Chapter 6

Confirmation

When evidence supports a hypothesis, philosophers of science say that the evidence "confirms" that hypothesis. Bayesians place this confirmation relation at the center of their theory of induction. But confirmation is also closely tied to such epistemological notions as justification and reasons. Bayesian Epistemology offers a systematic theory of confirmation (and its opposite, disconfirmation) that not only deepens our understanding of the relation but also provides specific answers about which hypotheses are supported in particular on-the-ground evidential situations.

Since its early days the analysis of confirmation has been driven by a perceived analogy to deductive entailment. In Chapter 4 we discussed "evidential standards" that relate a body of evidence (represented as a proposition) to the doxastic *attitudes* it supports. But confirmation—though intimately linked with evidential standards in ways we'll see shortly—is a different kind of relation: instead of relating proposition and attitude, it relates two propositions (evidence and hypothesis) themselves. It shares this feature with deductive entailment. Carnap, in fact, thought of confirmation as a generalization of standard logical relations, with deductive entailment and refutation as two extremes of a continuous confirmational scale.

In the late nineteenth and early twentieth centuries, logicians produced ever-more-powerful syntactical theories capable of answering specific questions about which propositions deductively entailed which. Impressed by this progress, theorists such as Hempel and Carnap envisioned a syntactical theory that would do the same for confirmation. As Hempel put it,

> The theoretical problem remains the same: to characterize, in precise and general terms, the conditions under which a body of evidence can be said to confirm, or to disconfirm, a hypothesis

> of empirical character. (1945a, p. 7)

Hempel identified various formal properties that the confirmation relation might or might not possess. Carnap then argued that we get a confirmation relation with exactly the right formal properties by identifying confirmation with positive probabilistic relevance.

This chapter begins with Hempel's formal conditions on the confirmation relation. Identifying the right formal conditions for confirmation will not only help us assess various theories of confirmation; it will also help us understand exactly what relation philosophers of science have in mind when they talk about "confirmation".[1] We then move on to Carnap's Objective Bayesian theory of confirmation, which roots confirmation in probability theory. While Carnap's theory has a number of attractive features, we will also identify two drawbacks: its failure to capture particular patterns of inductive inference Carnap found appealing; and the language-dependence suggested by Goodman's "grue" problem. We'll respond to these problems with a confirmation theory grounded in Subjective Bayesianism (in the normative sense).

Confirmation is fairly undemanding, in one sense: we say that evidence confirms a hypothesis when it provides *any* amount of support for that hypothesis, no matter how small. Probabilistic theories offer the advantage that when the confirmation relation obtains in a particular case, we can measure how strong the relation of support is. We will survey different measures of confirmational strength, assessing the pros and cons of each. Finally, we'll apply the probabilistic theory of confirmation to provide a Bayesian solution to Hempel's Paradox of the Ravens.

6.1 Formal features of the confirmation relation

6.1.1 Confirmation is weird! The Paradox of the Ravens

One way to begin thinking about confirmation is to consider the simplest possible cases in which a piece of evidence confirms a general hypothesis. For example, the proposition that a particular frog is green seems to confirm the hypothesis that all frogs are green. On the other hand, the proposition that a particular frog is not green disconfirms the hypothesis that all frogs are green. (In fact, it *refutes* that hypothesis!) If we think this pattern always holds, we will maintain that confirmation satisfies the following constraint:

Nicod's Criterion: For any predicates F and G and constant a of $\mathcal{L}$, $(\forall x)(Fx \supset Gx)$ is confirmed by $Fa \,\&\, Ga$ and disconfirmed by $Fa \,\&\, {\sim}Ga$.

Carl Hempel (1945a,b) names this condition after Jean Nicod (1930), who built his theory of induction around the criterion.

Yet Hempel worries about the Nicod Criterion, because of how it interacts with another principle he endorses:

Equivalence Condition (for hypotheses): Suppose H and H' in $\mathcal{L}$ are logically equivalent ($H \dashv\models H'$). Then any E in $\mathcal{L}$ that confirms H also confirms H'.

Hempel endorses the Equivalence Condition because he doesn't want confirmation to depend on the particular way a hypothesis is formulated; logically equivalent hypotheses say the same thing, so they should enter equally into confirmation relations. Hempel is also concerned with how working scientists *use* confirmed hypotheses; for instance, practitioners will often deduce predictions and explanations from confirmed hypotheses. Equivalent hypotheses have identical deductive consequences, and scientists don't hesitate to substitute logical equivalents for each other.

But combining Nicod's Criterion with the Equivalence Condition yields counterintuitive consequences, which Hempel calls the "paradoxes of confirmation". The most famous of these is the **Paradox of the Ravens**. Consider the hypothesis that all ravens are black, representable as $(\forall x)(Rx \supset Bx)$. By Nicod's Criterion this hypothesis is confirmed by the evidence that a particular raven is black, $Ra \,\&\, Ba$. But now consider the evidence that a particular non-raven is non-black, ${\sim}Ba \,\&\, {\sim}Ra$. By Nicod's Criterion this evidence confirms the hypothesis $(\forall x)({\sim}Ba \supset {\sim}Ra)$. By contraposition, this hypothesis is equivalent to the hypothesis that all ravens are black. So by the Equivalence Condition, ${\sim}Ba \,\&\, {\sim}Ra$ confirms $(\forall x)(Rx \supset Bx)$ as well. The hypothesis that all ravens are black is confirmed by the observation of a red herring, or a white shoe. This result seems counterintuitive, to say the least.

Nevertheless, Hempel writes that "the impression of a paradoxical situation... is a psychological illusion" (1945a, p. 18); on his view, we reject the confirmational result because we misunderstand what it says. Hempel highlights the fact that in everyday life people make confirmation judgments relative to an extensive corpus of background knowledge. For example, a candidate's performance in an interview may confirm that she'd be good for the job, but only relative to a great deal of background about how the

questions asked relate to the job requirements, how interviewing reveals qualities of character, etc. In assessing confirmation, then, we should always be explicit about the background we're assuming. This is especially important because background knowledge can dramatically alter confirmation relations. For example, in Section 4.3 we discussed a poker game in which you receive the cards that will make up your hand one at a time. At the beginning of the game, your background knowledge contains facts about how a deck is constructed and about which poker hands are winners. At that point the proposition that your last card will be a three of hearts does not confirm the proposition that you will win the hand. But as the game goes along and you're dealt some other threes, your total background knowledge changes such that the proposition that you'll receive the three of hearts now strongly confirms that you'll win.

While Nicod's Criterion states a truth about confirmation for some combinations of evidence, hypothesis, and background corpus, there are other corpora against which applying the Criterion is a bad idea. For instance, suppose I know I'm in the Hall of Atypically-Colored Birds. A bird is placed in the Hall only if the majority of his species-mates are one color but he happens to be another color. Against the background that I'm in the Hall of Atypically-Colored Birds, observing a black raven *disconfirms* the hypothesis that all ravens are black.[2] Hempel thinks the only background against which the Nicod Criterion states a general confirmational truth about all hypotheses and bodies of evidence is the **tautological background**. The tautological background corpus contains no contingent propositions; it is logically equivalent to a tautology T.

When we intuitively reject the Nicod Criterion's consequence that a red herring confirms the ravens hypothesis, we are sneaking non-tautological information into the background. Hempel thinks we're imgaining a situation in which we already know in advance (as part of the background) that we will be observing a herring and checking its color. Relative to that background—which includes the information $\sim Ra$—we know that whatever we're about to observe will have no evidential import for the hypothesis that ravens are black. So when we then get the evidence that $\sim Ba$, that evidence is confirmationally inert with respect to the hypothesis $(\forall x)(Rx \supset Bx)$.

But the original question was whether $\sim Ba \,\&\, \sim Ra$ (taken all together, at once) confirmed $(\forall x)(Rx \supset Bx)$. On Hempel's view, this is a fair test of the Nicod Criterion only against an empty background corpus (since that's the background against which he thinks the Criterion applies). And against that corpus, Hempel thinks the confirmational result is correct. Here's a way of understanding why: Imagine you've decided to test the hypothesis

that all ravens are black. You will do this by selecting objects from the universe one at a time and checking them for ravenhood and blackness. It's the beginning of the experiment, you haven't checked any objects yet, and you have no background information about the tendency of objects to be ravens and/or black. Moreover, you've found a way to select objects from the universe at random, so you have no background information about what kind of object you'll be getting. Nevertheless, you start thinking about what sorts of objects might be selected, and whether they would be good or bad news for the hypothesis. Particularly important would be any ravens that weren't black, since they would immediately refute the hypothesis. (Here it helps to realize that the ravens hypothesis is logically equivalent to $\sim(\exists x)(Rx \,\&\, \sim Bx)$.) So when the first object arrives and you see it's a red herring—$\sim Ba \,\&\, \sim Ra$—this is good news for the hypothesis (at least, moderately good news). After all, the first object could've been a non-black raven, in which case the hypothesis would've been sunk.

This kind of reasoning defuses the seeming paradoxicality of a red herring's confirming that all ravens are black, and the objection to the Nicod Criterion that results. As long as we're careful not to smuggle in illicit background information, a red herring confirms the ravens hypothesis to at least a small degree. Nevertheless, I.J. Good worries about the Nicod Criterion, even against a tautological background:

> [T]he closest I can get to giving [confirmation relative to a tautological background] a practical significance is to imagine an infinitely intelligent newborn baby having built-in neural circuits enabling him to deal with formal logic, English syntax, and subjective probability. He might now argue, after defining a crow in detail, that it is initially extremely unlikely that there are any crows, and therefore that it is extremely likely that all crows are black. "On the other hand," he goes on to argue, "if there are crows, then there is a reasonable chance that they are of a variety of colors. Therefore, if I were to discover that even a black crow exists I would consider [the hypothesis that all crows are black] to be less probable than it was initially."
>
> I conclude from this that the herring is a fairly deep shade of pink. (Good 1968, p. 157)[3]

Here Good takes advantage of the fact that $(\forall x)(Rx \supset Bx)$ is true if there are no ravens (or crows, in his example).[4] Before taking any samples from the universe, the intelligent newborn might consider four possibilities: there

are no ravens; there are ravens but they come in many colors; there are ravens and they're all black; there are ravens and they all share some other color. The first and third of these possibilities would make $(\forall x)(Rx \supset Bx)$ true. When the baby sees a black raven, the first possibility is eliminated; this might be such a serious blow to the ravens hypothesis that the simultaneous elimination of the fourth possibility would not be able to compensate.

6.1.2 Further adequacy conditions

We have already seen two general conditions (Nicod's Criterion and the Equivalence Condition) that one might take the confirmation relation to satisfy. We will now consider a number of other such conditions, most of them discussed (and given the names we will use) by Hempel. Sorting out which of these are genuine properties of confirmation has a number of purposes. First, Hempel thought the correct list provided a set of adequacy conditions for any positive theory of confirmation. Second, sorting through these conditions will help us understand the abstract features of evidential support. These are features about which epistemologists, philosophers of science, and others (including working scientists and ordinary folk!) often make strong assumptions—many of them wrong. Finally, we are going to use the word "confirmation" in subsequent sections as a somewhat technical term, distinct from some of the ways "confirm" is used in everyday speech. Working through the properties of the confirmation relation will help illustrate exactly how we're using the term.

The controversy between Hempel and Good leaves it unclear whether the Nicod Criterion should be endorsed as a constraint on confirmation, even when it's restricted to tautological background. On the other hand, the Equivalence Condition can be embraced in a fairly strong form:

Equivalence Condition (full version): Suppose $H \dashv\models H'$, $E \dashv\models E'$, and $K \dashv\models K'$ in $\mathcal{L}$. Then E confirms (/disconfirms) H against background K just in case E' confirms (/disconfirms) H' against background K'.

Here we can think of K as a conjunction of all the propositions in an agent's background corpus, just as E is often a conjunction of multiple pieces of evidence.

Our next candidate constraint is the

Entailment Condition: For any E, H, and K in $\mathcal{L}$, if $E \,\&\, K \models H$ but $K \not\models H$, then E confirms H relative to K.

This condition enshrines the idea that entailing a hypothesis is one way to support, or provide evidence for, that hypothesis. If E entails H in light of background corpus K (in other words, if E and K together entail H), then E confirms H relative to K. The only exception to this rule is when K already entails H, in which case the fact that E and K together entail H does not indicate any particular relation between E and H.[5] Notice that a tautological H will be entailed by every K, so the restriction on the Entailment Condition keeps it from saying anything about the confirmation of tautologies. Hempel thinks of his adequacy conditions as applying only to empirical hypotheses and bodies of evidence, so he generally restricts them to logically contingent Es and Hs.

Hempel considers a number of adequacy conditions motivated by the following intuition:

Confirmation Transitivity: For any A, B, C, and K in $\mathcal{L}$, if A confirms B and B confirms C relative to K, then A confirms C relative to K.

It's tempting to believe confirmation is transitive, as well as other nearby notions such as justification or evidential support. This temptation is buttressed by the fact that logical entailment is transitive. Confirmation, however, is not in general transitive. Here's an example of Confirmation Transitivity failure. Suppose our background is the fact that a card has just been selected at random from a standard 52-card deck. Consider these three propositions:

A: The card is a spade.

B: The card is the Jack of spades.

C: The card is a Jack.

Relative to our background, A would confirm B, at least to some extent. And relative to our background, B clearly would confirm C. But relative to the background that a card was picked from a fair deck, A does nothing to support the conclusion that C.

The failure of Confirmation Transitivity has a number of important consequences. First, it explains why in the study of confirmation we take evidence to be propositional rather than objectual. In everyday language we often use "evidence" to refer to objects rather than propositions; police don't store propositions in their Evidence Room. But as possible entrants into confirmation relations, objects have an ambiguity akin to the Reference

Class Problem (Section 5.1.1). Should I consider this bird evidence that all ravens are black? If we describe the bird as a black raven, the answers might be yes. But if we describe it as a black raven found in the Hall of Atypically-Colored Birds, the answer seems to be no. Yet a black raven in the Hall of Atypically-Colored birds is still a black raven. If confirmation were transitive, knowing that a particular description of an object confirmed a hypothesis would guarantee that more precise descriptions confirmed the hypothesis as well. Logically stronger descriptions (it's a black raven in the Hall of Atypically-Colored Birds) entail logically weaker descriptions (it's a black raven) of the same object; by the Entailment Condition, the logically stronger description confirms the logically weaker; so if confirmation were transitive anything confirmed by the weaker description would be confirmed by the stronger as well.

But confirmation isn't transitive, so adding more or less information to our description *of the very same object* can alter what's confirmed. (Black raven? Possibly confirms ravens hypothesis. Black raven in Hall of Atypically-Colored Birds? Disconfirms. Black raven mistakenly placed in the Hall of Atypically-Colored Birds when it shouldn't have been? Perhaps confirms again.) We solve this problem by letting propositions rather than objects enter into the confirmation relation. If we state our evidence as a proposition—such as the proposition *that I observed a black raven in the Hall of Atypically-Colored Birds*—there's no question at what level the objects involved are being described.

Confirmation's intransitivity also impacts epistemology more broadly. For instance, it may cause trouble for the principle that "evidence of evidence is evidence". (Feldman 2007) Suppose I read in a magazine that anthropologists have reported evidence that Neanderthals cohabitated with *homo sapiens*. I don't actually have the anthropologists' evidence for that hypothesis—the body of information that they think supports it. But the magazine article constitutes evidence that they have such evidence; one might think that the magazine article therefore also constitutes evidence that Neanderthals and *homo sapiens* cohabitated. (After all, reading the article seems to provide me with some justification for that hypothesis.) Yet we cannot adopt this "evidence of evidence is evidence" principle with full generality. Suppose I've randomly picked a card from a standard deck and examined it carefully. If I tell you my card is a spade, you have evidence that I know my card is the Jack of spades. If I know my card is the Jack of spades, I have (very strong) evidence that my card is a Jack. Yet your evidence that my card is a spade is not evidence that my card is a Jack.

Finally, the failure of Confirmation Transitivity shows what's wrong with

two confirmation constraints Hempel embraces:

Consequence Condition: If an E in $\mathcal{L}$ confirms every member of a set of propositions relative to K and that set jointly entails H' relative to K, then E confirms H' relative to K.

Special Consequence Condition: For any E, H, H', and K in $\mathcal{L}$, if E confirms H relative to K and $H \,\&\, K \models H'$, then E confirms H' relative to K.

The Special Consequence Condition is so named because it's entailed by the Consequence Condition. Yet each of these is a bad idea, as can be demonstrated by our earlier Jack of spades example. (In that example the proposition B that the card is the Jack of spades entails the proposition C that the card is a spade.) In fact, we can even create examples in which H entails H' relative to K, but evidence E which confirms H *dis*confirms H' relative to K. Bradley (ms, §1.3) offers an example relative to the background corpus most of us have concerning the kinds of animals people keep as pets. Relative to that K, the evidence E that Bob's pet is hairless confirms (at least slightly) the hypothesis H that Bob's pet is a Peruvian Hairless Dog. Yet relative to K that same evidence E disconfirms the hypothesis H' that Bob's pet is a dog.[6]

Why might the Special Consequence Condition seem plausible? It certainly looks tempting if one reads "confirmation" in a particular way. In everyday language it's a fairly strong claim that a hypothesis has been "confirmed"; this suggests our evidence is sufficient for us to accept the hypothesis. (Consider the sentences "That confirmed my suspicion" and "Your reservation has been confirmed.") We may then follow Glymour's view that "when we accept a hypothesis we commit ourselves to accepting all of its logical consequences." (1980, p. 31) This would tell us that evidence confirming a hypothesis also confirms its logical consequences, as the Special Consequence Condition requires. But hopefully the discussion to this point has indicated that we are not using "confirms" in this fashion. On our use, evidence confirms a hypothesis if it provides *any* amount of support for that hypothesis; the support need not be decisive. We will often possess evidence that confirms a hypothesis without requiring or even permitting us to accept it—if your only evidence about a card is that it's a spade, it's a bad idea to accept that the card is the Jack of spades.

Another motivation for the Special Consequence Condition—perhaps this was Hempel's motivation—comes from the way we often treat hypotheses in science. Suppose we make a set of atmospheric observations confirming

a particular global warming hypothesis. Suppose further that in combination with our background knowledge, the hypothesis entails that average global temperatures will increase by five degrees in the next fifty years. It's very tempting then to say that the atmospheric observations support the conclusion that temperatures will rise five degrees in fifty years. Yet that's to unthinkingly apply the Special Consequence Condition.

I hope you're getting the impression that denying the Confirmation Transitivity can have serious consequences for the way we think about everyday and scientific reasoning. Yet it's important to realize here that denying the Special Consequence Condition as a *general* principle does not mean these relationships *never* hold. It just means that we need to be careful about assuming they will, and perhaps also that we need a precise, positive theory of confirmation to help us understand when they do and when they don't.

Denying the Special Consequence Condition does open up some intriguing possibilities in epistemology. Consider these three propositions:

E: I am having a perceptual experience as of a hand before me.

H: I have a hand.

H': There is a material world.

This kind of evidence figures prominently in Moore's proof of the existence of an external world. (Moore 1939) Yet for some time it was argued that E could not possibly be evidence for H. The reasoning was, first, that E could not discriminate between H' and various skeptical hypotheses (such as Descartes' evil demon), and therefore could not provide evidence for H'. Next, H entails H', so if E were evidence for H it would be evidence for H' as well. But this step assumes the Special Consequence Condition. Recent positions in epistemology allowing E to support H without supporting H' deny Special Consequence.[7]

Hempel's unfortunate endorsement of the Consequence Condition also pushes him towards the:

Consistency Condition: For any E and K in $\mathcal{L}$, the set of all hypotheses confirmed by E relative to K is logically consistent with $E \,\&\, K$.

In order for the set of all hypotheses confirmed by E to be consistent with $E\&$ K, it first has to be a logically consistent set in its own right. It seems easy to generate confirmational examples that violate this stricture: evidence that a randomly drawn card is red confirms both the hypothesis that it's a heart and the hypothesis that it's a diamond, but these two confirmed hypotheses

are inconsistent with each other. Hempel also notes that in scientific contexts we often entertain a variety of mutually exclusive hypotheses; a particular experiment may eliminate some from contention while confirming the ones that remain. Yet Hempel is trapped into the Consistency Condition by his allegiance to the Consequence Condition. Taken together, the propositions in an inconsistent set entail a contradiction; so if evidence could confirm all the members of an inconsistent set it would (by the Consequence Condition) also confirm a contradiction. Hempel refuses to grant that anything could confirm a contradiction! So he tries to make the Consistency Condition work.[8]

Hempel rightly rejects the

Converse Consequence Condition: For any E, H, H', and K in $\mathcal{L}$, if E confirms H relative to K and $H' \,\&\, K \models H$, then E confirms H' relative to K.

Here's a counterexample. Suppose our background knowledge is that a fair six-sided die has been rolled, and our propositions are:

E: The roll outcome is prime.

H: The roll outcome is odd.

H': The roll outcome is one.

In this case E confirms H relative to our background, H' entails H, yet E refutes H'. (Recall that 1 is not a prime number!)

Still, there's a good idea in the vicinity of Converse Consequence. Suppose our background consists of the fact that we are going to run a certain experiment. A particular scientific theory, in combination with that background, entails that the experiment will produce a particular result. If this result does in fact occur when the experiment is run, we take that to support the theory. This is an example of the

Converse Entailment Condition: For any E, H, and K in $\mathcal{L}$, if $H \,\&\, K \models E$ but $K \not\models E$, then E confirms H relative to K.

Again, this condition rules out cases in which the background K entails the experimental result E all on its own, because such cases need not reveal any connection between H and E.

Converse Entailment doesn't give rise to examples like the die roll case above (because in that case E is not *entailed* by either H or H' in combination with K). But because deductive entailment is transitive, Converse

Entailment does generate the **problem of irrelevant conjunction**. Consider the following propositions:

E: My pet is a flightless bird.

H: My pet is an ostrich.

H': My pet is an ostrich and beryllium is a good conductor.

Here H entails E, so by the Converse Entailment Condition E confirms H, which seems reasonable.[9] Yet despite the fact that H' also entails E (because H' entails H), it seems worrisome that E would confirm H'. What does my choice in pets indicate about the conductivity of beryllium?

Nothing—and that's completely consistent with the Converse Entailment Condition. Just because E confirms a conjunction one of whose conjuncts concerns beryllium doesn't mean E confirms that beryllium-conjunct all on its own. To assume that it does would be to assume the Special Consequence Condition, which we've rejected. So facts about my pet don't confirm any conclusions that are about beryllium but not about birds. On the other hand, it's reasonable that E would confirm H' at least to some extent, by virtue of eliminating such rival hypotheses as "beryllium is a good conductor and my pet is an iguana."

Rejecting the Special Consequence Condition therefore allows us to accept Converse Entailment. But again, all this should make us very careful about how we reason in our everyday lives. A scientific theory, for instance, will often have wide-ranging consequences, and might be thought of as a massive conjunction. When the theory entails a prediction and that prediction occurs, this confirms the theory. But it doesn't necessarily confirm each of the conjuncts within the theory, taken in isolation.[10]

Finally, we should say something about disconfirmation. Hempel takes the following position:

Disconfirmation Duality: For any E, H, and K in $\mathcal{L}$, E confirms H relative to K just in case E disconfirms $\sim H$ relative to K.

Disconfirmation Duality allows us to immediately convert many of our constraints on confirmation into constraints on disconfirmation. For example, the Entailment Condition now tells us that if $E \,\&\, K$ deductively refutes H (yet K doesn't refute H all by itself), then E disconfirms H relative to K. (See Exercise 6.2.) We should be careful, though, not to think confirmation and disconfirmation are exhaustive categories: for many propositions E, H, and K, E will neither confirm nor disconfirm H relative to K.

Figure 6.1: Accepted and rejected conditions on confirmation

Name	Brief, Somewhat Imprecise Description	Verdict
Equivalence Condition	equivalent hypotheses, evidence, backgrounds behave same confirmationally	accepted
Entailment Condition	evidence confirms what it entails	accepted
Converse Entailment Condition	a hypothesis is confirmed by what it entails	accepted
Disconfirmation Duality	a hypothesis is confirmed just when its negation is disconfirmed	accepted
Confirmation Transitivity	anything confirmed by a confirmed hypothesis is also confirmed	rejected
Consequence Condition	anything entailed by a set of confirmed hypotheses is also confirmed	rejected
Special Consequence Condition	anything entailed by a confirmed hypothesis is also confirmed	rejected
Consistency Condition	all confirmed hypotheses are consistent	rejected
Converse Consequence Condition	anything that entails a confirmed hypothesis is also confirmed	rejected
Nicod's Criterion	Fa & Ga confirms $(\forall x)(Fx \supset Gx)$	???

Figure 6.1 summarizes the formal conditions on confirmation we have accepted and rejected. The task now is to find a positive theory of which evidence confirms which hypotheses relative to which backgrounds that satisfies the right conditions and avoids the wrong ones.

6.2 Carnap's Theory of Confirmation

6.2.1 Confirmation as relevance

Carnap saw that we could get a confirmation theory with exactly the right properties by basing it on probability. Begin by taking *any* probabilistic distribution Pr over $\mathcal{L}$. (I've named it "Pr" because we aren't committed at this stage to its being any *kind* of probability in particular—much less a credence distribution. All we know is that it's a distribution satisfying the Kolmogorov axioms.) Define Pr's background corpus K as the conjunction of all propositions X in $\mathcal{L}$ such that $\Pr(X) = 1$.[11] Given an E and H in $\mathcal{L}$, we apply the Ratio Formula to calculate $\Pr(H \mid E)$. Two distinct theories of confirmation now suggest themselves: (1) E confirms H relative to K just in case $\Pr(H \mid E)$ is high; (2) E confirms H relative to K just

in case $\Pr(H \mid E) > \Pr(H)$. In the preface to the second edition of his *Logical Foundations of Probability*, Carnap calls the first of these options a "firmness" concept of confirmation and the second an "increase in firmness" concept.[12] (1962, p. xvff.)

The firmness concept of confirmation has a number of problems. First, there are questions about where exactly the threshold for a "high" value of $\Pr(H \mid E)$ falls, what determines that threshold, how we discover it, etc. Second, there will be cases in which E is irrelevant to H, yet $\Pr(H \mid E)$ is high because $\Pr(H)$ is already high. For example, take the background K that a fair lottery with a million tickets has been held, the hypothesis H that ticket 942 did not win, and the evidence E that elephants have trunks. In this example $\Pr(H \mid E)$ may very well be high, but that need not be due to any confirmation of lottery results by the endowments of elephants. Finally, the firmness concept doesn't get the confirmation conditions we identified in the previous section right. Wherever the threshold for "high" is set, whenever E confirms H relative to K it will also confirm any H' entailed by H. As a probability distribution, Pr must satisfy the Entailment rule and its extension to conditional probabilities (see Section 3.1.2), so if $H \models H'$ then $\Pr(H' \mid E) \geqslant \Pr(H \mid E)$. If $\Pr(H \mid E)$ surpasses the threshold, $\Pr(H' \mid E)$ will as well. But that means the firmness concept of confirmation satisfies the Special Consequence Condition, to which we've already seen counterexamples.

Warning: Conflating firmness and increase in firmness, or just blithely assuming the firmness concept is correct, is one of the most frequent mistakes made in the confirmation literature and more generally in discussions of evidential support.[13] For example, it is often claimed that an agent's evidence supports or justifies a conclusion just in case the conclusion is probable on that evidence. But for conclusions with a high prior, the conclusion may be probable on the evidence not because of anything the evidence is doing, but instead because the conclusion was probable all along. Then it's not *the evidence* that's justifying anything!

Increase in firmness has none of these disadvantages; it is the concept of confirmation we'll work with going forward. Given a probability distribution Pr with background K (as defined above), E confirms H relative to K just in case $\Pr(H \mid E) > \Pr(H)$. In other words, given Pr evidence E confirms

H relative to K just in case E is *positively relevant* to H. We identify disconfirmation with *negative relevance*: Given Pr, E disconfirms H relative to K just in case $\Pr(H \mid E) < \Pr(H)$. If $\Pr(H \mid E) = \Pr(H)$, then E is irrelevant to H and neither confirms nor disconfirms it relative to K.

This account of confirmation meets exactly those conditions we endorsed in the previous section: Disconfirmation Duality and the Equivalence, Entailment, and Converse Entailment Conditions. Disconfirmation Duality follows immediately from our definitions of positive and negative relevance. The Equivalence Condition follows from the Equivalence rule for probability distributions; logically equivalent propositions will always receive identical Pr-values. We get the Entailment Condition because if $E \,\&\, K \models H$ but $K \not\models H$, then $\Pr(H \mid E) = 1$ while $\Pr(H) < 1$. (If $\Pr(H)$ were 1, then H would be a conjunct of K, which would contradict $K \not\models H$.) The key result for Converse Entailment was established in Exercise 4.4. Identifying confirmation with positive relevance yields an account of confirmation with exactly the general contours we want, without our having to commit on the specific numerical values of Pr.

6.2.2 Finding the right function

Yet Carnap wants more than the general contours of confirmation—he wants a substantive theory that says which bodies of evidence support which hypotheses relative to which backgrounds. A theory like that seems obtainable to Carnap because he sees confirmation as a *logical* relation. As with other logical relations, whether E confirms H relative to K is independent of the truth-values of those propositions and of any particular attitudes individuals adopt toward them. Like Hempel, Carnap thinks confirmation relations emerge from the logical form of propositions, and therefore can be captured by a syntactical theory working with strings of symbols representing those forms. (Nicod's Criterion is a good example of a confirmation principle that works with logical form.) Enormous progress in formal deductive logic in the decades just before *Logical Foundations* makes Carnap confident that a formalism for inductive logic is within reach.

To construct the formalism Carnap wants, we begin with a formal language $\mathcal{L}$.[14] We then take each consistent corpus K and associate it with a particular Pr distribution over $\mathcal{L}$. That done, we can test whether evidence E confirms hypothesis H relative to a particular K by seeing whether E is positively relevant to H on the Pr associated with that K.

The crucial step for Carnap is to associate each K with the unique, correct distribution Pr. Of course Pr will assign an unconditional value of

1 to each conjunct of K, but that leaves a lot of latitude with respect to the members of $\mathcal{L}$ that aren't conjuncts of K. Yet a full Pr distribution must be specified for each K so that for any E, H, and K we might select in $\mathcal{L}$, there will be a definite answer to the question of whether E confirms, disconfirms, or is irrelevant to H on K. (Just as there's always a definite answer as to whether a given P deductively entails a given Q, refutes it, or neither.) And it's important to get the *right* Pr for each K; the wrong Pr distribution could make evidential support counterinductive, or could have everyday evidence confirming skeptical hypotheses.

Since there are infinitely many possible consistent background corpora K, specifying a Pr-distribution for each one could be a great deal of trouble. Carnap simplifies the process by constructing every Pr from a single, regular probability distribution he calls $\mathfrak{m}$. As a regular probability distribution, $\mathfrak{m}$ contains no contingent evidence. ($\mathfrak{m}$ has a tautological background corpus.) The $\Pr(\cdot)$ distribution relative to any consistent, non-tautological K is then specified as $\mathfrak{m}(\cdot \mid K)$. (This guarantees that $\Pr(K) = 1$.) Evidence E confirms hypothesis H relative to K just in case $\Pr(H \mid E) > \Pr(H)$, which is equivalent to $\mathfrak{m}(H \mid E \,\&\, K) > \mathfrak{m}(H \mid K)$. So instead of working with particular Pr-distributions we can now focus our attention on $\mathfrak{m}$.[15]

$\mathfrak{m}$ also fulfills a number of other roles for Carnap. Carnap thinks of an agent's background corpus at a given time as her total evidence at that time. If an agent's total evidence is E, Carnap thinks $\mathfrak{m}(H \mid E)$ provides the logical probability of H on her total evidence. Moreover, if the agent is rational she will assign credence $\mathrm{cr}(H) = \mathfrak{m}(H \mid E)$ for any H in $\mathcal{L}$. Since $\mathfrak{m}$ is the unique logical probability function, this means there is a unique credence any agent is required to assign a particular proposition H given body of total evidence E. So Carnap endorses the Uniqueness Thesis (Section 5.1.2), with $\mathfrak{m}$ playing the role of the uniquely rational hypothetical prior function. On Carnap's view, logic provides the correct evidential standards all rational agents should apply, represented numerically by the function $\mathfrak{m}$. Carnap is thus an Objective Bayesian in both senses of the term: in the normative sense, because he thinks there's a unique rational hypothetical prior; and in the semantic sense, because he defines "probability" as an objective concept independent of agents' particular attitudes.[16]

Carnap doesn't just talk about this hypothetical distribution $\mathfrak{m}$; he provides a recipe for calculating its numerical values. To see how, let's begin with a very simple language, containing only one predicate F and two constants a and b. This language has only two atomic propositions (Fa and Fb), so we can specify distribution $\mathfrak{m}$ over the language using a stochastic truth-table with four rows. Carnap runs through a few candidates for

distribution $\mathfrak{m}$; he calls the first one $\mathfrak{m}^\dagger$:

Fa	Fb	$\mathfrak{m}^\dagger$
T	T	1/4
T	F	1/4
F	T	1/4
F	F	1/4

$\mathfrak{m}^\dagger$ captures the natural thought that a tautological background should treat each of the available possibilities symmetrically. So $\mathfrak{m}^\dagger$ applies a principle of indifference and assigns each state-description the same value.[17]

Yet $\mathfrak{m}^\dagger$ has a serious drawback:

$$\mathfrak{m}^\dagger(Fb \,|\, Fa) = \mathfrak{m}^\dagger(Fb) = 1/2 \tag{6.1}$$

On $\mathfrak{m}^\dagger$, Fa is irrelevant to Fb; so according to $\mathfrak{m}^\dagger$, Fa does not confirm Fb relative to the empty background. Carnap thinks the fact that one object is an F should confirm that the next object will be an F, yet $\mathfrak{m}^\dagger$ does not yield that result. Even worse, this failure is carried over as $\mathfrak{m}^\dagger$ is extended to larger languages. $\mathfrak{m}^\dagger$ makes each proposition Fa, Fb, Fc, etc. independent not only of each of the others but also of logical combinations of the others; even the observation that 99 objects all have property F will not confirm that the 100th object is an F. (See Exercise 6.3.) This is an especially bad result because $\mathfrak{m}^\dagger$ is proposed as the unique hypothetical prior for rational agents. If $\mathfrak{m}^\dagger$ were correct, then an agent whose total evidence consisted of the fact that 99 objects all had property F would nevertheless be 50-50 on whether the next object would have F. $\mathfrak{m}^\dagger$ does not allow "learning from experience"; as Carnap puts it,

> The choice of [$\mathfrak{m}^\dagger$] as the degree of confirmation would be tantamount to the principle never to let our past experiences influence our expectations for the future. This would obviously be in striking contradiction to the basic principle of all inductive reasoning.(1950, p. 565)

Carnap wants a theory of confirmation that squares with commonsense notions of rational inductive reasoning; $\mathfrak{m}^\dagger$ is clearly failing in that role.

To address this problem, Carnap proposes distribution $\mathfrak{m}^*$. According to $\mathfrak{m}^*$, logical probability is indifferent not among the state-descriptions in a language but instead among its **structure-descriptions**. To understand structure-descriptions, start by thinking about property profiles. A property profile specifies exactly which of the language's predicates an object does

or does not satisfy. In a language with the single predicate F, the two available property profiles would be "this object has property F" and "this object lacks property F"; in a language with two predicates the property profiles would include "this object lacks property F but has property G"; etc. Given language $\mathcal{L}$, a structure-description describes how *many* objects in the universe of discourse possess each of the available property profiles, but doesn't say which *particular* objects possess which profiles. For example, the language containing one property F and two constants a and b has the two property profiles just mentioned. Since there are two objects, this language allows three structure-descriptions: "both objects have F", "one object has F and one object lacks F", and "both objects lack F". Written in disjunctive normal form, the three structure-descriptions are:

$$\begin{gathered} Fa \,\&\, Fb \\ (Fa \,\&\, {\sim}Fb) \vee ({\sim}Fa \,\&\, Fb) \\ {\sim}Fa \,\&\, {\sim}Fb \end{gathered} \tag{6.2}$$

Note that one of these structure-descriptions is a disjunction of multiple state-descriptions.[18] $\mathfrak{m}^*$ works by assigning equal value to each structure-description in a language. If a structure-description contains multiple state-description disjuncts, $\mathfrak{m}^*$ then divides the value of that structure-description equally among its state-descriptions. For our simple language, the result is:

Fa	Fb	$\mathfrak{m}^*$
T	T	1/3
T	F	1/6
F	T	1/6
F	F	1/3

Each structure-description receives $\mathfrak{m}^*$-value 1/3; the structure-description containing the middle two lines of the table divides its $\mathfrak{m}^*$-value between them.

$\mathfrak{m}^*$ allows learning from experience. From the table above, we can calculate

$$\mathfrak{m}^*(Fb \,|\, Fa) = 2/3 > 1/2 = \mathfrak{m}^*(Fb) \tag{6.3}$$

On $\mathfrak{m}^*$, the fact that a possesses property F confirms that b will have F relative to the tautological background.

Nevertheless, $\mathfrak{m}^*$ falls short in a different way. Suppose our language contains two predicates F and G and two constants a and b. Carnap thinks that on the correct, logical $\mathfrak{m}$ distribution we should have

$$\mathfrak{m}(Fb \,|\, Fa\&Ga\&Gb) > \mathfrak{m}(Fb \,|\, Fa) > \mathfrak{m}(Fb \,|\, Fa\&Ga\&{\sim}Gb) > \mathfrak{m}(Fb) \tag{6.4}$$

While evidence that a has F should increase a rational agent's confidence that b has F, that rational confidence should increase even higher if we throw in the evidence that a and b share property G. If a and b both have G, in some sense they're the same kind of object, so one should expect them to be alike with respect to F as well. On the other hand, information that a and b are unalike with respect to G should make the fact that a has F less influential on one's confidence that b has F than if one knew nothing about how things stood with G.

To see if Equation (6.4) holds for $\mathfrak{m}^*$, one would need to identify the structure-descriptions in this language. The available property profiles are: object has both F and G, object has F but not G, object has G but not F, object has neither. Some examples of structure-descriptions are: both objects have F and G, one object has both F and G while the other has neither, one object has F but not G while the other object has G but not F, etc. I'll leave the details to the reader (see Exercise 6.4), but suffice it to say that Mary Hesse demonstrated to Carnap that $\mathfrak{m}^*$ is unable to capture **analogical effects** such as Equation (6.4).

Carnap eventually responded to this problem (Carnap 1952) by introducing a continuum of $\mathfrak{m}$-distributions with properties set by two adjustable parameters. The parameter λ was an "index of caution", controlling how reluctant $\mathfrak{m}$ made an agent to learn from experience. $\mathfrak{m}^\dagger$ was the $\mathfrak{m}$-distribution with λ-value ∞ (because it made the agent infinitely cautious and forbade learning from experience), while $\mathfrak{m}^*$ had λ-value 2. Adjusting the other parameter, γ, made analogical effects possible. Carnap suggested the values of these parameters be set by "pragmatic" considerations, yet this threatened the Objective Bayesian aspects of his project. At the same time, Carnap found other, more subtle learning effects that even his parameterized $\mathfrak{m}$-distributions were unable to represent.

6.3 Grue

Nelson Goodman (1946, 1979) offered another kind of challenge to Hempel and Carnap's theories of confirmation. Here is the famous passage:

> Suppose that all emeralds examined before a certain time t are green. At time t, then, our observations support the hypothesis that all emeralds are green; and this is in accord with our definition of confirmation. Our evidence statements assert that emerald a is green, that emerald b is green, and so on; and each

> confirms the general hypothesis that all emeralds are green. So far, so good.
>
> Now let me introduce another predicate less familiar than "green". It is the predicate "grue" and it applies to all things examined before t just in case they are green but to other things just in case they are blue. Then at time t we have, for each evidence statement asserting that a given emerald is green, a parallel evidence statement asserting that that emerald is grue. And the statements that emerald a is grue, that emerald b is grue, and so on, will each confirm the general hypothesis that all emeralds are grue. Thus according to our definition, the prediction that all emeralds subsequently examined will be green and the prediction that all will be grue are alike confirmed by evidence statements describing the same observations. But if an emerald subsequently examined is grue, it is blue and hence not green. Thus although we are well aware which of the two incompatible predictions is genuinely confirmed, they are equally well confirmed according to our definition. Moreover, it is clear that if we simply choose an appropriate predicate, then on the basis of these same observations we shall have equal confirmation, by our definition, for any prediction whatever about other emeralds. (1979, pp. 73–4)

The target here is any theory of confirmation on which the observation that multiple objects all have property F confirms that the next object will have F as well. As we saw, Carnap built this "learning from experience" feature into his theory of confirmation. It was also a feature of Hempel's positive theory of confirmation, so Goodman is objecting to both Carnap's and Hempel's theories. We will focus on the consequences for Carnap, since I did not present the details of Hempel's approach.

Goodman's concern is as follows: Suppose we have observed 99 emeralds before time t, and they have all been green. On Carnap's theory, this evidence confirms the hypothesis that the next emerald observed will be green. So far, so good. But Goodman says this evidence can be re-expressed as the proposition that the first 99 emeralds are grue. On Carnap's theory, this evidence confirms the hypothesis that the next emerald observed will be grue. But for the next emerald to be grue it must be blue. Thus it seems that on Carnap's theory our evidence confirms both the prediction that the next emerald will be green and the prediction that the next emerald will be blue. Goodman thinks it's intuitively obvious that the former prediction

is confirmed by our evidence while the latter is not, so Carnap's theory is getting things wrong.

Let's look more carefully at the details. Begin with a language $\mathcal{L}$ containing constants a_1 through a_{100} representing objects, and predicates G and O representing the following properties:

Gx: x is green
Ox: x is observed by time t

To simplify our analysis and equations, I will assume that the following facts are part of our background corpus, then suppress mention of that background in what follows: (1) objects a_1 through a_{100} are all emeralds (so we don't have to bother with an "is an emerald" predicate); (2) each object is observed exactly once (so we can partition the objects into "is observed by t" and "is observed after t"); and (3) each object is either green or blue (so blue can be treated simply as the negation of green).[19] Against this background, we can define "grue" as follows:

$Gx \equiv Ox$: x is grue; it is either green and observed by time t or non-green (blue) and observed after time t

The grue predicate says that the facts about whether an emerald is green match the facts about whether it was observed by t. Goodman claims that according to Carnap's theory, our evidence in the example confirms $(\forall x)Gx$ and Ga_{100} (which is good), but also $(\forall x)(Gx \equiv Ox)$ and $Ga_{100} \equiv Oa_{100}$ (which are supposed to be bad).

But what exactly *is* our evidence in the example? Goodman agrees with Hempel that in assessing confirmation relations we must explicitly and precisely state the contents of both our total evidence and the background corpus. Evidence that the first 99 emeralds are green would be:

E: $Ga_1 \,\&\, Ga_2 \,\&\, \ldots \,\&\, Ga_{99}$

But E neither entails nor is equivalent to the statement that the first 99 emeralds are grue (because it doesn't say anything about whether those emeralds' G-ness matches their O-ness), nor does E confirm $(\forall x)(Gx \equiv Ox)$ on Carnap's theory.

A better statement of the total evidence would be:

E': $(Ga_1 \,\&\, Oa_1) \,\&\, (Ga_2 \,\&\, Oa_2) \,\&\, \ldots \,\&\, (Ga_{99} \,\&\, Oa_{99})$

Here we've added an important fact that was included in the example: that emeralds a_1 through a_{99} were observed by t. This evidence statement entails

both that all those emeralds were green and that they all were grue. A bit of technical work with Carnap's theory[20] will also show that according to that theory, E' confirms $(\forall x)Gx$, Ga_{100}, $(\forall x)(Gx \equiv Ox)$, and $Ga_{100} \equiv Oa_{100}$.

It looks like Carnap is in trouble. As long as his theory is willing to "project" past observations of any property onto future predictions that that property will appear, it will confirm grue predictions alongside green predictions. The theory seems to need a way of preferring greenness over grueness for projection purposes; it seems to need a way to play favorites among properties.

Might this need be met by a technical fix? One obvious difference between green and grue is the more complex logical form of the grue predicate in $\mathcal{L}$. There's also the fact that the definition of "grue" involves a predicate O that makes a reference to times; perhaps for purposes of induction predicates referring to times are suspicious. Yet Goodman points out that we can turn all these comparisons around by re-expressing the problem in an alternate language $\mathcal{L}'$, built on the following two predicates:

GRx: x is grue
Ox: x is observed by time t

We can define the predicate "green" in language $\mathcal{L}'$; it will look like this:

$GRx \equiv Ox$: x is green; it is either grue and observed by time t or non-grue and observed after time t

Expressed in $\mathcal{L}'$, the evidence E' is

$$E'\text{: } (GRa_1 \,\&\, Oa_1) \,\&\, (GRa_2 \,\&\, Oa_2) \,\&\, \ldots \,\&\, (GRa_{99} \,\&\, Oa_{99})$$

This expression of E' in $\mathcal{L}'$ is true in exactly the same possible worlds as the expression of E' we gave in $\mathcal{L}$. And once more, when applied to $\mathcal{L}'$ Carnap's theory has E' confirming both that all emeralds are grue and that they are green, and that a_{100} will be grue and that it will be green.

But in $\mathcal{L}'$ all the features that were supposed to help us discriminate against grue now work against green—it's the definition of greenness that is logically complex and mentions the predicate O referring to time. If you believe that it's logical complexity or reference to times that makes the difference between green and grue, you now need a reason to prefer the expression of the problem in language $\mathcal{L}$ over its expression in $\mathcal{L}'$. This is why Goodman's grue problem is sometimes described as a problem of **language dependence**: We could build a formal confirmation theory that projected logically simple predicates but not logically complex, yet such a

theory would yield different answers when applied to the *very same problem* expressed in different languages (such as $\mathcal{L}$ and $\mathcal{L}'$).

Why is language dependence such a concern? Recall that Hempel endorsed the Equivalence Condition in part because he didn't want confirmation to depend on the particular way hypotheses and evidence were presented. If propositions in alternative languages say the same thing, they should enter into the same confirmation relations. Especially for theorists like Hempel and Carnap who take confirmational relations to be objective, how particular subjects choose linguistically to represent certain propositions should not make a difference.[21] (The language a scientist speaks isn't supposed to be relevant to the conclusions she draws from her data!) Notice, by the way, that satisfying the probability axioms is language-independent: If a distribution over a particular language satisfies the axioms, copying its values from propositions in that original language to the equivalent propositions in a different language will yield a new distribution that satisfies the axioms as well.

Hempel and Carnap sought a successful theory of confirmation that worked exclusively with the syntactical forms of propositions represented in language. Goodman charges that such theories can yield consistent verdicts only if appropriate languages are selected for them to operate within. Since a syntactical theory operates only once a language has been provided, it cannot choose among languages for us. Goodman concludes that "Confirmation of a hypothesis by an instance depends rather heavily upon features of the hypothesis other than its syntactical form." (1979, pp. 72–3)

Warning: It is sometimes suggested that—although this is certainly not a *syntactical* distinction—the grue hypothesis can be dismissed out of hand on the grounds that it is "metaphysically weird". This involves reading "All emeralds are grue" as being true just in case all the emeralds in the universe are green before time t then switch to being blue after t. But that reading is neither required to get the problem going nor demanded by anything in (Goodman 1946) or (Goodman 1979). Suppose, for instance, that each emerald in the universe is either green or blue, and no emerald ever changes color. By an unfortunate accident, it just so happens that the emeralds you observe by t are all and only the green emeralds. In that case it will be true that all emeralds are grue, and no metaphysical sleight-of-hand was required.

As the previous warning suggests, the metaphysical details of Goodman's grue example have sometimes obscured its philosophical point. "Grue" indicates a correlation between two properties: being green and being observed before time t. It happens to be a perfect correlation, expressed by a biconditional. Some such correlations are legitimately projectible in science: If you observe that fish are born with a fin on the left side whenever they are born with a fin on the right, this bilateral symmetry is a useful, projectible biconditional correlation. The trouble is that any sizable body of data will contain many correlations, and we need to figure out which ones to project as regularities that will actually extend into the future. (The women in this meeting hall all have non-red hair, all belong to a particular organization, and all are under 6 feet tall. Which of those properties will also be exhibited by the next woman to enter?) Grue is a particularly odd, particularly stark example of a spurious correlation, but is emblematic of the problem of sorting projectible from unprojectible hypotheses.[22]

Goodman offers his own proposal for detecting projectible hypotheses, and many authors have made further proposals since then. Instead of investigating those, I'd like to examine exactly what the grue problem establishes about Carnap's theory (and others). The first thing to note is that although evidence E' confirms on Carnap's theory that emerald a_{100} is grue, it does *not* confirm that emerald a_{100} is blue. E' confirms Ga_{100}. Carnap's theory interprets confirmation as positive relevance on the probability distribution $\mathfrak{m}^*$, so in Carnap's theory

$$\mathfrak{m}^*(Ga_{100} \mid E') > \mathfrak{m}^*(Ga_{100}) \tag{6.5}$$

But if that's true, E' must be *negatively* relevant to $\sim Ga_{100}$, the proposition that emerald a_{100} is blue. So while E' confirms that a_{100} is green and confirms that a_{100} is grue, it does not confirm that a_{100} is blue.

How is this possible, given that $\sim Oa_{100}$ (i.e. a_{100} is not observed by t)? The key point is that $\sim Oa_{100}$ is not stated in the evidence E' being analyzed. E' says that every emerald a_1 through a_{99} was observed by t and is green. If that's all we put into the evidence, that evidence is going to confirm that a_{100} both is green *and was observed by* t. After all, if every object described in the evidence has the property $Gx \,\&\, Ox$, Carnapian "learning from experience" will confirm that other objects have that property as well. Once we understand that Carnap's theory is predicting from E' that a_{100} bears both Ox and Gx, the prediction that a_{100} will have $Gx \equiv Ox$ is no longer so startling.

In fact, the assessment of E' one gets from Carnap's theory is intuitively plausible. If *all you knew* about the world was that there existed 99 objects

and all of them were green and observed before t, you would expect that if there were a 100th object it would be green and observed before t as well. In other words, you'd expect the 100th object to be grue—by virtue of being green (and observed), not blue![23] We read the prediction that the 100th object is grue as a prediction that it's not green because we are smuggling covert background knowledge into the case. (Similar to what happened in Hempel's analysis of the Paradox of the Ravens.) We assume that a_{100} is an *unobserved* emerald; so when E' confirms that a_{100} is grue we take that to be tantamount to confirming that a_{100} is blue. What happens if we explicitly state in the evidence that a_{100} is not observed by t?

$$E'': \quad (Ga_1 \,\&\, Oa_1) \,\&\, (Ga_2 \,\&\, Oa_2) \,\&\, \ldots \,\&\, (Ga_{99} \,\&\, Oa_{99}) \,\&\, \sim Oa_{100}$$

Skipping the calculations (see Exercise 6.5), it turns out that

$$\begin{aligned} \mathfrak{m}^*(Ga_{100} \equiv Oa_{100} \,|\, E'') &= \mathfrak{m}^*(\sim Ga_{100} \,|\, E'') \\ &= \mathfrak{m}^*(Ga_{100} \,|\, E'') = \mathfrak{m}^*(Ga_{100}) \\ &= 1/2 \end{aligned} \tag{6.6}$$

Relative to Carnap's probabilistic distribution $\mathfrak{m}^*$, E'' confirms neither that a_{100} will be grue, nor that a_{100} will be green, nor—for that matter—that all emeralds are grue or that all emeralds are green.

Perhaps the lack of confirmation here for some hypotheses that intuitively should be confirmed is a problem for Carnap's theory. Or perhaps the willingness of $\mathfrak{m}^*$ to have E' confirm that all emeralds are grue—even if that doesn't have the consequence of confirming that the next emerald will be blue—is a problem for Carnap's theory. Suffice it to say that while language-dependence problems can be found for Carnap's theory as well as various other positive theories of confirmation,[24] it's very subtle to determine exactly where these problems lie and what their significance is.

6.4 Subjective Bayesian confirmation

Carnap was an Objective Bayesian, in the normative sense: He believed that given any body of total evidence, there was a unique credence a rational agent would assign any given proposition in light of that evidence. These unique rational credences could be determined from $\mathfrak{m}$, the hypothetical prior distribution representing Carnap's "logical probabilities". $\mathfrak{m}$ would also allow us to determine which bodies of evidence supported which hypotheses; a hypothesis H was supported by body of total evidence E just in case E was positively relevant to H on $\mathfrak{m}$.

Subjective Bayesians do not believe in a distribution $\mathfrak{m}$ that determines the correct attitudes relative to evidence for all rational agents. They are willing to let different rational agents construct their credences using different hypothetical priors, encoding those agents' differing evidential standards. Yet Subjective Bayesians retain Carnap's insight that if we define confirmation as positive probabilistic relevance, the confirmation relation winds up having all the desirable features we identified in Section 6.1.2.

Just as Hempel said confirmation relations are always relative to a background, Subjective Bayesians hold that confirmation relations between evidence and hypotheses are always relative to some probability distribution. If I ask you whether a particular E supports a particular H, you shouldn't answer without first determining which probability distribution the question is relative to. For a Subjective Bayesian, it may be different distributions on different occasions, or for different agents on the same occasion.[25]

For example, in Section 4.3 we discussed a game of five-card stud in which you receive your cards one at a time. At the beginning of the game your credence distribution about the possible outcomes of the hand is dictated by the chances of receiving various cards at random from a standard deck. Relative to this distribution (and the rules of poker), the proposition that the last card dealt to you will be the three of hearts disconfirms the hypothesis that you will win the hand. But as the hand is dealt and you receive all the other threes in the deck, your credence distribution changes. Relative to this updated distribution, the proposition that your last card will be the three of hearts confirms that you'll win. (Four threes is an almost unbeatable hand.) However, you may have a friend observing the game who is very suspicious that the dealer is dealing from the bottom of the deck. A cheating dealer would deal one of the players an almost unbeatable hand, just to make sure that player (namely, you) would bet a great deal of money before inevitably being defeated by the dealer's superior hand. Relative to your friend's credence distribution concerning the game, your receiving that three of hearts strongly *dis*confirms that you'll win the hand.

Often, an agent will make confirmation judgments relative to her own credence distribution. But the central claim of the Subjective Bayesian approach to confirmation is that confirmation is relative to *some* probability distribution; the determining distribution need not always be an agent's credence function. For example, a scientist may assess her experimental data relative to a commonly-accepted probabilistic model of the phenomenon under examination (such as a statistical model of gases), even if that model doesn't match her personal credences about the events in question. Similarly, a group may agree to judge evidence relative to a probability distribu-

tion distinct from the credence functions of each of its members. Whatever probability distribution we consider, the Kolmogorov axioms and Ratio Formula ensure that the confirmation relation relative to that distribution will meet the general conditions we desire.

Warning: Bayesian theories of confirmation make confirmation relative to a probability distribution Pr. $\Pr(H \mid E)$ tells us how probable H is given E (relative to Pr). Also, if Pr represents a particular agent's hypothetical prior, $\Pr(H \mid E)$ tells us the degree of confidence that agent should have in H when her total evidence is E. It is important to distinguish this position from the following claims one sometimes finds in the literature:

- Some authors describe $\Pr(H \mid E)$ as "the degree to which E justifies H" relative to Pr. This is a mistake—it's another example of the firmness/increase in firmness conflation we discussed in Section 6.2.1. The value of $\Pr(H \mid E)$ can be affected just as much by the value of $\Pr(H)$ as by the influence of E, so $\Pr(H \mid E)$ is not solely reflective of the relationship between E and H.
- It's sometimes suggested that $\Pr(H \mid E)$ is "the degree to which an agent with total evidence E would be justified in believing or accepting H". (Notice that this is different from asking how much E *itself* justifies accepting H, again for firmness/increase in firmness reasons.) Now there are various formal theories of how much a body of evidence justifies an agent in accepting a hypothesis; some even attach numbers to how much an agent is justified. Yet this is a very different project from the standard Bayesian analysis of confirmation; for instance, it's unclear why degrees of justification for acceptance should have anything to do the probability axioms. (See (Shogenji 2012).)
- Finally, there is the view that an agent is justified in believing or accepting H just in case $\Pr(H \mid E)$ is high (where E represents her total evidence). Here $\Pr(H \mid E)$ is not supposed to measure how justified such an acceptance would be; it's simply part of a necessary condition for such acceptance to be justified. If $\Pr(H \mid E)$ is the credence an agent with total evidence E is rationally required to assign H, then this proposal depends on one's views about rational relations between credences and

binary acceptances/beliefs.

The most common objection to the Subjective Bayesian view of confirmation is that for confirmation to play the objective role we require in areas like scientific inquiry, it should never be relative to something so subjective as an agent's degrees of belief about the world. We will return to this objection—and some theories of confirmation that try to avoid it—in Chapter XXX. For now I want to consider another objection to the Subjective Bayesian view, namely that it is so empty as to be near-useless. There are so many probability distributions available that for any E and H we will be able to find *some* distribution on which they are positively relevant (except in extreme cases when $E \models \sim H$). It looks, then, like the Subjective Bayesian view tells us almost nothing substantive about which particular hypotheses are confirmed by which bodies of evidence.

While the Subjective Bayesian denies the existence of a unique probability distribution to which all confirmation relations are relative, the view need not be anything-goes.[26] Often we are interested in confirmation relations relative to some rational agent's credences, and Chapter 5 proposed a number of plausible constraints beyond the Kolmogorov axioms and Ratio Formula that such credences will satisfy. These constraints, in turn, impose some substantive shape on any confirmation relation defined relative to a rational agent's credences. For example, David Lewis shows at his (1980, p. 285ff.) that if a credence distribution satisfies the Principal Principle, then relative to that distribution the evidence that a coin has come up heads on x percent of its tosses will confirm that the objective chance of heads on a single toss is close to x.

This result of Lewis's has the form: if your credences have features such-and-such, then confirmation relative to those credences will have features so-and-so. The fact that features such-and-such are required by rationality is neither here nor there. For example, if you assign equal credence to each possible outcome of the roll of a six-sided die, then relative to your credence distribution the evidence that the roll came up odd will confirm that it came up prime. This will be true regardless of whether your total evidence rationally required such equanimity over the outcomes. Subjective Bayesianism can yield interesting, informative results about which bodies of evidence confirm which hypotheses once the details of the relevant probability distribution are specified.[27]

The theory can also work in the opposite direction: it can tell us what features in a probability distribution will generate particular kinds of con-

firmational relations. But before I can outline some of Subjective Bayesianism's more interesting results on that front, I need to explain how Bayesians measure the strength of evidential support.

6.4.1 Confirmation measures

We have been considering a *classificatory* question: Under what conditions does a body of evidence E confirm a hypothesis H? But related to that classificatory question are various *comparative* confirmational questions: Which of E or E' confirms H more strongly? Is E better evidence for H or H'? etc. These comparative questions could obviously be answered if we had the answer to an underlying *quantitative* question: To what degree does E confirm H? (Clearly if we knew the degree to which E confirms H and the degree to which E' confirms the same H, we could say whether E or E' confirms H more strongly.) Bayesian **confirmation measures** take propositions E and H and probability distribution Pr and try to quantify how much E confirms H relative to Pr.

There is a sizable literature on confirmation measures. Almost all of the measures that have been seriously defended are **relevance measures**: They agree with our earlier analysis that E confirms H relative to Pr just in case $\Pr(H \mid E) > \Pr(H)$. In other words, the relevance measures all concur that confirmation goes along with positive probabilistic relevance (and disconfirmation goes with negative probabilistic relevance). Yet there turn out to be a wide variety of confirmation measures satisfying this basic constraint. The following have all been extensively discussed in the historical literature:[28]

$$
\begin{aligned}
d(H, E) &= \Pr(H \mid E) - \Pr(H) \\
s(H, E) &= \Pr(H \mid E) - \Pr(H \mid {\sim}E) \\
r(H, E) &= \log \left[\frac{\Pr(H \mid E)}{\Pr(H)} \right] \\
l(H, E) &= \log \left[\frac{\Pr(E \mid H)}{\Pr(E \mid {\sim}H)} \right]
\end{aligned}
$$

These measures are to be read such that, for instance, $d(H, E)$ is the degree to which E confirms H relative to Pr on the d-measure. Each of the measures has been defined such that if H and E are positively relevant on Pr, then the value of the measure is positive; if H and E are negatively relevant, the value is negative; and if H is independent of E then the value is 0. In other words: positive values represent confirmation, negative values

represent disconfirmation, and 0 represents irrelevance.[29] For example, if Pr assigns each of the six faces on a die equal probability of coming up on a given roll, then

$$\begin{aligned} d(2, \text{prime}) &= \Pr(2 \,|\, \text{prime}) - \Pr(2) \\ &= 1/3 - 1/6 \\ &= 1/6 \end{aligned} \tag{6.7}$$

This value is positive because evidence that the die roll came up prime would confirm the hypothesis that it came up 2. Beyond the fact that it's positive, the particular value of the d-measure has little significance here. (It's not as if a d-value of, say, 10 has any particular meaning.) But the specific values do allow us to make comparisons. For example, $d(3 \vee 5, \text{prime}) = 1/3$. So according to the d-measure, evidence that the die roll came up prime more strongly supports the disjunctive conclusion that it came up 3 or 5 than the conclusion that the roll came up 2.

Since they are all relevance measures, the confirmation measures I listed will agree on *classificatory* facts about whether a particular E supports a particular H relative to a particular Pr. Nevertheless, they are distinct measures because they disagree about various *comparative* facts. A bit of calculation will reveal that $r(2, \text{prime}) = \log(2)$. Again, that particular number has no special significance, nor is there really much to say about how an r-score of $\log(2)$ compares to a d-score of $1/6$. (r and d measure confirmation on different scales, so to speak.) But it *is* significant that $r(3 \vee 5, \text{prime}) = \log(2)$ as well. According to the r-measure (sometimes called the "log ratio measure"), evidence that the roll came up prime confirms the hypothesis that it came up 2 to the *exact same degree* as the hypothesis that it came up either 3 or 5. That is a substantive difference with the d-measure on a comparative confirmation claim.

Since the various confirmation measures can disagree about comparative confirmation claims, to the extent that we are interested in making such comparisons we will need to select among the measures available. Arguing for some measures over others takes up much of the literature in this field. What kinds of arguments can be made? Well, we might test our intuitions on individual cases. For instance, it might just seem intuitively obvious to you that the primeness evidence favors the $3 \vee 5$ hypothesis more strongly than the 2 hypothesis, in which case you will favor the d-measure over the r-measure. Another approach parallels Hempel's approach to the qualitative confirmation relation: We first identify abstract features we want a confirmation measure to have, then we test positive proposals for each of those

features.

For example, suppose E confirms H strongly while E' confirms H only weakly. If we let c represent the "true" confirmation measure (whichever measure that turns out to be), $c(H, E)$ and $c(H, E')$ will both be positive numbers (because E and E' both confirm H), but $c(H, E)$ will be the larger of the two. Intuitively, since E is such good news for H it should also be very *bad* news for $\sim H$; since E' is only weakly good news for H it should be only weakly bad news for $\sim H$. This means that while $c(\sim H, E)$ and $c(\sim H, E')$ are both negative, $c(\sim H, E)$ is the lesser (farther from zero) of the two. That relationship is guaranteed by the following formal condition:

Hypothesis Symmetry: For all H and E in $\mathcal{L}$ and every probabilistic Pr, $c(H, E) = -c(\sim H, E)$.

Hypothesis Symmetry says that evidence which favors a hypothesis will disfavor the negation of that hypothesis just as strongly. It guarantees that if $c(H, E) > c(H, E')$ then $c(\sim H, E) < c(\sim H, E')$.[30]

Hypothesis Symmetry won't do all that much work in narrowing our field; of the confirmation measures under consideration, only r is ruled out by this condition. A considerably stronger condition can be obtained by following Carnap's thought that entailment and refutation are the two extremes of confirmation. On this line of thought, entailment is the strongest kind of confirmation one can get, and all entailments are equally-strong confirmations. Similarly, refutation is the strongest disconfirmation, and all refutations are equally disconfirming. It's only when we move from deductive arguments to inductive that we need fine-grained measures of intermediate degrees of support.[31] If this is right, then the correct confirmation measure $c(H, E)$ should satisfy:

Logicality: All entailments have the same degree of confirmation; all refutations have the same degree of confirmation; and all other confirmation cases fall strictly in-between.[32]

Logicality is violated by, for instance, confirmation measure d (often called the "difference measure"). It's easy to see why. d subtracts the prior of H from its posterior. Since the posterior can never be more than 1, the prior will therefore put a cap on how high d can get. For example, if $\Pr(H) = 9/10$, then no E will be able to generate a d-value greater than $1/10$, which is the value one will get when $E \models H$. On the other hand, we saw in Equation (6.7) that d-values greater than $1/10$ are possible even for evidence that doesn't entail the hypothesis (e.g. $d(2, \text{prime}) = 1/6$), simply

because the prior of the hypothesis in question begins so much lower. As with the firmness concept of confirmation, the prior of H interferes with the d-score's assessment of the relation *between* E and H. This interference generates a violation of Logicality.

Out of all the confirmation measures prominently defended in the historical literature (including all the measures described above), only measure l (the "log likelihood-ratio measure") satifies Logicality.[33] However, a new confirmation measure has recently been proposed (Crupi, Tentori, and Gonzalez 2007) that also satisfies Logicality:

$$z(H,E) = \begin{cases} \frac{\Pr(H \mid E) - \Pr(H)}{1 - \Pr(H)} & \text{if } \Pr(H \mid E) \geqslant \Pr(H) \\ \\ \frac{\Pr(H \mid E) - \Pr(H)}{\Pr(H)} & \text{if } \Pr(H \mid E) < \Pr(H) \end{cases}$$

This measure is particulary interesting because it measures confirmation differently from disconfirmation (hence the piecewise definition). That means confirmation and disconfirmation may satisfy different general conditions under the z-measure. For example, the following condition is satisfied for cases of disconfirmation but not for cases of confirmation:

$$z(H,E) = z(E,H) \tag{6.8}$$

Interestingly, Crupi, Tentori, and Gonzalez have conducted empirical studies in which subjects' comparative judgments seem to track z-scores better than the other confirmation measures. In particular, subjects seem intuitively to treat disconfirmation cases differently from confirmation cases. (See Exercise 6.8.)

6.4.2 Subjective Bayesian solutions to the Paradox of the Ravens

Earlier (Section 6.1.1) we saw Hempel endorsing conditions on confirmation according to which the hypothesis that all ravens are black would be confirmed not only by the observation of a black raven but also by the observation of a red herring. Hempel explained this result—the so-called Paradox of the Ravens—by arguing that its seeming paradoxicality results from background assumptions we illicitly smuggle into the question. Hempel set our immediate intuitive reactions aside and defended a positive theory of confirmation on which black ravens and red herrings stand symmetrically to the ravens hypothesis.

Subjective Bayesians take the paradox in exactly the opposite direction. They try to explain why our intuitive confirmation judgment makes sense, given our background assumptions about what the world is like. As Chihara puts it (in a slightly different context), the problem is "that of trying to see why we, who always come to our experiences with an encompassing complex web of beliefs", assess the paradox the way we do. (Chihara 1981, p. 437)

Take the current knowledge you actually have of what the world is like. Now suppose that against the background of that knowledge, you are told that you will soon be given an object a to observe. You will record whether it is a raven and whether it is black; you are not told in advance whether a will have either of these properties. Recall that on the Subjective Bayesian view of confirmation, evidence E confirms hypothesis H relative to probability distribution Pr just in case E is positively relevant to H on Pr. In this situation it's plausible that, when you gain evidence E about whether a is a raven and whether it is black, you will judge the confirmation of various hypotheses by this evidence relative to your personal credence function. So we will let your cr play the role of Pr.

The key judgment we hope to explain is that the ravens hypothesis (all ravens are black) is more strongly confirmed by the observation of a black raven than by the observation of a non-black non-raven (a red herring, say). One might go further and suggest that observing a red herring shouldn't confirm the ravens hypothesis at all. But if we look to our considered judgments (rather than just our first reactions) here, we should probably grant that insofar as a non-black raven would be absolutely disastrous news for the ravens hypothesis, any observation of a that doesn't reveal it to be a non-black raven should be at least *some* good news for the hypothesis.[34]

Expressing our key judgment formally requires us to measure confirmation strength, a topic we discussed in the previous section. If $c(H, E)$ measures the degree to which E confirms H relative to cr, the Bayesian claims that

$$c(H, Ba \,\&\, Ra) > c(H, {\sim}Ba \,\&\, {\sim}Ra) \tag{6.9}$$

where H is the ravens hypothesis $(\forall x)(Rx \supset Bx)$. Again, the idea is that relative to the credence function cr you assign before observing a, observing a to be a black raven would confirm H more strongly than observing a to be a non-black non-raven.

Fitelson and Hawthorne (2010b) show that Equation (6.9) will hold rel-

ative to cr if both the following conditions are met:

$$\mathrm{cr}(\sim Ba) > \mathrm{cr}(Ra) \tag{6.10}$$

$$\frac{\mathrm{cr}(\sim Ba \,|\, H)}{\mathrm{cr}(Ra \,|\, H)} \leqslant \frac{\mathrm{cr}(\sim Ba)}{\mathrm{cr}(Ra)} \tag{6.11}$$

These conditions are proposed as jointly sufficient for the confirmational result in Equation (6.9). They are not necessary; in fact, Bayesians have proposed a number of different sufficient sets over the years.[35] But these have the advantage of being simple and compact; they also work for every construal of c canvassed in the previous section except for confirmation measure s.

What do these conditions *say*? You satisfy Equation (6.10) if you are more confident prior to observing the object a that it will be non-black than you are that a will be a raven. This would make sense if, for example, you thought a was going to be randomly selected for you from a universe that contained more non-black things than ravens.[36] Equation (6.11) then considers the *ratio* of your confidence that a will be non-black to your confidence that it will be a raven. Meeting condition (6.10) makes this ratio greater than 1; now we want to know how the ratio would *change* were you to suppose all ravens are black. Equation (6.11) says that when you make this supposition the ratio doesn't go up—supposing all ravens are black wouldn't, say, dramatically increase how many non-black things you thought were in the pool or dramatically decrease your count of ravens. (It turns out from the math that for the confirmational judgment in Equation (6.9) to go false, the left-hand ratio in (6.11) would have to be *much* larger than the right-hand ratio; hence my talk of *dramatic* changes.) This constraint seems sensible. Under normal circumstances, for instance, supposing that all ravens are black should if anything increase the number of black things you think there are, not increase your count of *non*-black items.

Subjective Bayesians suggest that relative to our real-life knowledge of the world, were we to confront a selection situation like the one proposed in the ravens scenario, our credence distribution would satisfy Equations (6.10) and (6.11). Relative to such a credence distribution, the observation of a black raven confirms the ravens hypothesis more strongly than the observation of a red herring. This is how a Subjective Bayesian explains the key intuitive judgment that the ravens hypothesis is better confirmed by a black raven than by a red herring: by showing how that judgment follows from more general assumptions about the composition of the world. Given that people's outlook on the world typically satisfies Equations (6.10) and

(6.11), it follows from the Subjective Bayesian's quantitative theory of confirmation that if they are rational they will take the black raven observation to be more highly confirmatory.[37]

Now one might object that people who endorse the key ravens judgment have credence functions that don't actually satisfy the conditions specified (or other sets of sufficient conditions Bayesians have proposed). Or an Objective Bayesian might argue that a confirmation judgment can be vindicated only by grounding it in something firmer than personal credences. I am not going to take up those arguments here. But I hope to have at least fought back the charge that Subjective Bayesianism about confirmation is empty. The Subjective Bayesian account of confirmation tells us when evidence E confirms hypothesis H relative to credence distribution cr. You might think that because it does very little to constraint the values of cr, this account can tell us nothing interesting about when evidence confirms a hypothesis. But we have just seen a substantive, unexpected result. It was not at all obvious at the start of our inquiry that any rational credence distribution satisfying Equations (6.10) and (6.11) would endorse the key ravens judgment. Any Subjective Bayesian result about confirmation will have to take the form, "If your credences are such-and-such, then these confirmation relations follow," but such conditionals can nevertheless be highly informative.

For instance, the result we've just seen not only reveals what confirmation judgments agents will make in typical circumstances, but also which atypical circumstances may legitimately undermine those judgments. Return to the Hall of Atypically-Colored Birds, where a bird is displayed only if the majority of his species-mates are one color but his color is different. Suppose it is part of an agent's background knowledge (before she observes object a) that a is to be selected from the Hall of Atypically-Colored Birds. If at that point—before observing a—the agent were to suppose that all ravens are black, that would dramatically decrease her confidence that a will be a raven. If all ravens are black, there are no atypically-colored ravens, so there should be no ravens in the Hall.[38] Thus given the agent's background knowledge about the Hall of Atypically-Colored Birds, supposing the ravens hypothesis H decreases her confidence that a will be a raven (that is, Ra). This makes the lefthand side of Equation (6.11) greater than the righthand side, and renders Equation (6.11) false. So one of the sufficient conditions in our ravens result fails, and Equation e:ravconf cannot be derived. This provides a tidy explanation of why, if you know you're in the Hall of Atypically-Colored Birds, observing a black raven should not necessarily be better news for the ravens hypothesis than observing a non-black non-raven.

Besides this account of the Paradox of the Ravens, Subjective Bayesians have offered solutions to various other confirmational puzzles. For example, we can approach the problem of irrelevant conjunction (Section 6.1.2) by specifying conditions under which adding an irrelevant conjunct to a confirmed hypothesis yields a new hypothesis that—while still confirmed—is less strongly confirmed than the original. (Hawthorne and Fitelson 2004) Similarly, Chihara (1981) and Eells (1982, Ch. 2) respond to Goodman's grue example (Section 6.3) by specifying credal conditions under which a run of observed green emeralds more strongly confirms the hypothesis that all emeralds are green than the hypothesis that all emeralds are grue.

Even more intriguingly, the Subjective Bayesian account of confirmation has recently been used to explain what look like *irrational* judgments on the part of agents. The idea here is that sometimes when subjects are asked questions about probability, they respond with answers about confirmation. In Tversky and Kahneman's Conjunction Fallacy experiment (Section 2.2.3), the hypothesis that Linda is a bank teller is entailed by the hypothesis that Linda is a bank teller and active in the feminist movement. This entailment means that an agent satisfying the probability axioms must be at least as confident in the former hypothesis as the latter. But it does *not* mean that evidence must confirm the former as strongly as the latter. Crupi, Fitelson, and Tentori (2008) outline credal conditions under which the evidence presented to subjects in Tversky and Kahneman's experiment would confirm the feminist-bank-teller hypothesis more strongly than the bank-teller hypothesis. It may be that subjects who rank the feminist-bank-teller hypothesis more highly in light of that evidence are reporting confirmational judgments instead of credences.

Similarly, in analyzing the Base Rate Fallacy (Section 4.1.2) we noted the strong Bayes factor of the evidence one gets from a highly reliable disease test. Since the Bayes factor tracks the likelihood ratio measure of confirmation, this tells us that a positive result from a reliable test strongly confirms that the patient has the disease (as it should!). When doctors are asked for their confidence that the patient has the disease in light of such a positive test result, the high values they report may reflect their confirmational judgments.

The Subjective Bayesian account of confirmation may therefore provide an explanation of what subjects are doing when they seem to make irrational credence reports. Nevertheless, having an explanation for subjects' behavior does not change the fact that these subjects may be making serious *mistakes*. It's one thing when a doctor is asked in a study to report a credence value and reports a confirmation value instead. But if the doctor goes on to

make treatment decisions based on the confirmation value rather than the posterior probability, this can have significant consequences. Confusing how probable a hypothesis is on some evidence with how strongly that hypothesis is confirmed by the evidence is a version of the firmness/increase-in-firmness conflation. If the doctor recommends a drastic treatment for a patient on the basis that the test applied was highly reliable (even though, with the base rates taken into account, the posterior probability that a disease is present remains quite low), her confusion about probability and confirmation may prove highly dangerous for her patient.

6.5 Exercises

Unless otherwise noted, you should assume when completing these exercises that the distributions under discussion satisfy the probability axioms and Ratio Formula. You may also assume that whenever a conditional probability expression occurs, the needed proposition has nonzero unconditional credence so that conditional probabilities are well-defined.

Problem 6.1. Suppose the Special Consequence Condition and Converse Consequence Condition were both true. Show that under those assumptions, if evidence E confirms some proposition H relative to K, then relative to K evidence E will also confirm any other proposition X we might choose.* (Hint: Start with the problem of irrelevant conjunction.)

Problem 6.2. For purposes of this problem, assume that the Equivalence Condition, the Entailment Condition, and Disconfirmation Duality are all true of the confirmation relation.

(a) Show that if $E \,\&\, K$ deductively refutes H but K does not refute H on its own, then E disconfirms H relative to K.

(b) Show that if $H \,\&\, K$ deductively refutes E but K does not refute H on its own, then E disconfirms H relative to K.

Problem 6.3. Suppose we have a language whose only atomic propositions are $Fa_1, Fa_2, \ldots, Fa_n$ for some integer $n > 1$. In that case, $\mathfrak{m}^\dagger(Fa_n) = 1/2$.

(a) Show that for any proposition E expressible solely in terms of Fa_1 through Fa_{n-1}, $\mathfrak{m}^\dagger(Fa_n \,|\, E) = 1/2$.

*For purposes of this problem you may assume that E, H, X, and K stand in no special logical relationships.

(b) What does the result you demonstrated in part (a) have to do with Carnap's point that $\mathfrak{m}^\dagger$ does not allow "learning from experience"?

Problem 6.4. (a) Make a stochastic truth-table for the four atomic sentences Fa, Fb, Ga, Gb. In the right-hand column, enter the values Carnap's $\mathfrak{m}^*$ assigns to each state-description. (Hint: Keep in mind that $Fa \,\&\, {\sim}Fb \,\&\, Ga \,\&\, {\sim}Gb$ belongs to a different structure-description than $Fa \,\&\, {\sim}Fb \,\&\, {\sim}Ga \,\&\, Gb$.)

(b) Use your table to show that $\mathfrak{m}^*(Fb \,|\, Fa \,\&\, Ga \,\&\, Gb) > \mathfrak{m}^*(Fb \,|\, Fa)$.

(c) Use your table to show that $\mathfrak{m}^*(Fb \,|\, Fa \,\&\, Ga \,\&\, {\sim}Gb) = \mathfrak{m}^*(Fb)$.

(d) For each of problem (b) and (c) above, explain how your answer relates to $\mathfrak{m}^*$'s handling of "analogical effects".[†]

Problem 6.5. Suppose E'' is the proposition

$$(Ga_1 \,\&\, Oa_1) \,\&\, (Ga_2 \,\&\, Oa_2) \,\&\, \ldots \,\&\, (Ga_{99} \,\&\, Oa_{99}) \,\&\, {\sim}Oa_{100}$$

Without actually making a stochastic truth-table, argue convincingly that on Carnap's confirmation theory:

(a) $\mathfrak{m}^*(Ga_{100} \equiv Oa_{100} \,|\, E'') = \mathfrak{m}^*({\sim}Ga_{100} \,|\, E'')$

(b) $\mathfrak{m}^*(Ga_{100} \equiv Oa_{100} \,|\, E'') = \mathfrak{m}^*(Ga_{100} \,|\, E'')$

(c) $\mathfrak{m}^*(Ga_{100} \,|\, E'') = 1/2$

(d) $\mathfrak{m}^*(Ga_{100}) = 1/2$

(e) $\mathfrak{m}^*(Ga_{100} \equiv Oa_{100} \,|\, E'') = \mathfrak{m}^*(Ga_{100} \,|\, E'') = \mathfrak{m}^*(Ga_{100})$

Problem 6.6. Provide examples showing that the r-measure of confirmation violates each of the following constraints:

(a) Hypothesis Symmetry

(b) Logicality

Problem 6.7. Statisticians have favored the l-measure of confirmation because it has a convenient mathematical property: If bodies of evidence E_1 and E_2 are independent conditional on both H and ${\sim}H$, then the degree to which their conjunction confirms H can be found by summing the degrees to which E_1 and E_2 confirm H individually. Prove that the l-measure has this property. (Hint: Remember that $\log(x \cdot y) = \log x + \log y$.)

[†]I owe this entire problem to Branden Fitelson.

Problem 6.8. Crupi, Tentori, and Gonzalez think it's intuitive that on whatever measure c correctly gauges confirmation, the following constraint will be satisfied for cases of disconfirmation but not confirmation:

$$c(H, E) = c(E, H)$$

(a) Provide an example of a real-world E and H such that, intuitively, E confirms H but H does not confirm E to the same degree. (Don't forget to specify what Pr distribution you're relativizing your confirmation judgments to!)

(b) Provide an example of a real-world E and H such that, intuitively, E disconfirms H and H disconfirms E to the same degree. (Don't make it too easy on yourself—pick an E and H that are not logically equivalent to each other!)

(c) Does it seem to you intuitively that for any E, H, and Pr such that E disconfirms H, H will disconfirm E to the same degree? Explain why or why not.

Problem 6.9. The solution to the Paradox of the Ravens presented in Section 6.4.2 is not the only Subjective Bayesian solution that has been defended. An earlier solution invoked the following four conditions (where H abbreviates $(\forall x)(Rx \supset Bx)$):

(i) $\Pr(Ra \,\&\, {\sim}Ba) > 0$

(ii) $\Pr({\sim}Ba) > \Pr(Ra)$

(iii) $\Pr(Ra \,|\, H) = \Pr(Ra)$

(iv) $\Pr({\sim}Ba \,|\, H) = \Pr({\sim}Ba)$

Assuming Pr satisfies these conditions, complete each of the following. (Hint: Feel free to write H instead of the full, quantified proposition it represents, but don't forget what H entails about Ra and Ba.)

(a) Prove that $\Pr({\sim}Ra \,\&\, {\sim}Ba) > \Pr(Ra \,\&\, Ba)$.

(b) Prove that $\Pr(Ra \,\&\, Ba \,\&\, H) = \Pr(H) \cdot \Pr(Ra)$.

(c) Prove that $\Pr({\sim}Ra \,\&\, {\sim}Ba \,\&\, H) = \Pr(H) \cdot \Pr({\sim}Ba)$.

(d) Show that on confirmation measure d, if Pr satisfies conditions (i) through (iv) then $Ra \,\&\, Ba$ confirms H more strongly than ${\sim}Ra \,\&\, {\sim}Ba$ does.

(e) Where in your proofs did you use condition (i)?

(f) Suppose Pr is your credence distribution when you know you are about to observe an object a drawn from the Hall of Atypically-Colored Birds. Which of the conditions (i) through (iv) will Pr probably not satisfy? Explain.

6.6 Further reading

Introductions and Overviews

> Ellery Eells (1982). *Rational Decision and Causality*. Cambridge Studies in Philosophy. Cambridge: Cambridge University Press

The latter part of Chapter 2 (pp. 52–64) offers an excellent discussion of Hempel's adequacy conditions for confirmation, how the correct ones are met by a probabilistic approach, and Subjective Bayesian solutions to the Paradox of the Ravens and Goodman's grue puzzle.

> Alan Hájek and James M. Joyce (2008). Confirmation. In: *The Routledge Companion to Philosophy of Science*. Ed. by Stathis Psillos and Martin Curd. New York: Routledge, pp. 115–128

Besides providing an overview of much of the material in this chapter, suggests that there may not be one single correct function for measuring degree of confirmation.

Classic Texts

> Carl G. Hempel (1945a). Studies in the Logic of Confirmation (I). *Mind* 54, pp. 1–26

> Carl G. Hempel (1945b). Studies in the Logic of Confirmation (II). *Mind* 54, pp. 97–121

Hempel's classic papers discussing his adequacy conditions on the confirmation relation and offering his own positive, syntactical account of confirmation.

> Rudolf Carnap (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press

While much of the material earlier in this book is crucial for motivating Carnap's probabilistic theory of confirmation, his discussion of functions $\mathfrak{m}^\dagger$ and $\mathfrak{m}^*$ occurs in the Appendix. (Note that the preface distinguishing "firmness" from "increase in firmness" conceptions of confirmation does not appear until the second edition of this text, in 1962.)

> Janina Hosiasson-Lindenbaum (1940). On Confirmation. *Journal of Symbolic Logic* 5, pp. 133–148

Early suggestion that the Paradox of the Ravens might be resolved by first admitting that both a black raven and a red herring confirm that all ravens are black, but then second arguing that the former confirms more strongly than the latter.

> Nelson Goodman (1979). *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press

Chapter III contains Goodman's "grue" discussion.

Extended Discussion

> Branden Fitelson (2012). Evidence of Evidence is Not (Necessarily) Evidence. *Analysis* 72, pp. 85–88

Discusses counterexamples to the "evidence of evidence is evidence" principle, based on counterexamples to Confirmation Transitivity.

> Michael G. Titelbaum (2010). Not Enough There There: Evidence, Reasons, and Language Independence. *Philosophical Perspectives* 24, pp. 477–528

Proves a general language-dependence result for all objective accounts of confirmation (including accounts that are Objective Bayesian in the normative sense), then evaluates the result's philosophical significance.

> Katya Tentori, Vincenzo Crupi, and Selena Russo (2013). On the Determinants of the Conjunction Fallacy: Probability versus Inductive Confirmation. *Journal of Experimental Psychology: General* 142, pp. 235–255

Assessment of various explanations of the Conjunction Fallacy in the psychology literature, including the explanation that subjects are reporting confirmation judgments.

Notes

[1]Scientists—and philosophers of science—are interested in a number of properties and relations of evidence and hypotheses besides confirmation. These include predictive power, informativeness, simplicity, unification of disparate phenomena, etc. An interesting ongoing Bayesian line of research asks whether and how these various other notions relate to confirmation.

[2](Good 1967) offers a more detailed example in the same vein. Good describes the population distributions of two worlds such that observing a black raven confirms that one is in the world in which not all ravens are black.

[3]As I pointed out in Chapter 4's note 11, this passage may have been the inspiration for David Lewis's referring to hypothetical priors (credence distributions reflecting no contingent evidence) as "superbaby" credences.

[4]In discussing the Paradox of the Ravens, one might wonder in general whether $(\forall x)(Rx \supset Bx)$—especially with its material conditional, and its curious existential import—is a faithful translation of "All ravens are black." Strictly speaking, Hempel's discussion is an examination of what confirms the sentence in logical notation, rather than the sentence in English. But if the two come apart, intuitions about "All ravens are black" may be less relevant to Hempel's discussion.

[5]Despite his attention to background corpora, Hempel isn't careful about backgrounds in the adequacy conditions he proposes. So I will add those background specifications as we work through the various conditions, and explain their motivations as we go along.

[6]One might want a restriction to keep the Special Consequence Condition from applying when $K \models H'$, but in the stated counterexamples H' is not entailed by the background. Out of desperation we could try to save Special Consequence by claiming it holds only relative to tautological backgrounds (as Hempel did with Nicod's Criterion). But we can recreate our cards counterexample to Special Consequence by emptying out the background and adding facts about how the card was drawn as conjuncts to each of A, B, and C. Similar remarks apply to the counterexamples we'll soon produce for other putative confirmation constraints.

[7]For one of many recent articles on confirmational intransitivity and skepticism, see (White 2006).

[8]Another argument for the Consistency Condition would be if you thought confirmation of a hypothesis meant we should accept that hypothesis, and also that one should never accept inconsistent propositions. But we've already rejected that interpretation of confirmation for our purposes.

[9]I'm assuming the definition of an ostrich includes its being a flightless bird, and whatever K is involved doesn't entail E, H, or H' on its own.

[10]**Hypothetico-deductivism** is a positive view of confirmation that takes the condition in Converse Entailment to be not only sufficient but also necessary for confirmation: E confirms H relative to K *just in case* $H \,\&\, K \models E$ and $K \not\models E$. This is implausible for a number of reasons (see (Hempel 1945b)). Here's one: Evidence that a coin of unknown bias has come up heads on exactly half of a huge batch of flips supports the hypothesis that the coin is fair; yet that evidence isn't *entailed* by that hypothesis.

[11]Strictly speaking there will be infinitely many X in $\mathcal{L}$ such that $\Pr(X) = 1$, so we will take K to be a proposition in $\mathcal{L}$ logically equivalent to the conjunction of all such X. I'll ignore this detail in what follows.

[12]Carnap's preface to the second edition distinguishes the firmness and increase in firmness concepts because he had equivocated between them in the first edition. Carnap was

roundly criticized for this by Popper.

[13]Despite my intense awareness of the issue I even made this mistake in an article once myself—I was lucky to have my misstep caught before the offending piece was published!

[14]Here we assume that, as pointed out in Chapter 2's note 5, the atomic sentences of $\mathcal{L}$ are logically independent.

[15]A word about Carnap's notation in his (1950). Carnap actually introduces two confirmation functions, $\mathfrak{m}(\cdot)$ and $\mathfrak{c}(\cdot,\cdot)$. For any non-contradictory proposition K in $\mathcal{L}$, $\mathfrak{c}(\cdot, K)$ is just the function I've been describing as $\Pr(\cdot)$ relative to K; in other words, $\mathfrak{c}(\cdot, K) = \mathfrak{m}(\cdot \mid K) = \mathfrak{m}(\cdot \,\&\, K)/\mathfrak{m}(K)$. As I've just mentioned in the main text, this makes $\mathfrak{c}$ somewhat redundant in the theory of confirmation, so I won't bring it up again.

[16]As I mentioned in Chapter 5, note 8, Carnap actually thinks "probability" is ambiguous between two meanings. What he calls "probability$_1$" is the logical notion of probability we've been discussing. Carnap's "probability$_2$" is based on frequencies, and is therefore objective as well.

[17]Among other things, $\mathfrak{m}^\dagger$ represents the probability distribution Ludwig Wittgenstein proposed in his *Tractatus Logico-Philosophicus*. (Wittgenstein 1921/1961)

[18]Formally, two state-descriptions are disjuncts of the same structure-description just in case one state-description can be obtained from the other by permuting its constants.

[19]In his (1979, p. 73, n. 9), Goodman says the grue problem is "substantially the same" as the problem he offered in (Goodman 1946). The earlier version of the problem is both more clearly laid-out and cleaner from a logical point of view. For instance, instead of green and blue, he there uses red and not-red. The earlier paper also makes clearer exactly whose positive theories of confirmation Goodman takes the problem to target.

[20]For simplicity's sake I'm going to assume Goodman is criticizing the version of Carnap's theory committed to $\mathfrak{m}^*$; subsequent changes Carnap made to handle analogical effects make little difference here.

[21]Compare the difficulties with partition selection we encountered for indifference principles in Section 5.3.

[22]Hume's (1739–40/1978) problem of induction asked what justifies us in projecting *any* correlations that have occurred in the past into the future. Goodman's "new riddle of induction" asks, given that we are justified in projecting some correlations, how to sort out *which* ones to project.

[23]Hempel's theory of confirmation displays a similar effect. And really, any close reader of Hempel should've known that some of Goodman's claims against Hempel were overstated. I mentioned that Hempel endorses the Consistency Condition (Section 6.1.2); he goes on to prove that it is satisfied by his positive theory of confirmation. On Hempel's theory, the hypotheses confirmed by any piece of evidence must be consistent both with that evidence and with each other. So *contra* Goodman, it just can't be that on Hempel's theory we get the "devastating result that any statement will confirm any statement." (1979, p. 81)

[24]For more on this topic, see (Hooker 1968), (Fine 1973, Ch. VII), (Maher 2010), and (Titelbaum 2010).

[25]Earlier I defined the background corpus of a probability distribution as the conjunction of propositions to which that distribution assigns unconditional probability 1. Since we can always recover background K from distribution Pr in this way, the Subjective Bayesian need not make confirmation of H on E relative to both a probability distribution Pr *and* a background corpus K. Pr takes care of the K, so to speak. Going the other way, Carnap used his $\mathfrak{m}$-functions to generate a specific Pr-distribution for each K, so on his view K took care of Pr and confirmation of H on E needed be relative only to K.

[26]As we put it in Chapter 5, the Subjective Bayesian need not be an extreme Subjective Bayesian, who denies any constraints on rational hypothetical priors beyond the probability axioms.

[27]Here's another way to put the same point: Carnap offered a substantive theory of confirmation as a *two-place* relation; there were interesting facts to be had about which bodies of evidence (first relatum) confirmed which hypotheses (second relatum). Subjective Bayesians, on the other hand, see confirmation as a *three-place* relation, between bodies of evidence, hypotheses, and probability distributions. Sometimes (when rational requirements narrow the permissible range of probability distributions) this will generate interesting two-place facts. But in other cases, the only substantive confirmation facts to be found take into account all three relata—they are relative to the particular probability distribution specified. As we will see in Section 6.4.2, these three-place facts can still be highly informative.

[28]For citations to various historical authors who defended each measure, see (Eells and Fitelson 2002).

[29]The logarithms have been added to the r- and l-measures to achieve this centering on 0. Removing the logarithms would yield measures ordinally equivalent to their logged versions, but whose values ran from 0 to infinity (with a value of 1 indicating probabilistic independence). Notice also that the base value of the logarithms is irrelevant for our purposes.

[30]Hypothesis Symmetry was defended as a constraint on degree of confirmation by (Kemeny and Oppenheim 1952); see also (Eells and Fitelson 2002), who gave it that particular name.

[31]Carnap thought of confirmation as a "generalization of entailment" in a number of senses. Many Subjective Bayesians are happy to accept Carnap's idea that deductive cases are limiting cases of confirmation. But they aren't willing to follow Carnap in taking those limiting cases as a model for the whole domain. Whether E entails H relative to K depends just on the content of those propositions, and Carnap thought matters should be the same for all confirmatory relations. To a Subjective Bayesian, though, whether E confirms H relative to K depends on something more—a probability distribution Pr.

[32]See (Fitelson 2006) for Logicality.

[33]A few technical notes: First, when $E \models H$ the denominator in l goes to zero. We think of l as assigning an infinite positive value in these cases, and an infinite negative value when E refutes H. Second, any confirmation measure ordinally equivalent to l (such as l without the logarithm out front) will satisfy Logicality as well. Third, in discussing Logicality I am restricting my attention to "contingent" cases, in which neither E nor H is entailed or refuted by the K associated with Pr.

[34]Another thought one might have is that while a red herring confirms that all ravens are black, its degree of confirmation of that hypothesis is exceedingly weak in absolute terms. While some Bayesian analyses of the paradox also try to establish that result, we won't consider it here. (See (Vranas 2004) for discussion and citations on the proposal that a red herring confirms the ravens hypothesis to a degree that is "positive but minute.")

[35]For citations to many historical proposals, see (Fitelson and Hawthorne 2010a, esp. n. 10). (Fitelson and Hawthorne 2010b) goes beyond these historical sources by also proposing necessary conditions for Equation (6.9), which are unfortunately too complex to delve into here.

[36]The result assumes you assign non-extreme unconditional credences to the proposition that a is black and to the proposition that it's a raven. This keeps various denominators in the Ratio Formula positive. We also assume you have a non-extreme prior in H.

[37]Why "if they are rational"? The mathematical result assumes not only that the credence distribution in question satisfies Equations (6.10) and (6.11), but also that it satisfies the probability axioms and Ratio Formula. (This allows us to draw out conclusions about values in the credence distribution beyond what is directly specified in Equations (6.10) and (6.11).) The probability axioms and Ratio Formula are among the constraints Subjective Bayesians assume a rational credence distribution must satisfy.

[38]Perhaps even with the supposition that all ravens are black, the agent's confidence that a will be a raven is slightly above zero because once in a long while the Hall's curators make a mistake.