use of a well-known probabilistic measure of confirmation, namely Carnap's (1950: 362) so-called relevance measure $\mathfrak{r}$:[4]

$$\mathfrak{r}(h,e) = P(h \wedge e) - P(h)\,P(e). \tag{9}$$

4 Notice that Carnap's difference measure $d(h,e) = P(h|e) - P(h)$ (1950: 361), or even other Bayesian confirmation measures, could be used to make the same point (see Crupi et al. 2007b).

This measure is larger than 0 if e confirms h, equal to 0 if e is irrelevant for h and smaller than 0 if e disconfirms h. Given this measure, the new sufficient condition for the Alan Author Effect can be stated as follows:

$$\mathfrak{r}(h_1, e) + \mathfrak{r}(h_2, e) + \mathfrak{r}(\neg h_1 \wedge \neg h_2, e) > 0 \\ \text{where } \mathfrak{r}(h_1, e), \mathfrak{r}(h_2, e) < 0. \tag{10}$$

Less technically, this condition tells us the Alan Author Effect will occur whenever the evidence e confirms $\neg h_1 \wedge \neg h_1$ to a degree that trumps the summed negative degrees of support that e provides for h_1 and h_2 individually. To see that (10) is in fact sufficient for the Alan Author Effect, first observe that $\mathfrak{r}$has the following additivity property that is analogous to general additivity for conditional probability and that was already proven by Carnap (1950: 370):

$$\mathfrak{r}(h_1 \wedge h_2, e) = \mathfrak{r}(h_1, e) + \mathfrak{r}(h_2, e) - \mathfrak{r}(h_1 \vee h_2, e). \tag{11}$$

Carnap (1950: 364) also showed that $\mathfrak{r}$ has a symmetry property that Eells and Fitelson (2002) later called *hypothesis symmetry*, a property that is widely endorsed in the confirmation theory literature (see Crupi 2015):

$$\mathfrak{r}(h_1 \vee h_2, e) = -\mathfrak{r}(\neg (h_1 \vee h_2), e) = -\mathfrak{r}(\neg h_1 \wedge \neg h_2, e). \tag{12}$$

This symmetry property is helpful in our context because from (11) and (12) we see that the conjunction $h_1 \wedge h_2$ will be confirmed by e whenever the following inequality is satisfied:

$$\mathfrak{r}(h_1, e) + \mathfrak{r}(h_2, e) + \mathfrak{r}(\neg h_1 \wedge \neg h_2, e) > 0. \tag{13}$$

In Alan Author cases, the first two summands will be negative, and so will be their sum. But as long as the third summand exceeds the sum of their negative values, the total sum will become positive and hence, e will confirm the conjunction $h_1 \wedge h_2$. All of this is nicely captured by condition (10). And notice that (10) is also a necessary condition for the Alan Author Effect. For if (10) is false, then either e does not disconfirm h_1 or h_2 and thus condition (2) is violated, or the inequality (13) and thus condition (1) are violated. In both cases, the Alan Author Effect can no longer occur.

5. *Rachel Reviewer's reasoning*

In our first story, Rachel Reviewer argued that since it is common knowledge that the evidence e disconfirms the two popular hypotheses h_1 and h_2 individually, it cannot confirm their conjunction $h_1 \wedge h_1$. And, in our second story, she argued that e cannot confirm the conjunction $h_1 \wedge h_1$ because e is known to disconfirm h_1 and h_2 individually and to confirm

$\neg h_1 \wedge \neg h_2$. In both cases, Rachel Reviewer was wrong. But was her reasoning *irrational*? Intuitively, it does not seem so: who would expect that e can still confirm the conjunction $h_1 \wedge h_2$? So, in defence of Rachel Reviewer, one might wonder how *likely* it actually is that e confirms $h_1 \wedge h_2$ in both stories.

Fortunately, this question can be addressed with the help of a computational method known as Monte Carlo integration (see Metropolis and Ulam 1949). Instead of determining the prevalence of our two target effects *analytically*, we can do this *numerically* by first generating a sufficiently large number of probability functions over an algebra generated by three variables and by then calculating the proportion of functions that instantiate our two target effects. Let us call the proportion of probability functions that *jointly* satisfy conditions (1) and (2) the *conjunctive* prevalence of the Alan Author Effect, and the proportion of functions that satisfy condition (1) *among* the functions that satisfy condition (2) the *conditional* prevalence of the effect. Since the Alan Author Effect and the Strong Alan Author Effect are coextensional, their conjunctive prevalence will be identical. However, their conditional prevalence values can diverge: for the Alan Author Effect, the value is a proportion of probability functions that satisfy (2), while for the Strong Alan Author Effect, it is a proportion of probability functions that jointly satisfy (2) *and* (3).

Our integration is based on 10 million regular probability distributions P over the algebra generated by three variables e, h_1 and h_2. The distributions were generated randomly using the Mersenne Twister algorithm (see Matsumoto and Nishimura 1998), which corresponds to sampling from a uniform Dirichlet distribution on the unit 8-simplex with concentration parameters $\alpha_1, \ldots, \alpha_8$ equal to 1 (see Kotz et al. 2000). As Figure 1 shows, the results are already quite stable for 10,000 distributions. And, more importantly, they vindicate Rachel Reviewer: Alan Author cases and thus Strong Alan Author cases are *very rare* with a conjunctive prevalence of around 2.5%. And the conditional prevalence of the Alan Author Effect and the Strong Alan Author Effect is also very low, namely 9.99% and 11.1% respectively. So Rachel Reviewer's reasoning is far from being irrational: in both stories, she was *much* more likely to be right than to be wrong in thinking that the evidence e does not confirm the conjunction $h_1 \wedge h_2$, namely around 90% in the first story and around 89% in the second.[5]

5 An anonymous referee suggested that it would be good to have *analytical* instead of merely *numerical* results for the prevalence of the two effects. I agree. But for probability functions over an algebra generated by three variables, the required integrals are already too difficult to work out analytically. See also Wagner (2013) or Douven and Rott (2018), who used Monte Carlo methods for the same reason.

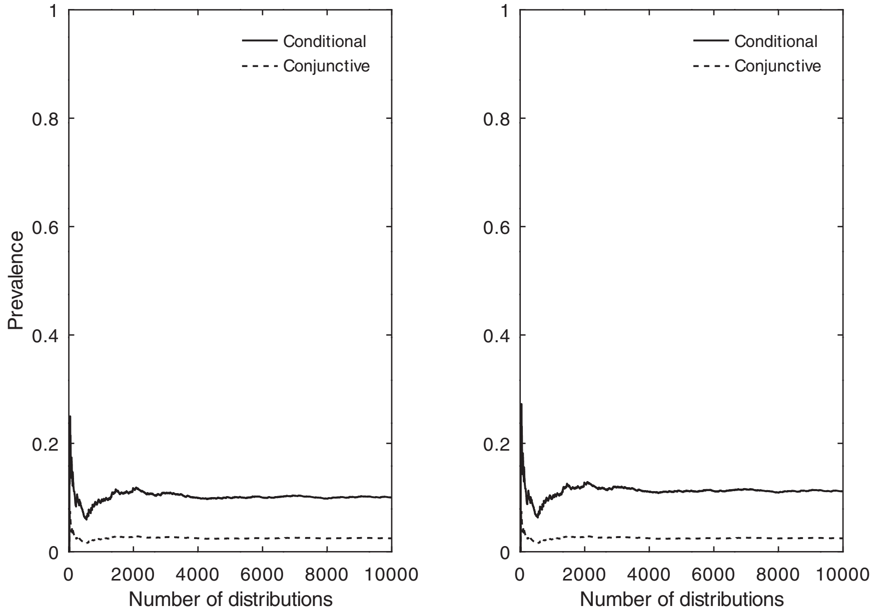

Figure 1 Alan Author Effect (left) and Strong Alan Author Effect (right).

6. *Conclusion*

Many scholars have emphasized that probabilistic reasoning is slippery and that our intuitions regarding confirmation and disconfirmation are often unreliable:[6]

> the superficial appearance of plausibility in inductive logic is very often entirely misleading. (Carnap 1950: 368)

> these relationships do not behave in a systematic or 'orderly' fashion, and fail to fulfill several natural expectations. (Falk and Bar-Hillel 1983: 240)

> just about all of one's naïve expectations regarding positive relevance turn out to be false. (Wagner 2013: 1456)

6 An anonymous referee pointed out an interesting clinical reasoning study by Crupi et al. (2018). The study includes a scenario in which a piece of medical evidence e disconfirms two medical hypotheses individually, namely thalassemia h_1 and alcoholism h_2, but at the same time confirms their conjunction $h_1 \wedge h_2$. This is of course an Alan Author scenario. And, interestingly, the authors of the study observe a striking kind of double conjunction fallacy in this scenario: about 50% of expert physicians judge the conjunction $h_1 \wedge h_2$ more probable than each of the conjuncts h_1 and h_2. This underlines that the Alan Author Effect is not just a formal phenomenon, but has substantial real-world implications.

The Alan Author Effect is another case in point: one might naïvely expect that a piece of evidence e disconfirms a conjunction of two hypotheses $h_1 \wedge h_2$ if e is known to disconfirm the conjuncts h_1 and h_2 individually, and even more so if additionally e is known to confirm $\neg h_1 \wedge \neg h_2$. But this expectation is false. However, this does not mean that it is irrational: Alan Author cases and thus Strong Alan Author cases are very rare. And a piece of evidence e is very unlikely to confirm a conjunction $h_1 \wedge h_2$ if it already disconfirms h_1 and h_2 individually, and also if additionally e confirms $\neg h_1 \wedge \neg h_2$.[7]

Funding

This work was funded by the German Research Foundation (DFG) as part of grant ZI 1824/1-1 to Alexandra Zinke.

Goethe-University Frankfurt
Germany
Koscholke@em.uni-frankfurt.de

References

Atkinson, D., J. Peijnenburg and T. Kuipers. 2009. How to confirm the conjunction of disconfirmed hypotheses. *Philosophy of Science* 76: 1–21.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Crupi, V. 2015. Inductive logic. *Journal of Philosophical Logic* 44: 641–50.

Crupi, V., F. Elia, F. Aprà and K. Tentori. 2018. Double conjunction fallacies in physicians' probability judgment. *Medical Decision Making* 38: 756–60.

Crupi, V., B. Fitelson and K. Tentori. 2007a. Probability, confirmation, and the conjunction fallacy. *Thinking and Reasoning* 14: 182–99.

Crupi, V., K. Tentori and M. Gonzalez. 2007b. On Bayesian measures of evidential support: theoretical and empirical issues. *Philosophy of Science* 74: 229–52.

Douven, I. 2007. Fitch's paradox and probabilistic antirealism. *Studia Logica* 86: 149–82.

Douven, I. and H. Rott. 2018. From probabilities to categorical beliefs: going beyond toy models. *Journal of Logic and Computation* 28: 1099–124.

Eells, E. and B. Fitelson. 2002. Symmetries and asymmetries in evidential support. *Philosophical Studies* 107: 129–42.

Falk, R. and M. Bar-Hillel. 1983. Probabilistic dependence between events. *Two-Year College Mathematics Journal* 14: 240–7.

Fitelson, B. 2008. A decision procedure for probability calculus with applications. *Review of Symbolic Logic* 1: 111–25.

Fréchet, M. 1935. Généralisations du théorème des probabilités totales. *Fundamenta Mathematica* 25: 379–87.

7 I would like to thank Jeanne Peijnenburg, David Atkinson and Branden Fitelson for helpful feedback. Special thanks to Krista Lawlor for sponsoring my Visiting Scholarship to Stanford University in 2022.

Kahneman, D., P. Slovic and A. Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kotz, S., N. Balakrishnan and N.L. Johnson. 2000. *Continuous Multivariate Distributions*. New York: Wiley.

Künne, W. 2003. *Conceptions of Truth*. Oxford: Oxford University Press.

Matsumoto, M. and T. Nishimura. 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation* 8: 3–30.

Metropolis, N. and S. Ulam. 1949. The Monte Carlo method. *Journal of the American Statistical Association* 44: 335–41.

Salmon, W.C. 1983. Confirmation and relevance. In *The Concept of Evidence*, eds. P. Achinstein and P. Achinstein, 3-36. Oxford: Oxford University Press.

Sides, A., D. Osherson, N. Bonini and R. Viale. 2002. On the reality of the conjunction fallacy. *Memory and Cognition* 30: 191–8.

Wagner, C.G. 2013. The corroboration paradox. *Synthese* 190: 1455–69.