# The Unity of Robustness: Why Agreement Across Model Reports is Just as Valuable as Agreement Among Experiments

Corey Dethier
[Preprint; forthcoming in *Erkenntnis*]

**Abstract**

A number of philosophers of science have argued that there are important differences between robustness in modeling and experimental contexts, and—in particular—many of them have claimed that the former is non-confirmatory. In this paper, I argue for the opposite conclusion: robustness in modeling contexts is capable of providing confirmation, and the same analysis should be given in both contexts—that is, the degree to which robustness confirms depends on precisely the same factors in both situations. The positive argument turns on the fact that confirmation theory doesn't recognize a difference between different sources of evidence. Most of the paper is devoted to rebutting various objections designed to show that it should. I end by explaining why philosophers of science have (often) gone wrong on this point.

## 0 Introduction

Sometimes we believe a hypothesis because there are experiments that support it. Sometimes, we believe a hypothesis because our best model or models support it. A hypothesis might enjoy agreement—it might be 'robust'—in either context (or, indeed, across both). 'Robustness' in this sense is a *property*: it's something that a hypothesis *has* relative to a set of lines of evidence. In

this paper, I argue that (1) this property is confirmatory in the sense that the better the agreement between different lines of evidence, the more the hypothesis is confirmed by that evidence and (2) the difference between models and experiments has no general implications for the confirmation-theoretic evaluation of this property. I call this position 'unity':

> (U) Agreement across appropriately varied model reports confirms, and the degree of confirmation it provides depends on precisely the same factors as are operative in other contexts.

Unity is not a widely-defended position in the literature. To my knowledge, only Schupbach (2018) and, following him, Winsberg (2018) endorse it explicitly. By contrast, quite a few philosophers have argued against against unity. So, for example, Cartwright (1991) and Woodward (2006) argue that the property of robustness—that is, agreement—should be evaluated in different ways in experimental and modeling contexts. Similarly, a number of philosophers have argued that 'robustness analysis,' understood not as a property but as a strategy of testing hypothesis involving constructing varied models, 'does not' (Forber 2010, 37; Weisberg 2013, 167) and indeed 'is unable to' (Odenbaugh and Alexandrova 2011, 758) confirm hypotheses about the world (at best it confirms only claims about the models).[1] If these philosophers are right, however, that would imply that even where the method of robustness analysis yields multiple models across which a single hypothesis is robust, that hypothesis is not thereby confirmed.

The present paper defends unity against the arguments raised by its opponents: the *property* of robustness should be evaluated in the same way regardless of context. Confirmation of a hypothesis that is robust across models depends on precisely the same probabilistic features of the evidence as confirmation of a hypothesis that is robust across experiments. It's all a matter of filling in the same variables in the same formulas, and there's nothing about the difference between models and experiments that necessitates those variables taking on different values. The *strategy*, robustness analysis, is therefore

[1] I take it that sense in which strategies or methods 'provide confirmation-theoretic support,' (Forber 2010, 37), 'confirm' (Odenbaugh and Alexandrova 2011, 758), or 'bestow confirmation' (Weisberg 2013, 167) is that they deliver results that raise the probability of the hypothesis in question (though see §2). Note that while all of Forber, Weisberg, and Odenbaugh and Alexandrova explicitly understand 'robustness analysis' as limited to the modeling context, more recent discussions (e.g. Schupbach 2018; Winsberg 2018) have adopted a broader conception.

capable of providing confirmation insofar as it generates results that a hypothesis can be robust across, and while there are some reasons for thinking that the results found in experimental contexts will generally confirm more than those found in modeling ones, there's no ground for either the bright-line distinctions that Cartwright and Woodward argue for or the skepticism of Forber, Odenbaugh and Alexandrova, and Weisberg.

The positive argument for unity is simple. Model reports can serve as evidence. Confirmation theory doesn't recognize a difference between different sources of evidence: it tells us to calculate the degree of confirmation in the same way regardless. Since confirmation theory is not sensitive to the difference between modeling and experimental contexts, therefore, we have good reason to think that unity is true (§1). After laying out this argument, I consider four different objections in sections 2-5. Most of these objections can be understood as designed to show that confirmation theory *should* recognize this difference, because there is some feature of model reports—the nature of the independence relationships between them, their non-empirical character, their reliance on idealizations—that affects the relevant probabilities. But none of them are persuasive. Finally, I end the paper with a discussion of the relationship between robustness the property and the strategy of 'robustness analysis.' I argue that the focus on practical difficulties with the latter has led to confusion about unity and the ability of robustness / robustness analysis to confirm (§6).

One final note before I begin. The thesis of this paper is a general thesis about the relationship between robustness or agreement across multiple lines of evidence and confirmation. The approach that I'll take is equally general. Of course, actual cases of robustness are extremely varied in both the degree of variation between the different lines of evidence and how the varied tests fits into the evidence as a whole. To determine what we should take from robustness in any one case, examination of the details will be necessary. As case-studies like Parker (2018) and Winsberg (2018) illustrate, however, the evaluation of those details has to occur within some sort framework for evaluating how robustness works in general. My contention is simply that the fact that the hypothesis is robust across models rather than experiments should not enter into our general framework: we should analyze both kinds of cases in the same way.

# 1 Unity and confirmation theory

Consider a simple proposition that we'll call $H$: after falling for one second, an object will be traveling at $\sim$10 meters per second. One way to confirm $H$ is to drop an object and observe its velocity after exactly one second. To confirm $H$ in this manner, we we would need to employ an instrument such as a radar gun. Let $I$ indicate the proposition that the radar gun reads $\sim$10 meters per second, or in other words, the proposition that the radar gun represents the object as traveling at $\sim$10 meters per second. Under the right conditions—i.e., if we have sufficient reason to believe that the radar gun is well-designed and operated—$I$ confirms $H$ in the incremental sense that $p(H|I) > p(H)$.

Alternatively, we could confirm $H$ by using a model of a freely falling object. In this case, rather than an instrumental reading, we would conditionalize on what I'll call a 'model report'—essentially, on the fact that the model represents the target as being a particular way. In this case, we can think of the model report, which we'll denote $R$, as being the proposition that the analogue of the object in the model is traveling at $\sim$10, or, in other words, the proposition that the model represents the object as traveling at $\sim$10 meters per second. Under the right conditions—i.e., if we have sufficient reason to believe that the gravitational model is well-designed and operated—$R$ confirms $H$ in the incremental sense that $p(H|R) > p(H)$.[2]

I take it that the foregoing sketch provides us with some reason to accept the following claim:

(P1) Model reports are evidence.[3]

where it should be understood that here we mean that such reports provide evidence not just about the nature of the model, but also about the nature

[2]Note that 'well-designed and operated' is doing the same work in the experimental and modeling cases: incremental confirmation fails in cases where (e.g.) we know that the radar gun is malfunctioning or that the model-user cannot reliably carry out calculations. We need not assume that model reports *always* incrementally confirm because instrumental readings do not either. As has been stressed by (e.g.) Cartwright (1999), Guala (2002), and Steel (2008) taking either experiments or models to tell us about real-world situations requires some kind of extrapolation.

[3]I mean 'evidence' here in the sense that that they are the kind of things that figure into the confirmation of a hypothesis. I am not presuming that model reports always raise the probability of a hypothesis.

of the target that the model represents. The reasoning here is simple: instrumental readings are a paradigm case of evidence for hypotheses about the world, and the sketch I've just given indicates that model reports can serve precisely the same confirmatory function as instrumental readings. Notably, I take it that (P1) is uncontroversial: throughout the sciences, we often take the fact that some hypothesis is true 'in the model' as a reason to increase our confidence in a given hypothesis (see Parker 2020a). So, for example, we consult a model of the solar system to determine the exact date of historical eclipses or run simulations using climate models to generate 'projections' of the climate under different forcing scenarios. That the eclipse happened on a certain date in the model gives us reason to believe that it actually happened on that date; that the average temperature increases by 3°C under a given forcing scenario in the model gives us reason to believe that if that scenario comes to pass, the average temperature would increase by 3°C. These model reports are *like* instrumental readings in that they may require interpretation and we might expect them to be more or less precise, accurate, or reliable; just like instruments, models can be broken, misapplied, or misread. Granting these complications, however, my claim here is a minimal one: in some cases, learning what a model reports raises the probability of propositions that we're interested in testing.

(P1) is interesting because there is a theorem of the probability calculus that tells us how the joint likelihood of *any* two pieces of evidence is related to their individual likelihoods:

$$\frac{p(E_1\&E_2|H)}{p(E_1\&E_2|\bar{H})} = \frac{S(E_1,E_2|H)}{S(E_1,E_2|\bar{H})} \times \frac{p(E_1|H)}{p(E_1|\bar{H})} \times \frac{p(E_2|H)}{p(E_2|\bar{H})} \tag{JL}$$

where $\bar{H}$ is the negation of $H$ and, following Myrvold (1996), $S$ is defined as

$$S(\phi,\psi|\chi) = \frac{p(\psi|\phi,\chi)}{p(\psi|\chi)} = \frac{p(\phi\&\psi|\chi)}{p(\phi|\chi)p(\psi|\chi)}$$

The idea is that $S$ measures the 'similarity' of the propositions $E_1$ and $E_2$ in the sense that where $S > 1$ the two propositions are positively correlated, where $S < 1$ they're negatively correlated, and where $S = 1$ they're probabilistically independent. For any two pieces of evidence, therefore, the degree to which they jointly confirm a hypothesis is a function of their individual likelihoods and their 'conditional' similarity. Since (JL) gives the joint likelihood, this fact is true for any incremental confirmation measure. As

Myrvold glosses the point: 'a diverse body of evidence confirms a hypothesis more strongly if the hypothesis renders the evidence less diverse' (Myrvold 1996, 663), though we can add that in fact the relationship is really one of *to the degree that* rather than merely *if* (see Wheeler and Scheines 2013).

Robustness is a special case: a hypothesis is robust across two pieces of evidence when they 'agree' that it is true. In confirmation theory, this condition is plausibly interpreted in terms of each of the pieces of evidence increasing the probability of the hypothesis when considered individually (i.e., as each $E_i$ being such that $p(H|E_i) > P(H)$).[4] Then the two pieces of evidence jointly raise the probability of the hypothesis just in case the ratio between $S(E_1, E_2|H)$ and $S(E_1, E_2|\bar{H})$ is not so negative as to offset the effect of the individual pieces of evidence, and further the larger this ratio the more evidence confirms. We can even use the ratio between $S$ terms as a measure of the value added by the variation between the different pieces of evidence. Nothing about this reasoning depends on where the evidence comes from—it doesn't matter whether we replace the $E$ variables in (JL) with $R$ variables or $I$ variables or indeed combinations of the two. The upshot is powerful motivation for the principle I'll call (P2):

> (P2) If model reports are evidence, then agreement across appropriately varied model reports confirms, and the degree of confirmation it provides depends on precisely the same factors as are operative in other contexts.[5]

Together, (P1) and (P2) provide an extremely simple argument for unity:

> (P1) Model reports are evidence.
>
> (P2) If model reports are evidence, then agreement across appropriately varied model reports confirms, and the degree

[4] I should note that while this interpretation is natural, it isn't universal. Parker (2018, 290–91, fn 2), for instance, offers a more pre-theoretic interpretation. These interpretations will come apart only where an individual result supports a hypothesis in some pre-theoretic sense while failing to confirm it—e.g., where what appears to be a result in favor of a hypothesis relies on the tacit assumption of that hypothesis. *Perhaps* such cases are more common in modeling contexts, but there's no reason that they couldn't arise in experimental ones as well, and so this is not a reason to reject unity.

[5] 'Appropriately varied' here should be understood simply as shorthand for the claim that the ratio between $S(E_1, E_2|H)$ and $S(E_1, E_2|\bar{H})$ is not so negative as to offset the effect of the individual pieces of evidence.

of confirmation it provides depends on precisely the same factors as are operative in other contexts.

$\therefore$ (U) Agreement across appropriately varied model reports confirms, and the degree of confirmation it provides depends on precisely the same factors as are operative in other contexts.

Essentially, whenever we have robust evidence from multiple sources, confirmation theory tells us to take account of only the likelihoods and probabilities, not whether the evidence comes from a model or an instrument. Or, in other words, if we hold fixed the relevant probabilities, there is no effect of varying whether the terms in the formalism represent instrumental readings or model reports.

Given the simplicity of this argument, anyone wanting to contest unity must either reject (P1) or (P2). Since (P2) is motivated by a theorem of the probability calculus, an argument against (P2) must consist of a demonstration that in fact some feature of modeling contexts makes them a special case that should be analyzed differently than the fully general case represented by (JL). As we'll see, a number of arguments to this effect have been advanced over the last few decades. In the next few sections, I'll argue that none of them are successful.

Before that, however, two comments. First, I want to be clear that I don't take myself to have established that (JL) provides *the* formal analysis of robustness. I think that the analysis that I've given—which is based in the work of Myrvold (1996) and Wheeler and Scheines (2013)—represents the most general and simple picture of variation in evidence. Other accounts—such as those put forward by Bovens and Hartmann (2003) and Schupbach (2018)—require stronger assumptions that I think are not always plausible. There's room for disagreement on this point, however. What's important is that *none* of these accounts of robustness or variation in evidence naturally recognize a difference between modeling and experimental contexts. Absent some strong argument for introducing one, therefore, it seems like there's good reason to think that the difference between the two cases is not relevant to how the contribution to confirmation should be analyzed. And, thus, if I can show that none of the extant arguments to this effect are persuasive, we will have good reason to accept unity.

Second, it might be thought that this argument shows too much. So, for example, a similar argument might be taken to show that robustness operates the same way in testimonial contexts as in experimental ones. I don't see a

problem with this result: structural relations like robustness *should* have the same implications in all domains. Note, however, that the real meat of my argument lies in the discussion of the proposed differences between models and experiments that follow. To show robustness works in similar ways with respect to testimony and experiments, we would need to carry out the same kind of evaluation of the arguments for building differences between testimony and experiments into the framework.

# 2 Arguments against (P1)

A large number of philosophers—Forber (2010) and Orzack and Sober (1993) perhaps most explicitly—have argued that robustness across model reports fails to confirm because it doesn't provide empirical evidence, just clarifies the nature of the models. Even some defenders of robustness—such as Kuorikoski, Lehtinen, and Marchionni (2010, 2012) and Weisberg (2013)—seem to accept this argument and instead argue for some facilitating but ultimately non-confirmatory role for robustness across model reports. Robustness across different experimental setups, by contrast, is supposed to provide empirical evidence and thus confirmation. Hence unity fails. Call this the argument from non-empiricality.

In the present context, the argument from non-empiricality can be read as rejecting either of the premises given in §1. If we read it as rejecting (P1), then the argument is simply that model reports do not confirm and thus are not evidence: models are theoretical constructs, not empirical information, and only the latter confirms. This position strikes me as unacceptable for a number of reasons. Most prominently, it would render a large proportion of our best sciences unconfirmed. Not only are there a wide variety of results that depend explicitly on models, but—as emphasized by recent work on measurement, such as Morrison (2015) and Tal (2012)—models often play crucial roles in even the most paradigmatic cases of empirical evidence. And it's worth noting as well how difficult it is to draw sharp ontological distinctions between models on the one hand and experiments on the other (Mäki 2005; Parker 2009; Winsberg 2010, chapter 4): after all, many models provide results by means of computer simulations, in which case they can be seen as experiments on the behavior of electrical signals through various metals.[6]

[6]This is an independent reason to suspect that unity is true, but I won't belabor the point here.

Perhaps I'm being too quick here, however. One way of pushing the idea that models don't themselves confirm is by seeing them as encoding prior knowledge (compare Beisbart 2012, 2018). A model, on this view, can't tell us anything *new* about the world. Rather than acting like empirical evidence, therefore, model reports are more like propositions deduced from what we already know. Suppose that's right. If it is, then the support offered by model reports is easily assimilated to the problem of old evidence: basically, we're learning that the prior empirical research that went into building the model supports the hypothesis in a way that we didn't previously recognize.[7] I'm not particularly concerned with whether or not it's the model report itself that is 'responsible' for the confirmation or if there's some important sense in which it's 'really' previously collected empirical evidence that does the confirming by way of some mechanism of accounting for old evidence. In either case, our confidence in the hypothesis can go up when we learn the model report. And that's all that's needed for the argument of the last section. That is, even if model reports aren't 'really' evidence thus can't 'really' confirm, they're capable of providing new information and thus of leading us to raise our probability in the hypothesis when we learn them, and that's enough.

One more way of objecting to (P1) is worth considering. Models—presumably unlike experiments, though I think this presumption is false—are often heavily idealized. One might think that the reports of heavily idealized models cannot confirm. Odenbaugh and Alexandrova suggest something along these lines when they say that 'there are assumptions we know to be false or whose truth we cannot evaluate. ... unless we can 'de-idealize' our Galilean assumptions ... we do not know that we have adequately represented a causal relationship' (Odenbaugh and Alexandrova 2011, 763; see also Odenbaugh 2011). The idea: unless we can show that the idealizations present in a model can be removed without changing its implications, we don't have good reason for increasing our confidence in a hypothesis based on the reports given by the model. Since we're rarely (if ever) in a position to remove all of the idealizations in a model, (P1) fails.

Whether or not this argument is what Odenbaugh and Alexandrova intend (more on that later), its conclusion is mistaken for the same reasons discussed above: rejecting (P1) in this manner is simply not plausible because it would leave far too much of our most successful sciences uncon-

[7]For an extended defense of this kind of thought, see Beisbart (2012, 2018) and Parker (2020a).

firmed. It's even commonly thought that all models are idealized (see, e.g., Teller 2001), which would render virtually all scientific results unconfirmed. Furthermore, the de-idealization method for showing that idealizations are harmless is clearly not the only one. For instance, we can show that an idealized model delivers consistently accurate predictions within a given domain. In such cases, even if we cannot build a model of the phenomenon with no idealizations, the past success of the model reports provides us with at least some reason to believe whatever is indicated by the model's next report. There's no persuasive reason to take the mere presence of idealizations in models as a reason to reject (P1).

Frankly, I take (P1) to be unassailable. A much more plausible route to rejecting the argument of the last section involves rejecting (P2) by arguing that agreement across model reports is special in a way that the argument I've given doesn't account for. I turn to objections along these lines now.

# 3 The argument from independence

The oldest argument against unity goes back to Nancy Cartwright, who urges that *unlike* what is the case when experiments or measurements agree, different models 'do not constitute independent instruments doing different things, but rather different ways of doing the same thing: instead of being unrelated, they are often *alternatives* to one another, sometimes even contradictory' (Cartwright 1991, 153). There is some important sense in which at most one of a set of models of the same phenomenon can be correct, and so variation among models must be given a very different analysis from variation among instruments, where the correctness of one instrumental reading does not preclude the correctness of another. Variations on this argument have been expressed as well by Woodward (2006), and similar ideas arguably lie behind at least some of the objections of critics like Houkes and Vaesen (2012), Odenbaugh and Alexandrova (2011), and Orzack and Sober (1993).

In the present context, we can see the argument just presented as undermining (P2). The critic of unity can grant the claim that the two types of robustness can be represented within the probability calculus in the same way. Their contention is that while in both cases we'll want to appeal to similarity measures—represented by $S(E_1, E_2)$ given either $H$ or $\bar{H}$—these similarity measures will behave in radically different ways in experimental and modeling contexts. In the experimental context, this measure represents

how well the hypothesis unifies the different pieces of evidence. In the modeling context, by contrast, any two models will be mutually exclusive, meaning that they can't be unified (at least not by the true hypothesis). As such, in any modeling context, we should expect the ratio between $S(E_1, E_2|H)$ and $S(E_1, E_2|\bar{H})$ to be small, arguably smaller than 1. So even if the *form* of (JL) is maintained on this picture, there's a fact about the nature of models that makes robustness across modeling a very special case—importantly, a special case that is much less interesting or powerful than robustness in general. Call this the argument from independence.

So understood, the argument fails for a straightforward reason. Recall that the $S$ terms in our model track the degree of correlation between *reports*, not the degree of independence between model assumptions. So the argument from independence simply misses the mark: even if we grant that there is a difference between models and instruments regarding when they're likely to be mutually exclusive, that difference is irrelevant to the behavior of robustness in the two contexts.[8]

The rejoinder just given rests on two observations about the relationship between model reports, confirmation, and model assumptions. First, the representational accuracy of a model report is not generally or typically dependent on the truth of the model assumptions. It's unproblematic that idealized models can in some cases deliver reports that accurately represent the target system. An eclipse can occur at the right time and location in the model even if the model misrepresents many other parts of the system. The straightforward implication is that two models whose assumptions are mutually inconsistent will not necessarily deliver mutually inconsistent reports. On the contrary, there are well-known cases where the opposite is the case: Newtonian physics and general relativity rest on mutually inconsistent assumptions, but there's a wide class of phenomena for which either will make reports that are accurate up to extremely high levels of precision.

The second observation is that in general, when we're evaluating what we should think given a particular report, the question to ask is how likely it is that said report accurately represents the target—*not* how likely that the model that produced the report has true assumptions. As Dethier (2019) has argued, the latter question is largely irrelevant, like asking how likely it is that thermometer outside my window would deliver an accurate report under

[8]Kuorikoski, Lehtinen, and Marchionni (2010) have advanced a similar argument based on distinguishing between different parts of the model.

conditions radically different from those currently under consideration.[9] This is particularly clear when we represent the effects of a model report within a Bayesian framework:

$$p(H|R) = \frac{p(H)p(R|H)}{p(H)p(R|H) + p(\bar{H})p(R|\bar{H})}$$

The likelihood of the report factors into the conclusion that we should draw from it, while the probability that the model has true assumptions doesn't appear in the formulation at all.

Given these two observations, it should be unsurprising that our account of agreement across model reports turns on relationships between different model reports rather than model assumptions. But this fact blocks the argument from independence: we have no reason to think that the relevant relationships between model reports systematically differ in any important way from the evidence found in experimental settings.

# 4 The argument from non-empiricality

Above, we encountered an argument to the effect that (P1) is false because model reports don't provide empirical evidence. (P2) can be challenged along the same lines; indeed, this may be a better reading of what the critics have in mind. On this reading, the critic grants that models can confirm by incorporating the treatment of models into the problem of old evidence: what the model report demonstrates is that there's a (previously unappreciated) connection between any empirical evidence that gives us a reason to think that the model is likely to be accurate and the hypothesis, and learning this fact provides a kind of confirmation (see Parker 2020a). This critic could even grant the point, recently stressed by Lehtinen (2016, 2018) and Lloyd (2015), that under some circumstances, robustness across model reports will confirm for this same reason: where learning the report of an additional model has the effect of bringing previously irrelevant (but 'old') empirical evidence to bear on the hypothesis, robustness across model reports will confirm. What the critic contends is that whereas robustness across experiments always brings new empirical evidence to bear, robustness across model reports does so only sometimes (at best), and so the two shouldn't be analyzed in the same

[9] See also Currie (2017) and Parker (2020b).

framework—at least as I read them, this is essentially the argument given Forber (2010) and Woodward (2006), who are concerned that agreement across model reports is radically different from paradigm cases of robustness across experiments.

I'm sympathetic to this line insofar as the point is that *usually* we have better overall evidence in experimental cases of robustness than in modeling cases (see §6). What I want to contest is that this phenomenon—if it actually exists—has implications for the analysis of the property of robustness. That is: understood as a critique of the position I've termed unity, I think it's mistaken, for the reason that it requires us to run together what seem to be two different phenomena.[10] The basic intuition behind robustness is that running a second, varied, experiment (or model) provides *better* evidence than simply repeating the same experiment again. Repeating an experiment is valuable: it generates more data and, as a consequence, lowers the probability of random sampling error. Intuitively, varying the experimental setup is more valuable than repeating the same experiment because it doesn't just provide more data and thus decrease the risk of sampling error, it also decreases the probability of systematic error due to instrumental bias. Change the instrument employed and any defects in the original instrument can no longer affect the results. As the data sets involved get larger, the probability of random error asymptotically approaches zero, and thus the amount of support offered by each additional run with the same experimental setup also asymptotically approaches zero. By contrast, the probability of error due to instrumental bias remains constant so long as the experimental setup is unchanged. In the infinite data limit, an additional data point produced by the same experiment provides no confirmation, while an additional data point produced by a different experiment provides some.

The point is the following: not only can we conceptually distinguish between the effects of robustness and the effects of gathering more data, we can—at least in principle—imagine experiments that lack the 'more data' effect that is supposed to distinguish robustness in experimental contexts from robustness across model reports. The natural conclusion is that experimental cases of robustness involve two separate phenomena: the pure more data phenomenon and the robustness phenomenon—the latter of which occurs not just in the experimental context but also in modeling contexts as well.

[10]For an extended discussion along the same lines offered in the rest of this paragraph, see Staley (2018).

If that's right, then we should treat robustness the same in both contexts; even granting the full strength of the critic's argument that there's something importantly non-empirical about robustness across model reports, the most natural conclusion is not that robustness in modeling and experimental contexts should be treated differently, but that there are additional empirical factors in experimental contexts above the 'non-empirical' (according to the critic) probability-raising effect of robustness. There's no reason to think that the alleged non-empirical character of robustness provides us with any reason to distinguish between robustness in modeling and experimental contexts.

# 5 The argument from idealization

Just as the non-empiricality argument can be understood as attacking either (P1) or (P2), so too can the argument from idealization. The second reading of the argument is suggested by Houkes and Vaesen (2012) and some passages in Odenbaugh and Alexandrova (2011). So, for instance, Odenbaugh and Alexandrova argue that in real life, groups of models will always share some set of idealizations, and this means that robustness across models has 'confirmatory value' only when we can remove all of the idealizations (Odenbaugh and Alexandrova 2011, 764).

I think that the most charitable way of reading this argument is as follows. Odenbaugh and Alexandrova in particular are concerned with a case in which (a) the models that generate the reports are all known to share idealizations and (b) we don't have any other information about the models, such as information indicating that the models, though idealized, are highly reliable with respect to similar hypotheses. In this case, they claim, robustness does not confirm. If we understand 'confirm' here in terms of providing the hypothesis with a high probability (what's sometimes called 'confirmation as firmness'; see, e.g., Fitelson 2017), then this claim is true: in the situation described, it's at least arguable that robustness across the different model reports does not provide us with sufficiently high confidence for (e.g.) knowledge. Houkes and Vaesen can be read similarly: when there are idealizations present across what they term the 'model family,' agreement across model reports can only raise our confidence up to the level of our confidence that one of the members of the family is accurate. Presumably, this feature of robustness in modeling contexts is in contrast to robustness in experimental contrasts; Odenbaugh and Alexandrova (2011), at least, explicitly position themselves as against

any analogy between the two.

I'll register two responses to this objection. The first is that in the scenario described, robustness can still offer confirmation in the sense of increasing probability; that is, adding an additional model report that supports the hypothesis can increase our confidence in the hypothesis, even if there remain serious idealizations that should prevent us from accepting or believing the hypothesis. Whether or not robustness confirms in the probability-raising sense is simply determined by (JL), and so there are clear sufficient conditions on robustness providing 'confirmatory value': for instance, if the model reports are sufficiently diverse in the sense that the ratio between $S$ terms is at least 1, then multiple reports, each of which supports the hypothesis when considered individually, will jointly increase the support for the hypothesis. Indeed, under the same conditions, adding an additional report will always serve to increase the degree of confirmation, as can be seen clearly in following trivial consequence of (JL):

$$\frac{p(E_2|H,E_1)}{p(E_2|\bar{H},E_1)} = \frac{S(E_1,E_2|H)}{S(E_1,E_2|\bar{H})} \times \frac{p(E_2|H)}{p(E_2|\bar{H})}$$

Since our stipulation of sufficient variation and robustness respectively guarantee that the two terms on the right-hand side are greater than 1, the left-hand side has to be greater than 1 as well. The implication is that robustness across model reports can provide confirmation in the sense of probability raising regardless of whether it provides confirmation in the sense of providing us with sufficiently high confidence for something like knowledge. In this respect, however, robustness across model reports is no different from any other empirical evidence, let alone from robustness in experimental contexts.

The other rejoinder is that there's nothing particular to *models* in the objection given above; indeed, insofar as it works, it works equally well in experimental contexts. Consider the case in which we vary our instruments or assumptions across a series of experiments but where a central instrument or assumption—and one whose reliability is questionable in the present context—is shared across each of the different instances. Precisely the same worries apply to this case as Odenbaugh and Alexandrova raise with respect to modeling: since the instrument or assumption in question could be leading us astray, we don't have knowledge until we show that it isn't. Furthermore, these different experiments should not raise our confidence in the hypothesis above our confidence in the 'experiment family' where this is understood in

the same way as 'model family' in Houkes and Vaesen (2012).[11]

When talking about robustness in experimental contexts, we tend to gloss over the marginally varied cases of variation in experiments such as the everyday use of multiple thermometers or observers and focus instead on the dramatic cases like Perrin's measurements of Avogadro's number. By contrast, the literature on robustness in modeling contexts has tended to focus on cases in which the models vary relatively little and share a large number of assumptions. If we're going to compare the value of robustness in the two contexts, however, the comparison should be with all other things being equal. The argument that I gave in the beginning of this essay indicates that when the *ceteris paribus* comparison is carried out, the two cases are identical: under the right circumstances, which are the same in both contexts, robustness increases probability and thus confirms.[12] There's no good reason to reject unity.

# 6 Robustness and robustness analysis

So far in this essay, I've gone to bat for the thesis I termed 'unity': robustness confirms in both experimental and modeling contexts, and the degree of confirmation that it provides depend on precisely the same factors in both contexts. The confirmatory value of robustness, while variable, doesn't depend on whether what's varied over are aspects of an experimental setup or modeling assumptions. I've offered an argument for this position and considered a number of possible objections against it, none of which are persuasive.

So what accounts for the widespread rejection of unity among philosophers of science? As indicated above, I think that part of the explanation is that there's a history of contrasting the best and most famous cases of ro-

[11]Well, not quite: that depends on the distribution of the prior. But the same point applies in the modeling case: we can imagine, for instance, that our prior is such that we're extremely confident that (a) all of the instruments/models are working, (b) the hypothesis is likely false, and (c) if one of instruments/models isn't working, then the hypothesis is true. Then learning that the model reports support the hypothesis should plausibly lead us to have higher confidence in the truth of the hypothesis than in the reliability of the 'family.'

[12]Of course, it's open to my opponent to argue that my way of cashing out the *ceteris paribus* condition in this case begs the question. What they owe us, then, is an alternative conception of what it means for 'all other things to be equal' in this case that vindicates the view that there's some important difference between the two cases.

bustness in experimental contexts—particularly Perrin's work on Avogadro's number—with the worst and least compelling cases of robustness in modeling contexts. Since the confirmatory value of robustness varies, this kind of comparison is guaranteed to mislead.

I think that there's another, perhaps deeper, explanation. Much of the literature on robustness in modeling contexts has been focused on 'robustness analysis,' which can be thought of as a *strategy* for testing a hypothesis: construct a number of different models or conduct a number of different experiments, and show that each of them supports this hypothesis.[13] My view is that philosophers have tended to focus on the difficulties with *applying* this strategy in modeling contexts, which are arguably more substantial than those associated with applying the strategy in experimental contexts. In making this point, however, they've expressed this fact by saying that in modeling contexts, 'Robustness analysis does not itself bestow confirmation' (Weisberg 2013, 167) or by claiming that robustness analysis is 'unable to confirm' (Odenbaugh and Alexandrova 2011, 758). Read in a straightforward way, these comments are misleading. It isn't the case that robustness analysis cannot confirm, because if the strategy reveals agreement across appropriately-related pieces of evidence, then it confirms. This is true regardless of whether the context is one of experiment or modeling. In other words, unity is perfectly compatible with the claim that robustness analysis *tends* to be less effective in modeling contexts; if I'm right, however, arguments for the latter claim have obscured the truth of the former.

Why might it be more difficult to apply robustness analysis in a modelling contexts? Two major reasons come to mind. The first concerns an important difference between experiments, modeling, and our knowledge of a target. The second concerns our ability to evaluate evidence. Beginning with the first. In general, the more we know about a phenomenon, the more stringent our requirements on models of that phenomenon and thus the harder it is to build distinct models of it. The solar system provides a nice exemplar: it's much easier to build a new model of the solar system if the criterion of success is replicating broad patterns in the movements of the planets than it is if the criterion includes accounting for the precession in Mercury's perihelion down to mere seconds of arc. The difficulty increases when we also want

[13]It's worth reiterating that much of the literature on 'robustness analysis'—virtually everyone prior to Schupbach (2018) explicitly limits their discussion to the modeling context.

the models to be varied or dissimilar. We can always create a new model by adding undetectable teacups to an old one, but I take it that robustness across models with varying numbers of undetectable teacups isn't very useful or interesting.[14]

By contrast, the ease of developing separate means of instrumental or experimental access increases with our knowledge of a phenomenon. Take Perrin's work on Avogadro's number. Crucial to Perrin's diverse means of measuring Avogadro's number was prior knowledge about its relationship to quantities like the mean kinetic energy of the system or the diameter of the component molecules. For example, as Perrin (1910, 51) explicitly notes, Einstein's work connecting the energy of a 'granule' suspended within a volume and the mean energy of the molecules within that volume allowed Perrin to use the displacement of a granule via Brownian motion as a proxy for the mean kinetic energy of the system and thus for Avogadro's number. The measurement of the quantity would not have been possible without the theory.[15] Similar comments apply to the measurement of the mass and charge of an electron. If all we know about electrons is that they make up cathode rays, there are only so many independent ways that we can measure quantities like charge or mass. The more knowledge we gain about electrons—the more testable phenomena in which they play an identifiable role—the more diverse our means of measurement can be.

These are generalities, not hard and fast laws. But the first generality makes it likely that most cases in which it pays to build multiple models of a phenomenon and test the robustness of a result across them will be cases in which we have relatively poor knowledge of the nature of said phenomenon—and thus makes it relatively likely that the amount of confirmation offered by a single model report will be low. By contrast, it's only when we have relatively substantial knowledge of a phenomenon that we're able to develop multiple lines of empirical access to it. The second generality thus makes it likely that the confirmation offered by each experiment is relatively high. So, the *strategy* of robustness analysis is likely to be more effective in experimental contexts than in modeling ones, because the overall quality of the evidence produced by the strategy is likely to be greater in the former. I stress again, however: that it is likely to be harder to show that a hypothesis

[14]The point made here is a variant of one familiar from debates about underdetermination, namely that it's quite hard to generate empirically adequate competitors to successful theories without relying on tricks (see Laudan and Leplin 1991).

[15]For an extended analysis of the implications of this point, see Smith and Seth (2020).

is true using agreement across varied models than it is to do the same with agreement across varied experiments does not mean that we should analyze actual cases of agreement differently in the two contexts. And, further, even if it is true that the strategy of robustness analysis is likely to be less effective in modeling contexts, that doesn't vindicate the claims that it does not or cannot confirm.

Turning now to the second reason. In order to use any sort of scientific evidence as part of an argument for accepting a hypothesis, we need to be able to evaluate what that evidence actually supports and to what degree. It might be the case that in experimental contexts, we can generally make these judgments with some confidence (I'm skeptical); it certainly isn't always true that we can make accurate evaluations of the quality of evidence in cases of robustness across model reports. As Parker (2018) emphasizes, for instance, we're just not in a position to evaluate the value of robustness across ensembles of climate models for hypotheses about the future of our climate: the models are too complex, their parts too interconnected, and their parameters too calibrated. It's one thing to say that robustness provides us with some evidence in these sorts of cases; it's another thing entirely to actually evaluate how good the evidence is or to derive actionable conclusions from it. As Odenbaugh and Alexandrova (2011) stress, for example, it's hard to know what we should conclude when two heavily idealized models lead to the same result. Perhaps the hypothesis is true; perhaps the shared idealizations are seriously problematic.

The fact that it's often very hard to judge or even estimate what's gained from an additional confirming model report clouds our judgment about cases of robustness across model reports. Cases of agreement across experiments are arguably easier to evaluate, at least insofar as we can usually say that this or that instrument is behaving as desired and thus rule out one possible source of error. In modeling contexts, by contrast, the presence of ineliminable idealizations and complex model structures make these sorts of simple judgments much more difficult and much less reliable. In the climate case, for instance, it's simply not clear what has been gained by replacing one idealized representation of cloud cover dynamics (say) by another when we know that both are flawed in various ways. It's plausible that there's some degree of confirmation when there's agreement in cases like these, but evaluating how much this agreement moves the needle is virtually impossible.

As I've indicated, the difficulties discussed in this section are difficulties for robustness *analysis*—for the practical project of using the evidence pro-

vided by robustness to argue for this or that hypothesis. They don't have have any bearing on unity, however, since unity is a claim about the evidence provided by the property of robustness itself, not our ability to find and evaluate it. Indeed, the arguments I've just sketched for the existence of problems with robustness analysis in modeling contexts rely on understanding robustness as fundamentally of the same kind in both contexts; it's precisely because the two contexts differ in ways that are relevant to applying the shared evaluative framework that makes robustness analysis more difficult to effectively employ in one than in the other. I suspect that the attention that has rightly been drawn to the difficulties inherent in evaluating robustness across model reports clouds our judgments about unity; attention to these difficulties makes it seem as though robustness across model reports is different in kind from robustness across experiments, when in reality what's going on is merely that the evidence is generally harder to find and interpret in the former context.

# 7 Conclusion

In this article, I've argued for unity: robustness across appropriately varied model reports confirms, and the degree of confirmation it provides depends on precisely the same factors as are operative in other contexts. I began by giving a straightforward argument to this effect, namely that if model reports can serve as evidence—which they can—then the probability calculus entails that the evidence that they provide has the same confirmation-theoretic effects as empirical evidence. I then considered a number of arguments against unity, all of which I've contended are unpersuasive. Finally, I argued that while it's plausible that the strategy of robustness analysis will be less effective in modeling contexts, this shouldn't be taken as evidence against unity: at best, these arguments show that there may be differences in the degree of confirmation provided that tend to correlate with the differences between models and experiments. But unity is a claim about whether robustness across models should be given the same confirmation-theoretic analysis as robustness across experiments; the way to show that it is false is to show that the alleged tendencies should be accounted for via applying a different (formal) framework in the two cases. And there's simply no good argument for that conclusion.

# References


Beisbart, Claus (2012). How can Computer Simulations Produce new Knowledge? *European Journal for the Philosophy of Science* 2: 395–434.

Beisbart, Claus (2018). Are Computer Simulations Experiments? And if not, how are they Related to Each Other? *European Journal for the Philosophy of Science* 8: 171–204.

Bovens, Luc and Stephan Hartmann (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Cartwright, Nancy (1991). Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy* 23: 143–55.

— (1999). *The Dappled World*. Cambridge: Cambridge University Press.

Currie, Adrian (2017). From Models-as-Fictions to Models-as-Tools. *Ergo* 4: 759–81.

Dethier, Corey (2019). How to Do Things with Theory: The Instrumental Role of Auxiliary Hypotheses in Testing. *Erkenntnis* (online first).

Fitelson, Branden (2017). Confirmation, Causation, and Simpson's Paradox. *Episteme* 14: 297–309.

Forber, Partick (2010). Confirmation and Explaining How Possible. *Studies in History and Philosophy of Science Part C* 41: 32–40.

Guala, Francesco (2002). Models, Simulations, and Experiments. In: *Model-Based Reasoning: Science, Technology, Values*. Ed. by Lorenzo Magnani and Nancy J. Nersessian. Dordrecht: Kluwer: 59–74.

Houkes, Wybo and Krist Vaesen (2012). Robust! Handle with Care. *Philosophy of Science* 79: 345–64.

Kuorikoski, Jaako, Aki Lehtinen, and Caterina Marchionni (2010). Economic Modelling as Robustness Analysis. *The British Journal for the Philosophy of Science* 61: 541–67.

— (2012). Robustness Analysis Disclaimer: Please Read the Manual Before Use! *Biology & Philosophy* 27: 891–902.

Laudan, Larry and Jarrett Leplin (1991). Empirical Equivalence and Underdetermination. *The Journal of Philosophy* 88: 449–72.

Lehtinen, Aki (2016). Allocating Confirmation with Derivational Robustness. *Philosophical Studies* 173: 2487–509.

— (2018). Derivational Robustness and Indirect Confirmation. *Erkenntnis* 83: 539–76.

Lloyd, Elisabeth (2015). *Model Robustness* as a Confirmatory Virtue: The Case of Climate Science. *Studies in History and Philosophy of Science Part A* 49: 58–68.

Mäki, Uskali (2005). Models are Experiments, Experiments are Models. *Journal of Economic Methodology* 12: 303–15.

Morrison, Margaret (2015). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford: Oxford University Press.

Myrvold, Wayne (1996). Bayesianism and Diverse Evidence: A Reply to Andrew Wayne. *Philosophy of Science* 63: 661–65.

Odenbaugh, Jay (2011). True Lies: Realism, Robustness, and Models. *Philosophy of Science* 78: 1177–88.

Odenbaugh, Jay and Anna Alexandrova (2011). Buyer Beware: Robustness Analyses in Economics and Biology. *Biology & Philosophy* 26: 757–71.

Orzack, Steven and Elliot Sober (1993). A Critical Assessment of Levins's 'The Strategy of Model Building in Population Biology" (1966). *The Quarterly Review of Biology* 68: 533–46.

Parker, Wendy S. (2009). Does Matter Really Matter? Computer Simulations, Experiments, and Materiality. *Synthese* 169: 483–96.

— (2018). The Significance of Robust Climate Projections. In: *Climate Modeling: Philosophical and Conceptual Issues*. Ed. by Elisabeth A. Lloyd and Eric Winsberg. Cham: Palgrave Macmillan: 273–96.

— (2020a). Evidence and Knowledge from Computer Simulation. *Erkenntnis*.

— (2020b). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science* 87: 457–77.

Perrin, Jean (1910). *Brownian Movement and Molecular Reality*. Trans. by Fredrick Soddy. London: Taylor and Francis.

Schupbach, Jonah (2018). Robustness Analysis as Explanatory Reasoning. *British Journal for the Philosophy of Science* 69: 275–300.

Smith, George E. and Raghav Seth (2020). *Brownian Motion and Molecular Reality: A Study in Theory-Mediated Measurement*. Oxford: Oxford University Press.

Staley, Kent W. (2018). Securing the Empirical Value of Measurement Results. *British Journal for the Philosophy of Science* (online first).

Steel, Daniel (2008). *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.

Tal, Eran (2012). The Epistemology of Measurement: A Model-Based Account. PhD dissertation. University of Toronto.

Teller, Paul (2001). Twilight of the Perfect Model Model. *Erkenntnis* 55: 393–415.

Weisberg, Michael (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Wheeler, Gregory and Richard Scheines (2013). Coherence and Confirmation through Causation. *Mind* 122: 135–70.

Winsberg, Eric (2010). *Science in the Age of Computer Simulation*. Chicago, IL: University of Chicago Press.

— (2018). What does Robustness Teach us in Climate Science: A Re-Appraisal. *Synthese* (online first).

Woodward, James (2006). Some Varieties of Robustness. *Journal of Economic Methodology* 13: 219–40.