Paradoxes of confirmation

Many sorts of reasoning - in both science and in everyday life - involve adjusting our view of the world based on evidence. Quite often, this is a matter of adjusting our views about general claims on the basis of evidence about particular things; for example, one might adjust one's views on the question of how good ND's 2014 football team is based on the evidence of their performance in the first game against Rice. In science, we adjust our views about the acceptability of various theories on the basis of the results of particular experimental findings.

For this to make sense, it must be the case that sometimes a particular piece of data counts in favor of a theory, and that sometimes it counts against it. As we'll use the terms, some data **confirms** a theory if it counts in favor of it, and **disconfirms** the theory if it counts against it. (In ordinary English, "confirms" means something stronger - it means something like "shows to be true" - it is important to keep in mind that evidence can count in favor of a theory without showing that theory to be true.)

One might, then, want to know **when** some evidence confirms a theory. The attempt to answer this question is known as **confirmation theory**. The paradoxes we'll be discussing today are paradoxes which arise within confirmation theory. These are thus analogous to Newcomb's problem and the prisoner's dilemma, which are paradoxes which arise within decision theory - that is, they are paradoxes which conflict with intuitively quite plausible principles about rational choice. The paradoxes we will discuss today are cases which conflict with intuitively quite plausible principles about when evidence counts in favor of -confirms - a theory.

Suppose we are interested in figuring out whether some general claim - for example,

All emeralds are green.

is true. What would count as evidence for this general claim? One might think that, whatever else we might think, this sort of general claim is confirmed by its instances - cases of individual emeralds being green. If we come across a green emerald, the thought goes, this confirms the generalization. (It does not, of course, show it to be true - but it does seem to count in favor of it.)

This suggests the following rule:

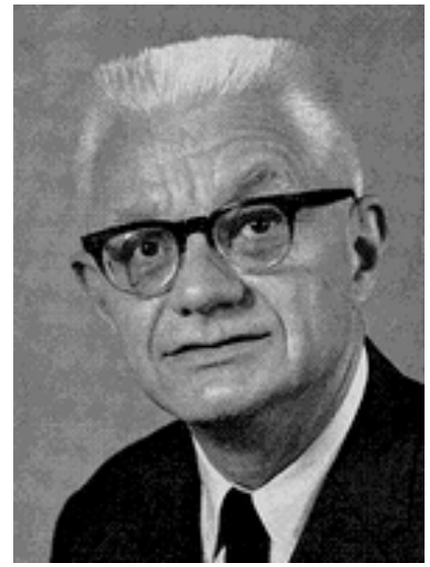
G. A generalization is confirmed by all of its instances.

This rule seems, at first glance to be obviously correct; indeed, if something like this is not correct, it is sort of hard to see how we could get started with confirmation theory at all. Nonetheless, this principle is the target of both of the paradoxes we'll be discussing today.

G. A generalization is confirmed by all of its instances.

This rule seems, at first glance to be obviously correct; indeed, if something like this is not correct, it is sort of hard to see how we could get started with confirmation theory at all. Nonetheless, this principle is the target of both of the paradoxes we'll be discussing today.

The first of these paradoxes is the **paradox of the ravens**, which was discovered by the German philosopher Carl Hempel, who was one of the many great European philosophers to leave the continent for America in the years leading up to World War II.



Hempel noticed that G, while seemingly innocuous, leads to quite surprising conclusions when combined with another thesis which, at first glance, seems obviously correct.

EQUIV. If two theories, T1 and T2, are known to be equivalent - in the sense that, necessarily, T1 is true if and only if T2 is - then any evidence which confirms T1 also confirms T2, and vice versa.

Why should EQUIV seem clearly correct? Well, if we are sure that T1 and T2 stand or fall together - that if one is true both are - then it seems like any evidence that one is true should also be evidence that the other is true. But that is just what EQUIV says.

Now suppose that we are interested in gathering evidence for the following theory:

R. All ravens are black.

I now triumphantly produce, in support of R, a white piece of chalk. Have I succeeded in providing data which confirms - counts in favor of - R? It seems not.

But now consider the following theory:

NR. All non-black things are non-ravens.

Have I provided evidence which confirms NR? It seems so; after all, I have provided a non-black thing which is also a non-raven, which means that I have provided an instance of the generalization NR. So by G my white piece of chalk confirms NR.

But now think about the relationship between R and NR. You may notice that they are logically equivalent: it is impossible for one to be true unless the other is. So, by EQUIV, it follows that I have, with my white piece of chalk, provided evidence that all ravens are black.

G. A generalization is confirmed by all of its instances.

EQUIV. If two theories, T1 and T2, are known to be equivalent - in the sense that, necessarily, T1 is true if and only if T2 is - then any evidence which confirms T1 also confirms T2, and vice versa.

R. All ravens are black.

I now triumphantly produce, in support of R, a white piece of chalk. Have I succeeded in providing data which confirms - counts in favor of - R? It seems not.

But now consider the following theory:

NR. All non-black things are non-ravens.

Have I provided evidence which confirms NR? It seems so; after all, I have provided a non-black thing which is also a non-raven, which means that I have provided an instance of the generalization NR. So by G my white piece of chalk confirms NR.

But now think about the relationship between R and NR. You may notice that they are logically equivalent: it is impossible for one to be true unless the other is. So, by EQUIV, it follows that I have, with my white piece of chalk, provided evidence that all ravens are black.

It appears that something has gone badly wrong. We have a paradox, because a pair of plausible premises - G and EQUIV - seems to imply an obvious falsehood - that a white piece of chalk is evidence for R.

One might respond to the paradox by denying EQUIV, and saying that generalizations are only ever confirmed by their instances - they are never confirmed by the instances of logically equivalent generalizations. But, as Sainsbury points out, this seems implausible. If we are wondering whether everyone who has a disease was exposed to chemical X, it seems as though we could provide support for this claim by showing that everyone **not** exposed to chemical X does not have the disease.

However, one can construct cases which seem to be counterexamples to principles like EQUIV. Consider the following piece of evidence:

(C) The card which is facedown on the table is a face card.

This seems to confirm:

(T1) The card on the table is a red jack.

But does not seem to confirm:

(T2) The card on the table is red.

G. A generalization is confirmed by all of its instances.

EQUIV. If two theories, T1 and T2, are known to be equivalent - in the sense that, necessarily, T1 is true if and only if T2 is - then any evidence which confirms T1 also confirms T2, and vice versa.

R. All ravens are black.

NR. All non-black things are non-ravens.

(C) The card which is facedown on the table is a face card.

This seems to confirm:

(T1) The card on the table is a red jack.

But does not seem to confirm:

(T2) The card on the table is red.

But (T2) is a logical consequence of (T1) - so why, if (C) confirms (T1), does not not also confirm every other theory that must be true if (T1) is? More to the point: if evidence can confirm a theory without confirming logical consequences of that theory, there seems no reason why EQUIV could not be false.

What's going on with the example of the card? The basic idea seems to be this. There are 52 possibilities for what that card can be and hence, before learning (C), one regards the probability of (T1) being true as 1 in 52. Upon learning (C), the relevant possibilities are reduced to the 12 face cards, which makes the probability of (T1) being true 1 in 12 - hence (C) confirms (T1), because it makes it more likely that - is evidence that - (T1) is true.

Before learning (C), there is a 26/52 chance - i.e., ½ - that (T2) is true. After (C) narrows the relevant possibilities to the 12 face cards, there is a 6/12 chance - still ½ - that (T2) is true. So (C) does not confirm (T2).

What this shows is that sometimes evidence can eliminate possibilities in such a way that it makes one theory A significantly more likely to be true but does not do this for a second theory B, even though every possibility in which A is true is also one in which B is true. **This is possible if B is true in some possibilities in which A is not, and the possibilities eliminated by the evidence eliminate a greater proportion of those in which A is false than those in which B is false**.

But this is also enough to show that the example of the cards is no challenge to principle EQUIV. After all, if two theories are logically equivalent, **then they are true in exactly the same possibilities** - and then no evidence can eliminate possibilities in such a way as to favor one over the other.

G. A generalization is confirmed by all of its instances.

EQUIV. If two theories, T1 and T2, are known to be equivalent - in the sense that, necessarily, T1 is true if and only if T2 is - then any evidence which confirms T1 also confirms T2, and vice versa.

R. All ravens are black.

NR. All non-black things are non-ravens.

Hence it seems that, because we don't have an explanation of how EQUIV could be false, we still lack any explanation of what is going on in the paradox of the ravens.

This might lead us to think that the paradox should be solved by rejecting G, rather than EQUIV. And in fact it is plausible to regard this as the lesson of the next paradox of confirmation we will be discussing: the puzzle of "grue", sometimes also called the "new riddle of induction."

This paradox is due to Nelson Goodman, one of the most important American philosophers of the 20th century.

Goodman's aim in his book *Fact, Fiction, and Forecast* was to show that simple principles of confirmation like G were false; he did this by inventing the predicate "grue", which is defined as follows:

x is grue if and only if either:
    x is green, and has been observed before 4/7/14
    x is blue, and has not been observed before 4/7/14

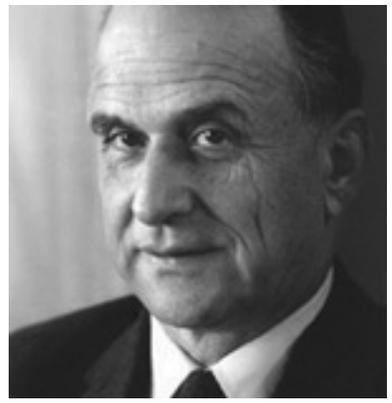
It is important to see, first, that this is a perfectly legitimate definition; it succeeds in classifying all objects as either grue or non-grue.

But suppose that we enumerate all of the emeralds which have been observed so far, and consider the following pieces of data:

G. A generalization is confirmed by all of its instances.

EQUIV. If two theories, T1 and T2, are known to be equivalent - in the sense that, necessarily, T1 is true if and only if T2 is - then any evidence which confirms T1 also confirms T2, and vice versa.

Goodman's aim in his book *Fact, Fiction, and Forecast* was to show that simple principles of confirmation like G were false; he did this by inventing the predicate "grue", which is defined as follows:



x is grue if and only if either:
x is green, and has been observed before 4/7/14
x is blue, and has not been observed before 4/7/14

But suppose that we enumerate all of the emeralds which have been observed so far, and consider the following pieces of data:

Emerald #1 is grue.
Emerald #2 is grue.
….
Emerald #N is grue.

If G were true, this would seem - given that N is a large number - to provide very strong evidence for the generalization that all emeralds are grue - which would in turn imply that an emerald discovered tomorrow will be blue. But of course our evidence confirms no such claim. Hence G must be false.

On the other hand, it is hard to see how G could **just** be false - don't instances in at least some cases confirm the corresponding generalizations? Perhaps what we need to do is to restrict G in some way so that it does not lead to conclusions like the above, but still is useful in explaining the relationship between confirming evidence and theory in the case of real scientific hypotheses.

This suggests a plausible thought: perhaps we should restrict G so that it applies only to generalizations which involve **suitable scientific vocabulary** - perhaps we should restrict G by excluding predicates which cause the sorts of trouble caused by "grue" ("gruesome predicates").

G. A generalization is confirmed by all of its instances.

x is grue if and only if either:
x is green, and has been observed before 4/7/14
x is blue, and has not been observed before 4/7/14

This suggests a plausible thought: perhaps we should restrict G so that it applies only to generalizations which involve **suitable scientific vocabulary** - perhaps we should restrict G by excluding predicates which cause the sorts of trouble caused by “grue” (“gruesome predicates”).

To pursue this thought, we need to be able to say what a gruesome predicate is - that is, we need to be able to say what, exactly, is so bad about “grue.” This turns out to be harder than you might think.

A first thought is that the problem is due to “grue” being a made-up word. But this won’t get us very far - after all, scientific theories introduce new scientific terms all the time, and these are “made up” in just the way that “grue” is - they are new terms defined in terms of existing vocabulary.

A more promising idea is that the problem with “grue” is that it is defined in terms of a particular **time**.

However, there are a few problems with this suggestion. One is that any predicate can be given a similarly time-indexed definition. For suppose that we define a new term, “bleen”, as follows: x is bleen if and only if either x is blue, and has been observed before 4/7/14, or x is green, and has not been observed before 4/7/14. Using “grue” and “bleen” we can then give the following definition of “blue”:

x is blue if and only if either:
x is bleen, and has been observed before 4/7/14
x is grue, and has not been observed before 4/7/14

But surely this should not be enough to rule out our using G to evaluate generalizations about which things are blue!

One might reply: “Yes, one **can** define “blue” this way - but we don’t **have** to. The difference between “grue” and “blue” is that no one could understand “grue” without this sort of time-indexed definition.” This suggests that we should exclude terms which are impossible to understand except via a time-indexed definition.

One might wonder why we should be so sure that, for example, aliens quite different from ourselves could not find “grue” quite easy to understand without such a definition, and find “blue” rather confusing. But set that aside; there are two further worries about the proposed restriction on admissible vocabulary.

G. A generalization is confirmed by all of its instances.

x is grue if and only if either:
x is green, and has been observed before 4/7/14
x is blue, and has not been observed before 4/7/14

This suggests a plausible thought: perhaps we should restrict G so that it applies only to generalizations which involve **suitable scientific vocabulary** - perhaps we should restrict G by excluding predicates which cause the sorts of trouble caused by “grue” (“gruesome predicates”).

To pursue this thought, we need to be able to say what a gruesome predicate is - that is, we need to be able to say what, exactly, is so bad about “grue.” This turns out to be harder than you might think.

One might reply: “Yes, one **can** define “blue” this way - but we don’t **have** to. The difference between “grue” and “blue” is that no one could understand “grue” without this sort of time-indexed definition.” This suggests that we should exclude terms which are impossible to understand except via a time-indexed definition.

One might wonder why we should be so sure that, for example, aliens quite different from ourselves could not find “grue” quite easy to understand without such a definition, and find “blue” rather confusing. But set that aside; there are two further worries about the proposed restriction on admissible vocabulary.

The first is that this restriction is **not restrictive enough**: one can concoct gruesome predicates which are not defined in terms of times - for example, if all the emeralds which have been observed are from 17 emerald mines, we could define “grue” in terms of place. Or, if all the emeralds in the world have been seen by one person, we could define “grue” in terms of what has been observed by that person.

The second worry is that it is **too restrictive**: after all, we might be interested in investigating theories which are only about particular times, and places, and people - we don’t want our theory of confirmation to simply fail to apply to such theories.

The idea that we can save G by restricting it to a certain privileged class of vocabulary is thus - while initially promising - hard to carry out.

Let’s pursue a different idea, which involves a more sweeping rejection of G.

G. A generalization is confirmed by all of its instances.

x is grue if and only if either:
x is green, and has been observed before 4/7/14
x is blue, and has not been observed before 4/7/14

Let's pursue a different idea, which involves a more sweeping rejection of G.

This idea involves the claim that whether a piece of evidence counts in favor of a theory depends partly on our background beliefs about the subject matter in question.

Consider, for example, the following piece of evidence:

> Every lobster I have seen has been pink.

Now suppose that every lobster I have seen has been in a restaurant; and I know that lobsters in restaurants are pink because they are boiled. Given this knowledge it would, it seems, be absurd for me to take my observations of lobsters to confirm the generalization:

> Every lobster is pink.

Why? A natural thought goes something like this: I know that all the instances of this generalization I have observed have a certain property - being boiled in a restaurant - which explains why they are instances of the generalization. Moreover, I know that not all lobsters have this property - some are still in the wild. Whenever this is the case, the instances of a generalization fail to confirm it. That is:

> G*. A generalization that all the As are B is confirmed by an instance of an A being a B so long as: we do not believe that there is some property F such that all the A's which are instances of the generalization are F, and that had they not been F, they would not have been B.

This is not adequate as it stands - one might want to stipulate that the relevant beliefs about F be justified, it is not obvious who the "we" is who are supposed to do the believing, we may need to require that we believe that there are some A's which are not F, and it may be enough if had the instances not been F they **might** not have been B. Nonetheless, the above is enough to illustrate a response to the new riddle of induction which does not depend on sorting vocabulary into good and bad.

On the above sort of view, the instances of "All emeralds are grue" don't confirm this generalization because we know that all of the instances have a certain property - having been observed before - which are such that, had they not had this property, they would not have been grue.

G. A generalization is confirmed by all of its instances.

x is grue if and only if either:
x is green, and has been observed before 4/7/14
x is blue, and has not been observed before 4/7/14

G*. A generalization that all the As are B is confirmed by an instance of an A being a B so long as: we do not believe that there is some property F such that all the A's which are instances of the generalization are F, and that had they not been F, they would not have been B.

This is not adequate as it stands - one might want to stipulate that the relevant beliefs about F be justified, it is not obvious who the "we" is who are supposed to do the believing, we may need to require that we believe that there are some A's which are not F, and it may be enough if had the instances not been F they **might** not have been B. Nonetheless, the above is enough to illustrate a response to the new riddle of induction which does not depend on sorting vocabulary into good and bad.

On the above sort of view, the instances of "All emeralds are grue" don't confirm this generalization because we know that all of the instances have a certain property - having been observed before - which are such that, had they not had this property, they would not have been grue.

One interesting consequence of this sort of approach - something which Goodman also took the example of "grue" to illustrate - is that there can be no such thing as the "logic" of confirmation. If the above is right, we can never tell when some evidence confirms a theory **just by looking at the evidence and the theory** - in the way that we **can** look at a deductive argument and tell, just by looking at the premises and conclusion, whether it is valid.

Does this replacement of G help with the paradox of the ravens? Unfortunately, it appears that it does not. Think about the white chalk - this really does seem, if G* is true, to confirm the generalization that all non-black things are non-ravens. Think about the collection of chalk in this room. Is there any property had by all of those pieces of chalk which is such that, had they not had that property, they would not have been non-ravens? If we exclude "cheating" properties like the property of not being a raven, it seems not. (It might, of course, be hard to tell what the difference between a cheating and a non-cheating property is.)

G. A generalization is confirmed by all of its instances.

G*. A generalization that all the As are B is confirmed by an instance of an A being a B so long as: we do not believe that there is some property F such that all the A's which are instances of the generalization are F, and that had they not been F, they would not have been B.

This suggests that we need a bigger departure from G than that offered by G*. Perhaps it is not enough to modify our view of confirmation by adding constraints to do with our prior beliefs; perhaps we should give up on the idea of treating the relationship between a generalization and its instances as special in any way, and instead define confirmation entirely in terms of prior beliefs. This, in effect, is done by attempts to define confirmation using Bayes' theorem, which we will discuss in more depth in connection with the Doomsday Paradox.

Intuitively, what it says is that if we want to know the probability of some theory given a bit of evidence, what we need to know are three things: (1) the probability of the evidence given the theory (i.e., how likely the evidence is to happen if the theory is true), (2) the prior probability of the theory, and (3) the prior probability of the evidence.

So let's consider the ravens. The relevant question is then, prior to having the color of the chalk revealed: Is the probability of R - the theory that all ravens are black - given that the chalk is white greater than the probability of R by itself? It seems not; which would explain why, despite the plausibility of EQUIV and G, the chalk does not confirm R. (You might want to think about whether a Bayesian can preserve EQUIV.)

This means that we have to depart even further from the idea of a "logic of induction" or "logic of scientific discovery", if these are thought of as purely formal disciplines which test the relationship between evidence and theory. If the Bayesian is right, confirmation is relative to individuals, and can only be understood in terms of their beliefs.

One interesting consequence of this sort of view is that it does not really make sense to ask, without specifying a person or set of background beliefs, whether some evidence supports a theory - or even whether the theory is, in general, well-supported by the evidence. In general, it will be true that evidence can confirm a theory relative to person A but not relative to person B. Does this undercut the idea that the scientific method provides a method of belief formation which is rational for everyone? What would "the scientific method" mean for a Bayesian?