Synthese
https://doi.org/10.1007/s11229-021-03434-8

ORIGINAL RESEARCH

The epistemic value of independent lies: false analogies and equivocations

Margherita Harris[1]

Received: 29 September 2020 / Accepted: 3 October 2021

Abstract
Here I critically assess an argument put forward by Kuorikoski et al. (Br J Philos Sci, 61(3):541–567, 2010) for the epistemic import of model-based robustness analysis. I show that this argument is not sound since the sort of probabilistic independence on which it relies is unfeasible. By revising the notion of probabilistic independence imposed on the models' results, I introduce a prima-facie more plausible argument. However, despite this prima-facie plausibility, I show that even this new argument is unsound in most if not all cases of model-based robustness analysis. This I do to demonstrate that the epistemic import of model-based robust analysis cannot be satisfactorily defended on the basis of probabilistic independence.

Keywords Robustness analysis · Models · Idealizations · Independence · Confirmation

1 Introduction

Any model of a real world phenomenon is bound to include idealizations of some sort (by disregarding some variables, or ignoring or simplifying interactions amongst variables, etc.). Yet we use models to learn about the world constantly, and shall not cease doing so any time soon. A question thus arises: why can we use models to learn about the world despite their idealizing assumptions? If no model is ever a complete and veridical representation of its target system, why do we think of them as 'vehicles for learning about the world' (Frigg and Hartmann 2020)?

There is an idea pertinent to this question that is popular amongst some scientists and philosophers (e.g. Levins 1966; Weisberg and Reisman 2008; Kuorikoski et al. 2010). This idea broadly consists in the following: we can increase our confidence in a model's conclusion by 'studying a number of similar but distinct models of the same

✉ Margherita Harris
m.harris2@lse.ac.uk

[1] Department of Philosophy, Logic and Scientific Method, London School of Economics, London, UK

Published online: 28 October 2021

phenomena' (Weisberg 2013, p. 156). Learning that all these models give the same conclusion, it is claimed, should make us more confident in that conclusion. This way of dealing with model results is usually referred to by its proponents as 'robustness analysis'.

The first explicit discussion of robustness analysis in the context of modelling is usually attributed to the scientist Richard Levins (1966). Below is a frequently quoted passage from Levins on the notion of robustness:

> Even the most flexible models have artificial assumptions. There is always room for doubt as to whether a result depends on the essentials of a model or on the details of the simplifying assumptions. [. . .] Therefore, we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results we have what we can call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies. (Levins 1966, p. 423)

Levins' suggestive remark that 'our truth is the intersection of independent lies' became something of a shibboleth for advocates of robustness analysis. However, what this shibboleth means, and whether or not there is any truth in it, is to this day a source of contention. On the one hand, there are those who argue that robustness analysis has a rightful claim as a method of confirmation of a model's conclusion (e.g. Weisberg 2006, 2013;[1] Lloyd 2009, 2015; Kuorikoski et al. 2010, 2012); and on the other, there are those who disagree (e.g. Cartwright 1991; Orzack and Sober 1993; Odenbaugh and Alexandrova 2011; Justus 2012).

The aim of this paper is to critically assess an argument put forward by Kuorikoski et al. (2010) for the epistemic import of model-based robustness analysis. This assessment is important for two reasons. First, I believe Kuorikoski et al.'s argument is a formal expression of a widely held, but what I believe to be an ultimately misleading intuition. This intuition is the following: a model's conclusion is more likely to hold in the target system if several models lead to that conclusion because it would be a remarkable coincidence if that were not the case. Kuorikoski et al. offer the best available defense of this intuition and that is why I believe it is important to rigorously assess it. Second, several arguments for the epistemic import of robustness analysis that have been offered so far are neither formulated nor defended with sufficient clarity and precision. Hence, in my view, a serious investigation into the epistemic import of robustness analysis must start with a careful reconstruction of those arguments, followed by a rigorous assessment of the tenability of the premises of those arguments. The purpose of this paper is to critically assess Kuoriskoski et al.'s argument in particular; I will conclude that the assumptions on which this argument relies are implausible.

[1] Weisberg might not belong to this camp after all since according to him '[r]obustness analysis helps to identify robust theorems, but it does not confirm them. Such theorems are confirmed via low-level confirmation, the sort of confirmation that licenses the use of a framework to construct models of phenomena in the first place' (Weisberg 2006, p. 742). However, his notion of low-level confirmation is not sufficiently clear which, in my view, leaves enough room for interpretation [see Houkes and Vaesen (2012, pp. 352–353) for some critical reflections on Weisberg's notion of low-level confirmation].

I must point out that I am not the first to object to Kuorikoski et al.'s (2010) argument. Odenbaugh and Alexandrova (2011) have also questioned the validity of some of its assumptions. However, in my view, their objections were insufficient, and thus so were Kuorikoski et al.'s (2012) responses. In this paper, I hope to show more forcefully that the assumptions that underscore Kuorikoski et al.'s argument are untenable.

2 Setting the scene

For the purpose of this discussion, I will adopt the following working definition of a model: a model is a (concrete or mathematical) entity that can be used to represent a target system and which includes substantial assumptions about the target system and various different idealizations. Following Kuorikoski et al.'s (2010) definition, I will assume that 'substantial assumptions identify a set of causal factors that in interaction make up the causal mechanism about which the modeller endeavours to make important claims' (ibid., p. 547). And I will take an idealization to mean any sort of known departure from a veridical representation of the target system. Of course there are various different kinds of idealizations; following Kuorikoski et al. (2010), I will assume that there are two conceptually distinct kinds: Galilean assumptions and tractability assumptions. Galilean assumptions 'serve to isolate the working of the core causal mechanism by idealising away the influence of the confounding factors' (ibid., p. 547). According to Kuorikoski et al., despite being unrealistic with respect to the model's target system, Galilean assumptions have a causal interpretation: 'they state that a factor known or presumed to have an effect is absent' (ibid., p. 547). Tractability assumptions, on the other hand, are assumptions that are introduced 'only for reasons of mathematical tractability' and, in contrast to Galilean assumptions, they often 'have no empirical merit on their own' (ibid., p. 548). This is why, according to Kuorikoski et al., 'unlike Galilean idealisations, for many tractability assumptions it is often unclear what it would mean to replace them with more realistic ones: if it were possible to do without this kind of assumptions they would not be introduced in the first place' (ibid., p. 548). Throughout the paper, I will denote the substantial assumptions by C, all the Galilean assumptions by G and all the tractability assumptions by T.

As an illustration of this working definition, take the Lotka Volterra model, one that is often used by proponents of robustness reasoning (e.g. Weisberg and Reisman 2008). This model is used to represent the behaviour of real-world predator-prey systems and is described by the following two coupled ordinary differential equations:

$$\frac{dV}{dt} = rV - (aV)P \tag{1}$$

$$\frac{dP}{dt} = b(aV)P - mP \tag{2}$$

Where $V(t)$ and $P(t)$ stand for the size of the pray and predator population at time t, respectively. The constant r stands for the growth rate of the prey population and the constant m stands for the death rate of the predator population. The constant a

stands for the predator attack rate and the constant b stands for the predator conversion efficiency.

In line with discussions of this model in the literature on robustness, I will take the substantial assumption in this model to be the assumption that the target predator-prey system is negatively coupled (i.e. increasing the size of the predator population decreases the size of prey population and increasing the size of the prey population increases the size of the predator population). And in line with the definition given above, an example of Galilean assumption could be the assumption that aside from the size of the predator population, there are no other factors that may affect the size of the pray population (such as limited resources). Notice that although this is an unrealistic assumption with respect any real-world predator-prey system, it could in principle be replaced with a more realistic assumption; for instance by replacing it with the assumption that there is a maximum carrying capacity to the growth rate of the pray population as done by Weisberg (2006). According to Kuorikoski et al. (2012), an example of a tractability assumption could be the *specific* functional form used to describe the rate of prey capture per predator (this model assumes that there is a linear increase in pray capture with pray density). Kuorikoski et al. (2012) consider this to be a tractability assumption in so far as *any* assumed functional form for the rate of prey capture will 'strictly speaking be false for any natural population' (ibid., p. 8).[2]

A famous result of the Lotka–Volterra model is what is known as the *Volterra property*: a general biocide will increase the size of the the prey population and decrease the size of the predator population. A special feature of this property is that it is robust across a variety of different predator-prey models, which share the Lotka–Volterra model's substantial assumption that the predator-prey system is negatively coupled, but that involve various different idealizations. This has given rise to the celebrated *Volterra principle*:

> *The Volterra principle*: Ceteris paribus, if a two-species predator-prey system is negatively coupled, then a general biocide will increase the size of the prey population and decrease the size of the predator population.

The Volterra principle is a hypothesis about the *actual* world, not about model land. But due to the use of the ceteris paribus clause it is, in my view, not very clear how one should interpret this principle. However, for the purpose of this paper and in line with discussions about the Volterra principle in the literature (e.g. Weisberg 2006; Kuorikoski et al. 2012), I will assume the Volterra principle is a causal hypothesis;

[2] This last claim may strike the reader as being a little strong since it certainly seems possible, in principle, that a particular assumed functional form could be true. Crucially, however, even if any assumed functional form for the rate of prey capture is unlikely to be strictly true, there is certainly a sense in which one particular functional form could be more approximately accurate than another. And if this is the case, then it is not clear why one should think of these assumptions (i.e. specific choices of functional forms) as being introduced 'only for reasons of mathematical tractability', as Kuorikoski et al. seem to suggest. It is also worth pointing out that Kuorikoski et al.'s (2010) case study is not the Lotka–Volterra model, but a model in geographical economics. According to them, examples of tractability assumptions in that case are 'specific functional forms of utility [...], production [...] and transformation technology [...]' (ibid., p. 556). It seems to me that the above considerations should apply to these examples too. I shall return to the question of how we should interpret tractability assumptions in Sect. 5.

that is, according to the Volterra principle, a two-species predator-prey system that is negatively coupled (C) has the efficacy to produce the Volterra property (R) as long as there is no other causal factor that is preempting this efficacy, and throughout the paper I will denote it as: in the actual world, R causally depends on C.[3]

According to Weisberg (2006) the discovery of the *Volterra principle* through the analysis of predator-prey models is a prime example of robustness analysis, which he characterizes as a four step procedure: (i) evaluate whether a group of models share a common result R; (ii) determine whether this set of models share a common substantial assumption C; (iii) formulate the robust theorem: a conditional statement linking the common substantial assumption C to the robust property R, prefaced by a ceteris clause; (iv) conduct "stability analysis" of the robust theorem, with the aim of finding out what conditions will defeat the connection between C and R.

Kuorikoski et al. (2010) largely agree with Weisberg's characterization of robustness analysis. However, they stress that it is only the failure of robustness with respect to tractability assumptions that is epistemically problematic 'because it suggests that the result is an artefact of the specific set of tractability assumptions, which in many cases have no empirical merit on their own' (ibid., p. 548). In contrast, in their view, the failure of robustness with respect to Galilean assumptions is *not* epistemically problematic because 'it [merely] suggests a new empirical hypothesis about a causally relevant feature in the modelled system' (ibid., p. 552). This is why, as we will see in the next section, Kuorikoski et al.'s argument for the epistemic import of robustness analysis focuses exclusively on models that involve different tractability assumptions, while keeping constant all Galilean assumptions.

3 An argument from coincidence?

According to Kuorikoski et al. (2010, p. 560):

> Levins' (1966) unclear but intuitively appealing claim that 'our truth is the intersection of independent lies' could be taken to mean that result R can be derived from mechanism-description C using multiple independent sets of untrue tractability assumptions. Various falsities involved in the different derivations do not matter if robustness analysis shows that result R does not depend on them.[4]

For Kuorikoski et al., the epistemic value of robustness analysis lies in the very *independence* of the different untrue tractability assumptions involved in the models, since if they are independent in the right sort of way, then (in their view) it can be shown that model-based robustness analysis is 'a species of general robustness analysis in

[3] More generally, this is how I will denote what Weisberg calls a 'robust theorem'. Hence, rather than writing 'Ceteris paribus, if C then R', which is Weisberg's own formulation of a robust theorem, I will write 'in the actual world, R causally depends on C'. What justifies the removal of the ceteris paribus clause is the fact that I am interpreting the robust theorem as a claim about (stable) capacities, which are introduced by Cartwright to explain causal laws and render them universal in character: if C has the capacity to produce R then C carries this capacity from situation to situation (Cartwright 1989, p. 145).

[4] Kuorikoski et al. (2010) use the notation R_M to refer to a model's result, but to be consistent with my notation I replaced all instances of R_M with R.

the sense discussed by Wimsatt (1981) and that the same epistemic rationale applies to it' (Kuorikoski et al. 2010, p. 559). However, aside from mentioning that according to Wimsatt,

> [robustness] provides epistemic support via triangulation: a result is more likely to be real or reliable if a number of different and mutually *independent* routes lead to the same conclusion. *It would be a remarkable coincidence if separate and independent forms of determination yielded the same conclusion if the conclusion did not correspond to something real* (ibid., p. 544, my emphasis),

they neither clarify *what* is the epistemic rationale on which Wimsatt relies in his defense of the epistemic value of general robustness analysis, nor (as we will see in the next section) do they rely on it for their own defense of the epistemic value of robustness analysis. Hence, the sole aim of this section is to reflect on what to make of the very last line of the quote above: that it would be a remarkable coincidence if separate and independent forms of determination yielded the same conclusion if the conclusion did not correspond to something real.

Indeed, it is not hard to find cases where it would be a remarkable coincidence if the same conclusion of distinct forms of determination did not correspond to something real. Suppose, for instance, that I weigh myself on several distinct scales from different manufacturers and different suppliers and they all show that I weigh 300 pounds, a lot more than I thought I would. Despite this, I think to myself 'it would be too remarkable a coincidence if all these scales showed that I weigh 300 pounds if I didn't really weigh 300 pounds. I must weigh 300 pounds!' No one should accuse me of irrationality here. But what kind of coincidence would this be? It would be the following: although each scale may mislead me, due to the possible presence of a faulty mechanism, I have no reason to suppose that these scales share the *same* faulty mechanism. Hence the fact that all these scales would mislead me in the same way for different reasons seems an extremely implausible concurrence of events. On the other hand, if my weight really was 300 pounds, and hence the scales' readings corresponded to something real (i.e. my weight), this concurrence of events would no longer seem a remarkable coincidence: under this hypothesis, all my scales are working well, and so through the right sort of causal mechanism my weight is causing the scales' readings to agree. Hence, it seems rational for me to opt for the hypothesis that does not involve a remarkable coincidence.[5]

Can one apply the same argument from coincidence that I applied to my scale example to the context of model-based robustness analysis? For this to be the case, if the same conclusion is implied by multiple models, each containing different tractability assumptions, one should be able to claim in this case too that it would be a remarkable coincidence if all these models implied the same conclusion, if the conclusion did not correspond to something real *and* that the coincidence would vanish if the conclusion *did* correspond to something real. However, this is not the case. For, without further

[5] Notice that this argument from coincidence crucially relies on the assumption that there is no *systematic error*, which seems reasonable in this case because all the scales come from different manufacturers and different suppliers. However, without this assumption, the convergence of the scales' readings would at best only entitle me to infer that that this convergence is not due to chance, but it would 'not indicate it is due to any specific cause.' (Mayo 1986, p. 45)

justification, the fact that these models all imply the same conclusion, *despite* each and every one of them containing false tractability assumptions, should still strike one as being a remarkable concurrence of events *even if* that conclusion were to correspond to something real. In other words, the fact that distinct models involving different false tractability assumptions give the same conclusion *is a coincidence*, but not one that seems to be explained away by the hypothesis that the conclusion corresponds to something real.

The crucial difference between my scale example and model-based robustness analysis is the following. In my scale example, we are able to postulate a process that links the cause (i.e. my weight) to the effect (i.e. the scale's readings) and it is the very postulation of this causal process that explains why the scales' readings are the same. But in the case of models, we cannot postulate a causal process that links the reality of the conclusion to the models' conclusions. Scales are measuring instruments, they *measure* things through a *causal* process. Models are *not* measuring instruments, they don't measure things through a causal process; hence postulating that a model's conclusion is real is not enough to explain why distinct models agree on that conclusion. So it seems to me that, in order for the reality of the models' conclusion to help us explain away this coincidence, we would also have to tell a story about why the models that we are considering in a given case must *all* agree on that conclusion if the conclusion were to correspond to something real. But whether that story can in fact be told does not seem to be something that can just be assumed.[6]

Perhaps, one could attempt to explain away this coincidence by merely appealing to the world of models and not the one outside them. But what would it mean to find a (non-causal) explanation for this coincidence in the world of models? As one of this article's reviewer pointed out, one may be tempted to answer this question by simply noting that 'all models share a common core, which could be the main driver of the common conclusion'. However, this assertion must be equivalent to the claim that a particular set of models which all share a common core all give the same conclusion. Now, if that set of models is the same set whose conclusion we have just observed, then this would be a tautological explanation: the explanation for why all the models in our ensemble give the same conclusion is that they all give the same conclusion. So this can't be right. If the claim is meant to appeal to a more general class of models of which our ensemble is but a subset, then this raises at least two questions: what is the relevant class of models? And in what sense would the fact that a more inclusive set of models all entail a conclusion provide an explanation for why a subset of it provides that conclusion? Alternatively, one might attempt to explain this coincidence by showing that the models in our ensemble are special cases of a more general model which gives the same conclusion. For instance, Räz (2017) demonstrates that as long as a condition that ensures that the average abundance of a system coincides with the relevant equilibrium is satisfied, the Volterra principle

[6] With this I am not suggesting that such a story can never be told. For instance, one might have reasons to believe that all models considered can be adequate (not by mere luck) representations of the target system for the purpose at hand, despite making incompatible or false assumptions about the target system (see Parker (2020) for an adequacy for purpose view of model evaluation). However, what I essentially argue is that the hypothesis that the models we are considering can indeed *all* be adequate (not by mere luck) is not something that can simply be assumed, but rather, it must be justified on a case-to-case basis.

holds for a more general model. However, there are of course many cases where it cannot be shown that different models are special cases of a more general model, especially when models involve different representational frameworks (e.g. Weisberg and Reisman (2008) also consider an individual based model version of the Lotka–Volterra model). In any case, I think it is important to note that whether or not it is possible to find an explanation for this coincidence in the world of models, this explanation *alone* would not help us infer anything about the world outside of them (which is ultimately what we are interested in).

So there seems to be a prima-facie clear difference between my scale example and the example involving models: in the former a causal argument from coincidence for the truth of the conclusion seems to be justified, whereas the same cannot be said of the latter. Although Kuorikoski et al. do not advocate a causal argument from coincidence to defend their view about the epistemic import of robustness analysis (as we will see in the next section), they nonetheless do make several equivocatory remarks that nudge the reader in that direction. Consider, for instance, this passage:

> Before conducting robustness analysis we do not know for sure which part of the models is responsible for the result, although modellers usually have strong intuitions about this issue. If a result is implied by multiple models, each containing different sets of tractability assumptions, we may be more confident that the result depends not on the falsities we have introduced into the modelling, but rather on the common components [. . .]. Robustness analysis thus increases our confidence in the claim that the modelling result follows from the substantial assumptions, i.e. that some phenomenon can be caused by the core mechanism. (ibid., p. 551)

In the above quote, Kuorikoski et al. are suggesting that if a result is implied by multiple models, each containing different sets of tractability assumptions, our confidence that the result R *depends* on the common components (i.e. the substantial assumptions C), rather than the various different false tractability assumptions, should increase. At first glance, this reasoning may seem analogous to the reasoning that I applied to my scale example (i.e a causal argument from coincidence). However on closer inspection, it is clearly based on an equivocation: one that, like most equivocations, has the potential to mislead. To see clearly why this is, it will be helpful to reconstruct Kuorikoski et al.'s above reasoning into a set of premises and a conclusion from those premises. Let M_i stand for a given model; the premises of Kuorikoski et al.'s argument are then the following:

$\boldsymbol{P_1}$: M_1 implies result R

$\vdots$

$\boldsymbol{P_n}$: M_n implies result R

By assumption, a model consists of substantial assumption C, Galilean assumptions G and tractability assumptions T. And, also by assumption, we are focusing on a class of models that all have the the same substantial assumptions and Galilean assumptions but that differ in their tractability assumptions. So the above premises can be rewritten as:

$\boldsymbol{P_1}$: $C\&G\&T_1$ implies result R

$\vdots$

$\boldsymbol{P_n}$: $C\&G\&T_n$ implies result R

According to Kuorikoski et al.'s above reasoning, from $P_1 \ldots P_n$, we are entitled to have more confidence in the following conclusion:[7]

Robustness conclusion (**R-C**): R depends on C.

In light of this argument, three observations are in place. First, notice that **R-C** is ambiguous between

R-C-model: In model land, R depends on C, and
R-C-world: In the actual world, R depends on C.

Second, given that all parts of a model are used in the derivation of a model's result, the only possible interpretation of **R-C-model** must be the following:

R-C-model: All models involving C in the relevant class imply result R.

But this interpretation of **R-C-model** is unclear without a specification of what is the relevant class of models. Are $M_1 \ldots M_n$ considered to be merely samples of this class or should we think of them as constituting the entire class? If the former, what is the relevant class? if the latter, why is this an interesting class? That is, why should we care about $M_1 \ldots M_n$? Without an answer to these questions it is really not clear how one should in fact interpret **R-C model**. As a side note, it is worth mentioning that according to Weisberg (2006, p. 739), 'if a *sufficiently heterogeneous* set of models for a phenomenon all have the common structure, then it is very likely that the real-world phenomenon has a corresponding causal structure' (my emphasis; Levins (1993) makes similar remarks). One might think that Weisberg's notion of sufficient heterogeneity is relevant to the questions I have just raised. However, and leaving aside the lack of clarity surrounding Weisberg's notion of sufficient heterogeneity, it seems to me that it is not in fact pertinent here. This is because, according to Weisberg, the purpose of robustness analysis is merely to 'identif[y] hypotheses' (ibid., p. 741) not to confirm them. Hence for Weisberg the only source of worry when it comes to evaluating the epistemic import of robustness analysis is the fact that the theorems generated by robustness analysis are 'conditional statements, further attenuated with ceteris paribus clauses' (ibid., p. 739). That is, the worry is that the robust theorem and its predictions hold only under certain conditions, but not in others. Hence the reason why we want a sufficiently heterogeneous set of models, according to Weisberg, is to address *this* worry: by considering models that satisfy various different conditions we can raise our confidence that the theorem holds more generally. However, Kuorikoski et al.'s concern is of an altogether different nature. Kuorikoski et al. worry that due to the presence of tractability assumptions (which are assumed to be strictly false for any target system) the ceteris paribus theorem might not be a theorem about the real world in the first place. This is not to say that there is no answer to the questions I raise above,

[7] More confidence than the one we would have if we only had P_1.

but it is to say that Weisberg's appeal to the notion of a sufficiently heterogeneous set of models should not be seen as an attempt to answer *those* questions. And we will see that Kuorikoski et al.'s argument for why robustness analysis should increase one's confidence in the ceteris paribus theorem also averts these questions all together, by relying on a concept of independence instead.

Third, even if Kuorikoski et al. could give a clear interpretation of **R-C-model**, the transition from **R-C-model** to **R-C-world** needs to be justified. Doing this silently (as done in this argument) is a petitio principii because what needs to be shown is precisely that the transition from model land to the actual world is legitimate.

In the next section I will turn to what I consider to be Kuorikoski et al.'s official argument for the epistemic import of model-based robustness analysis. Indeed their 'official argument' could be interpreted as indirectly addressing these criticisms, so I will now turn to it.

4 What is Kuorikoski et al.'s argument?

Here is what Kuorikoski et al. (2010) write:

> Modelling can be considered as an act of inference from a set of substantial assumptions to a conclusion [. . .]. Tractability assumptions are typically needed for the process of inference to be feasible, but these assumptions may induce errors in the modelling process: they may lead us to believe falsities about the world even if the substantial assumptions are true. [. . .] We thus propose that the modeller should have no positive reason to believe that if one tractability assumption induces a certain kind of error (due to its falsehood) in the result, so does another one. *Given that the modelling result of interest* (R) *is correct, prior probabilities concerning whether* R *can be derived from* $C\&T_1$ *or* $C\&T_2 \ldots C\&T_n$ *should be (roughly) independent. If the probabilities are independent in this way, then observing that the models lead to the same result rationally increases our degree of belief in the result*.[8,9] (ibid., p. 561 my emphasis)

There is a lot going on in this quote and I will need to introduce some new notation in order to unpack it. Let R_T be the proposition that result R is instantiated in the target system. And let R_k be the proposition that result R is derived by the kth model. From the above passage, the argument of Kuorikoski et al. for the epistemic import of model based robustness analysis seems to be the following:

[8] It is clear from Kuorikoski et al.'s (2010) general discussion that "the result" at the end of this quote is not meant to refer to the hypothesis that R holds in the target system, but to the hypothesis that in the actual world, R causally depends on C. For instance, robustness analysis is supposed to increase our degrees of belief that the Volterra principle is correct, not that the Volterra property is instantiated in the target system. Our confidence that the Volterra property is instantiated in the target system might increase as a result of this to the extent that we believe that the assumption that the predator-prey system is negatively coupled is correct *and* also to the extent that we believe that there are no disrupting factors in the target system. But this should be seen as merely a possible *by-product* of the confirmatory power of robustness analysis.

[9] Kuorikoski et al. (2010) use the notation V_i to refer to tractability idealizations. To be consistent with my notation I have replaced all instances of V_i with T_i.

> **The argument**. Assume that we observe that a model with substantial assumption C and tractability assumptions T_1 gives result R. Then we will have some degrees of belief that the hypothesis h: "in the actual world, R causally depends on C" is true. Suppose further that in addition to our first model, we observe that several other models sharing the same substantial assumptions C, but differing in their tractability assumptions T_i give the same result R. This should rationally increase our degrees of belief in the hypothesis h, because it is reasonable to assume that the models' results R are probabilistically independent conditional on R_T (and $\neg R_T$).[10](i.e. because it reasonable to assume that $Pr(R_1 \& \ldots \& R_n | R_T) = Pr(R_1|R_T) \times \ldots \times Pr(R_n|R_T)$ and $Pr(R_1 \& \ldots \& R_n | \neg R_T) = Pr(R_1|\neg R_T) \times \ldots \times Pr(R_n|\neg R_T)$).

To adequately assess this argument, I will need to make a couple of clarifications. First, as mentioned in Sect. 2, according to Kuorikoski et al. tractability idealizations are not the only kind of idealizations typically needed for the process of inference to be feasible; various Galilean idealizations will also be needed. So to be a little more rigorous one should say that a modeling result R can be derived from $C \& T_i \& G_i$ rather than just $C \& T_i$. But given that according to Kuorikoski et al., Galilean assumptions 'serve to isolate the working of the core causal mechanism by idealising away the influence of the confounding factors', I will assume for the sake of argument that Galilean assumptions, rather than being problematic, are always helpful in establishing causal dependencies. Therefore, I will assume that each model involves the same Galilean assumptions and I will set them aside for the time being.

Second, Kuorikoski et al. (2010, p. 545) reference Bovens and Hartman (2003, pp. 96–97) to justify that the sort of probabilistic independence invoked in this argument is enough to guarantee that our degrees of belief in the hypothesis h should rationally increase. Indeed, Bovens and Hartmann (2003, pp. 96–97) do show that under certain *specific* conditions, if distinct instruments' results are probabilistically independent conditional on the assumption that the testable consequence of a hypothesis is correct (or not correct),[11] then observing multiple positive results from distinct instruments should increase our degrees of belief in that hypothesis. But Bovens and Hartmann's demonstration depends on several *other* conditions being satisfied! One of these, for instance, is that an unreliable instrument must 'randomize at some level a':

> Our model does not apply to unreliable instruments that do not randomize, but rather provide accurate measurements of other features than the features they are supposed to measure. In effect, our model exploits the coherence of the reports as an indicator that the reports are obtained from reliable rather than unreliable instruments. But if unreliable instruments accurately measure features other than the ones they are supposed to measure, then they will also provide coherent

[10] In the above quote, Kuorikoski et al. do not explicitly claim that the models' results must also be probabilistically independent conditional on $\neg R_T$. But without this assumption this argument is not valid, so I am assuming this is just a slip of the hand.

[11] Bovens and Hartman's (2003) definition of a testable consequence of a hypothesis is as follows: 'the probability of the consequence given that the hypothesis is true is greater than the probability of the consequence given that the hypothesis is false' (ibid., p. 90).

> reports and so the coherence of the report is no longer an indicator that they were obtained from reliable instruments. (Bovens and Hartmann 2003, p. 95)

So if Kuorikoski et al. want to appeal to Bovens and Hartman's demonstration to justify the validly their argument, then they also *must* rely on the assumption that unreliable models (in contrast to reliable ones) do not tend to give coherent reports. In other words, they must rely on the assumption that unreliable models cannot be systematically biased. As mentioned earlier, for simple measurement devices like scales this seems to be an adequate assumption in some cases: for any unreliable scale (i.e. malfunctioning) from a different manufacture and supplier it can be reasonable to assume that whether or not it shows that I weigh 300 pounds (if I really weight 300 pounds) is a matter of chance (and the same if I do not really weigh 300 pounds). But in the case of models, this is a *substantial* assumption that would need to be further justified and nowhere in the paper do Kuorikoski et al. do so. The fact that the validity of this argument depends on substantial assumptions, that have not been made explicit by Kuorikoski et al., is in my view already an important weakness of the argument, one that is possibly strong enough to reject it. But for the sake of argument, in this paper I am going to assume this argument is valid and hence I will only critically assess whether, if valid, it is also sound.

For this argument to be sound Kuorikoski et al. need to convince us that it is reasonable to suppose that the probabilities of the models' results are independent conditional on R_T (and $\neg R_T$). But although this is a weaker notion of probabilistic independence than unconditional probabilistic independence, it is still an extremely strong notion of independence. This sort of independence demands that if I know that the models' result R is instantiated in the target system, then learning that a model gives result R should not at all affect my degrees of belief that another model would also give result R. But this is an unreasonable demand! To see why this is, suppose that I know that R is instantiated in the target system. If this is all I know, then there is no reason to think that prior to learning the models' results, I will have much confidence in the fact that result R will be derived by these models (even if I know that C is instantiated in the target system as I have no reason to suppose that R causally depends on C!). But now suppose that I learn that R can be derived by one of the models, consisting of substantial assumptions and tractability assumptions $C\&T_1$. Kuorikoski et al.'s notion of independence demands that my degrees of belief about whether R can be derived from another model $C\&T_2$ should not change. But this is implausible: I know that the two models share substantial assumptions C, so if I learn that R can be derived from the first model, my degrees of belief about whether R will be derived by the second model are bound to change: I now seem to be in a much better position than I was before to make an informed guess that result R will be derived by second model.

To make my objection more vivid, consider the Lotka–Volterra model alongside another model which shares the substantial assumption C that the system is negatively coupled, but that involves a different set of tractability assumptions. Suppose that all I know is that the Volterra Property is instantiated in the target system. Given that I have no knowledge regarding what the Volterra Property causally depends on, there is no reason to suppose that I should have much confidence in the fact that the Volterra property will be derived by these two models. But now suppose that I

learn that the Lotka–Volterra model has the Volterra property. Surely my degrees of belief that another model sharing the same substantial assumption C will also give the Volterra property will greatly increase. Why? Because the two models share substantial assumption C, and hence the fact that the first model had the Volterra property when assumption C was involved will greatly increase my confidence that the second model will also give the Volterra property.

Notice that the situation in my scale example is very different. Conditional on the fact that I really weigh 200 pounds, it seems reasonable to suppose that learning that a scale shows that I weigh 200 pounds will not affect my degrees of belief that another distinct scale will also show that I weigh 200 pounds. Effectively the difference consists in the following. Learning that a scale shows my weight does not affect my degrees of belief that another scale will also show my weight, because I already knew that scales are supposed to measure my weight *prior* to learning the first scale's reading. Whereas in the case of models, the situation is very different: If all I know is that R_T is true, learning that a model with substantial assumption C gives result R will affect my degrees of belief that another model sharing substantial assumptions C will also give result R, because learning that the first model gives result R when C is involved, gives me some reasons to expect that the second model, which also involves C, will also give result R; reasons that I didn't have prior to learning the first model's result.

It is worth mentioning that Schupbach (2018) has also objected to this notion of conditional independence in the context of model-based robustness analysis, but his objection relies on the assumption that the distinct models will share many unrealistic assumptions and so 'discovering that one of the models is unreliable should often greatly increase our confidence that the other is too' (ibid., p. 283). In other words his objection is the following: conditional on the result R not being correct, the probabilities of the models' results R cannot be independent. This is indeed a very good objection, but it is weaker than mine because it relies on the idea that models will invariably share many unrealistic assumptions. Although this is certainly true in most if not all cases, the reason why I object to this notion of conditional independence is because of the very fact that the distinct models share substantial assumptions C and so it will hold regardless of whether or not they share any unrealistic assumptions.

All said and done, it seems to me that models cannot be independent in the way required by Kuorikoski et al's argument. Hence this argument is not sound and should be rejected.

5 A prima-facie more plausible argument (and yet...)

It is worth noting that alternatively to Bovens and Hartman's demonstration, Kuorikoski et al. might want to appeal to Fitelson (2001)'s demonstration instead. Fitelson (2001) shows that with respects to several popular Bayesian measures of confirmation, if two results a_1 and a_2 individually confirm an hypothesis H and if a_1 and a_2 are confirmationally independent regarding H, i.e. $c(H, a_1|a_2) = c(H, a_1)$ and $c(H, a_2|a_1) = c(H, a_2)$, then a_1 and a_2 together confirm H to a greater extent than either a_1 or a_2 does separately, i.e. $c(H, a_2\&a_1) > c(H, a_1)$ and $c(H, a_2\&a_1) >$

$c(H, a_2)$.[12] Fitelson further *suggests* that a sufficient condition for a_1 and a_2 to be confirmationally independent regarding H is that they be probabilistically independent conditional on H (and $\neg H$), i.e. $Pr(a_1 \& a_2|H) = Pr(a_1|H)Pr(a_2|H)$ and $Pr(a_1 \& a_2|\neg H) = Pr(a_1|\neg H)Pr(a_2|\neg H)$.[13] If this is right then Kuorikoski et al. could rely on Fitelson's 'result' but only if they are willing to change the notion of conditional independence they demand on the models' results. That is if they want to rely on Fitelson's demonstration then their argument should be rephrased as follows:

> **A second argument**. Assume that we observe that a model with substantial assumption C and tractability assumptions T_1 gives result R. Then we will have some degrees of belief that the hypothesis h: "In the actual world, R causally depends on C" is true. Suppose further that in addition to our first model, we observe that several other models sharing the same substantial assumptions C, but differing in their tractability assumptions T_i give the same result R. This should rationally increase our degrees of belief in the hypothesis h, because it is reasonable to assume that the models' results are probabilistically independent conditional on the hypothesis h (and $\neg h$).

This argument strictly relies on the assumption that each model's result individually confirms h. This does not seem to be an unreasonable assumption in most cases, but it is still an assumption that needs to be acknowledged.[14]

For this argument to be justified Kuorikoski et al. need to convince us that it is reasonable to suppose that the probabilities of the models' results are independent conditional on the hypothesis h (and $\neg h$). For instance, in the case of the Volterra principle, we want our models' results to be probabilistically independent conditional on the *Volterra principle*, rather than conditional on the *Volterra property* as in Kuorikoski et al's original argument. This kind of conditional probabilistic independence of the models' results is prima facie more plausible: conditional on the hypothesis that in the actual world R causally depends on C, the fact that two models (consisting of $C \& T_1$ and $C \& T_2$ respectively) share substantial assumptions C seems less of a salient factor when assessing one's degrees of belief that one model will give R if one has learnt that another model has already given R. To see why this is, suppose that I know that in the actual world R causally depends on C (e.g. I know that the Volterra principle is correct). In this case it seems that already *prior* to learning the models' results, my confidence in the fact that R will be derived by these models is going to be relatively high, since I know that they both involve C. That is, in this case knowing that R causally depends on C seems to already put me in a good position to make an informed guess that result R will be derived by both models. But then in this case, learning that one model gives R does not seem to put me in a better position to make an informed guess about whether the second model will also give R. Hence it seems, prima facie,

[12] A confirmation measure $c(H, a)$ measures the degree of confirmation lent to H by a. I use the notation $c(H, a_i|a_j)$ to indicate the degree of confirmation lent to H by a_i, conditional on a_j.

[13] To the best of my knowledge, however, Fitelson (2001) does not actually prove this result.

[14] Although someone may very well question this assumption too: if all the models involve false assumptions, why would we have to accept that there is any confirmation relation at all? I.e. why should we think that $c(H, R_1)$, $c(H, R_2)$ etc. . . . are not all equal to zero? Indeed, if they are all equal to zero, the machinery does not get off the ground.

plausible to assume that if I know that in the actual world R causally depends on C, learning that the model consisting of $C\&T_1$ gives R should not change my degrees of belief that a model consisting of $C\&T_2$ will give R.

However, despite this prima-facie plausibility, this sort of independence is still unrealistically strong in most cases, if not all. And I see two reasons for this. First, despite differing in *some* tractability assumptions, models will more often than not share many other tractability assumptions. But then, in these cases, if I learn that result R can be derived from the first model, it is unreasonable to suppose that my degrees of belief that R will be derived by the second model are not going to change: even if I know that R causally depends on C in the actual world, if the second model shares some tractability assumptions with the first model, learning that the first model gives R *will* put me in a better position to make an informed guess about whether the second model will also give result R.[15]

So it seems that the only scenario in which it might be reasonable to assume that models' results satisfy this sort of independence is in those rare cases in which models share the same substantial assumptions C, but share no tractability assumptions. As far as the Lotka–Volterra model is concerned, Kuorikoski et al. (2012) argue that Weisberg and Reisman's (2008) individual based model - in which the Lotka Volterra model's variables, parameters and other assumptions are all translated into individual-based terms- is one such case:

> Weisberg and Reisman (2008) also discuss a way in which practically all the tractability assumptions can be expected to be independent: the derivation of the robust theorem in a completely different modelling framework. Whereas the class of Lotka–Volterra models described above are sets of differential equations relating population aggregates, the Volterra principle can also be demonstrated using agent based computational models. Such models represent the same core causal mechanisms, albeit describing them at an individual level. However, the radical difference in the modelling framework means that the tractability assumptions, although still unavoidable, are of an altogether different kind: they relate to the behavioural rules of individuals and the spatial representation of their environment, rather than to population-level generalisations as in the original Lotka–Volterra models. (Kuorikoski et al. 2012, p. 898)

This is an instance of what Weisberg and Reisman (2008) call representational robustness, which involves changing the representational framework of a model and assessing whether the same result still obtains.[16] If cases of representational robustness really are cases in which models share the same substantial assumptions C, but differ in *all* their tractability assumptions, as Kuorikoski et al. claim, then perhaps the sort of independence invoked in this argument is plausible in such cases. But notice that, if cases of representational robustness are the only kind of cases in which the sort of independence invoked in this argument is plausible (as I am suggesting) then the scope of this argument is clearly very restricted. Hence this argument is not applicable in

[15] Schupbach (2018, p. 285) makes essentially the same objection.

[16] As Weisberg and Reisman (2008, p. 120) note 'the representational framework of the model is a general description of the type of state variables and the type of transition rules the model employs.'

most instances of robustness analysis that are encountered in scientific and economic modelling.

But I think there is a second reason to doubt that this sort of independence is reasonable, even in the rare cases where models share the same substantial assumptions C but differ in all their tractability assumptions. And it has to do with the very nature of tractability assumptions. As mentioned in Sect. 2, for Kuorikoski et al., tractability assumptions are assumptions that are introduced 'only for reasons of mathematical tractability' and, in contrast to Galilean assumptions, they often 'have no empirical merit on their own'. According to Kuorikoski et al. (2012), as far as the Lotka–Volterra model is concerned, an example of a tractability assumption is the *specific* functional form used to describe the rate of prey capture per predator, since *any* assumed functional form for the rate of prey capture will 'strictly speaking be false for any natural population' (ibid., p. 8). But although it may be true that any assumed functional form will strictly speaking be false for any real-world predator-prey system, there is certainly a sense in which one particular functional form might be more adequate to describe the rate of prey capture than another, despite both of them being strictly false. And there is also a sense in which one might believe that at most one functional form amongst the ones one is considering is adequate, even if one lacks the knowledge to determine which one. But then, to the extent that this is the case, I think it is unreasonable to assume that the results of two distinct models that differ in their tractability assumptions are probabilistically independent conditional on h. And here is why. Suppose that I know that in the actual world R causally depends on C (e.g. I know that the Volterra principle is correct) and consider two distinct models that share substantial assumption C (e.g the assumption that the predator-prey system is negatively coupled) but that assume distinct functional forms for the rate of prey capture per predator. Suppose further that I believe that at most one of these two functional forms can adequately represent the actual rate of prey capture per predator. Prior to learning the models' results, I will have some degrees of belief in the fact that R (e.g. the Volterra property) will be derived by these models. But now suppose that I learn that one of these models gives result R. This should give me further reasons to suppose that the particular functional form assumed in this model can adequately describe the rate of prey capture per predator, reasons that I didn't have prior to learning the model's result. And if that's the case, then this should also give me further reasons to suppose that the functional form assumed in the other model is inadequate. But then, to the extent that this is the case, it is unreasonable to suppose that learning that the first model gives result R will not change my degrees of belief that the second model will give result R. I now seem to have some further reasons to suppose that the second model does not adequately represent predator-prey systems, which should reasonably decrease my degrees of belief that the second model will give result R. Hence, it is hard to see why it would be reasonable to assume that conditional on hypothesis h being correct, these two models' results are probabilistically independent.

Hence, due to the fact that in most cases of robustness analysis models will very often share many tractability assumptions, and due to the very nature of at least some tractability assumptions, I think it is in fact rather hard to justify the sort of probabilistic independence invoked in this argument in most if not all cases of model-based robustness analysis.

Before concluding, I would like to make a couple of remarks. First, throughout the paper, I have assumed that there is a clear distinction between Galilean assumptions on the one hand and tractability assumptions on the other. In particular, I assumed that Galilean assumptions are always helpful in establishing causal dependencies by idealizing away the influence of the confounding factors. This allowed me to assume that robustness failure in a modelling result with respect to Galilean assumptions is never epistemically problematic, since it merely suggests a new empirical hypothesis about a causally relevant feature. Without this assumption it would have been impossible to even begin to assess Kuorikoski et al.'s argument for the epistemic import of robustness analysis. This is because this assumption allowed me to give an *empirical* (causal) interpretation to the robust theorem. In other words, with this assumption I was able to interpret the robust theorem as a *causal* hypothesis about the *real* world, a hypothesis that one can both conditionalise on and confirm. However this distinction is, in my view, a lot less clear than Kuorikoski et al. suggest. Take the assumption that predators can consume infinite quantities of prey. This is arguably a Galilean assumption, since it assumes that there is no factor (e.g. a biological factor) that affects predator satiation. But a Volterra principle that only applies to target systems in which predators can consume infinite amount of food is clearly not a principle about real-world predators since no *real* predator can consume infinite amounts of food! Hence it seems to me that, at least as far as *some* Galilean assumptions are concerned, if they are not de-idealised from the model, then no matter how many different sets of tractability assumptions we might go through, the theorem that we are actually trying to confirm does not seem to be a theorem about the real world in the first place; indeed, it is not clear what kind of theorem it is at all. But without an unambiguous interpretation of the robust theorem as an empirical hypothesis, the whole idea that we can conditionalise on this theorem (as required by this argument), and thereby confirm it, is, I believe, brought into question.

Second, in this paper I have only discussed a particular type of robustness analysis, one in which the aim is to confirm a robust theorem, interpreted as a causal hypothesis, and the various models in an ensemble involve the same core assumption about the target system and only differ in their tractability assumptions, but are otherwise the same. I did this because the aim of this paper was to assess *Kuorikoski et al.*'s argument for the epistemic import of robustness analysis and that alone. However, there are of course many real cases in science in which models in an ensemble differ substantially from one another in several respects (they may make different empirical and theoretical assumptions about the target system; they may involve various different idealizations of all sorts; they may employ different numerical solution techniques etc.). Global climate model (GCM) ensembles are a good case in point. Although they all rely on well-established theories of mechanical, fluid, and thermodynamics and share many other commonalities, they significantly differ in various ways (e.g. distinct climate models often contain different parametrizations for processes that cannot be directly resolved by the models; they employ different numerical solution techniques etc.). The epistemic significance of agreement across climate models has been, and still is, the subject of considerable interest in the philosophical literature (Lloyd 2009, 2015; Justus 2012; Parker 2011; Vezér 2016; Winsberg 2018; etc.). The purpose of this paper is not to review all the various arguments that have been offered to defend or refute

the epistemic import of model agreement in the context of climate model ensembles. However, I think it will be instructive to have a look at a particular attempt to apply Fitelson (2001)'s account by Justus (2012).

In his reconstruction of Lloyd's (2009) variety-of-evidence argument for the epistemic import of model agreement in climate science, Justus also attempts to apply Fitelson's account of confirmational independence. In particular, Justus attempts to determine whether the assumption that the *models* are confirmationally independent regarding their *common core C* can be used to show that agreement across those models confirms C to a greater extent than either model does separately. He concludes that it cannot, and below is his reasoning:

> Returning to Fitelson's account and making the relevant substitutions, the generalization would require what follows:
> If GCM_1 and GCM_2 individually confirm C and are [confirmationally independent] regarding the (core) hypothesis C, then $c(C, GCM_2\ \&GCM_1) > c(C, GCM_1)$, and $c(C, GCM_2\&\ GCM_1) > c(C, GCM_2)$.[17]
> But this is flawed on many fronts. First, since C is part of GCM_i, the right side of each equation seems to be 0, and the first part of the preceding antecedent, false: GCM_i deductively entails C, but that certainly does not establish that it confirms C. And, second, since GCM_i and GCM_j $(i \neq j)$ are logically incompatible hypotheses about global climate, the left-hand side of each [inequality] seems undefined: the conditionalizations are predicated on an impossible circumstance. (Justus 2012, p. 805)

I agree with Justus that this analysis is flawed on many fronts, but I think these flaws are merely attributable to attempting to apply Fitelson's account in line with an approach (one arguably suggested by Lloyd (2009, p. 221)) which treats *models* themselves, rather than their results, as evidence for a hypothesis. Consider the first flaw that Justus points out, that it is very unclear why a model (interpreted as a complex hypothesis about the climate system) would confirm its core C since a model 'deductively entails C, but that certainly does not establish that it confirms C'. Indeed, given that C was introduced by the modelers in the first place, it is very hard to make sense of the idea that a complex hypothesis that was constructed so as to entail C would confirm it. The second flaw that Justus points out with this analysis is the fact that the distinct models 'are logically incompatible hypotheses about global climate, the left-hand side of each [inequality] seems undefined: the conditionalizations are predicated on an impossible circumstance.' Indeed, if we take *models* to serve as evidence for a hypothesis, and if we understand models to be complex incompatible hypotheses about the climate, then in order to apply Fitelson's account in this instance we would have to assume that a set of incompatible hypotheses about the climate can confirm a hypothesis. But this can't be right since if the probability of the 'evidence' is 0, conditionalization is undefined and hence under any plausible confirmation measure, the confirmatory value of learning this evidence will also be undefined.

Overall, what I (and Justus too) take this to show is that that an approach which assumes that *models* can serve as evidence for a hypothesis is conceptually problem-

[17] In the original quote those inequalities are equalities. But I changed them, as it is a typo.

atic. However, one can improve (if not save) Lloyd's argument by treating models' *results* as evidence for a hypothesis instead. Indeed, by treating models' results as evidence for a hypothesis both flaws in the above analysis seem to be resolved. First, one *might* be able to justify why the observation that the model gives a particular result confirms a hypothesis. But notice that if the aim is to confirm the common core C of the model, then whether the fact that a model gives a particular result could confirm it would seem to crucially depend on whether the result in question *matches* empirical observations in the target system (this seems to be a necessary condition, but certainly not a sufficient one!).[18] Second, by treating models' results as evidence for a hypothesis, we no longer have to conditionalize on incompatible hypotheses about the target system, hence conditionalizations are no longer predicated on an impossible circumstance. So returning to Fitelson's account and making the relevant substitutions, the generalization would now require what follows:

> If R_1 and R_2 (which are the results of GCM_1 and GCM_2 respectively) individually confirm a hypothesis C and are confirmationally independent regarding C, then $c(H, R_2\&R_1) > c(H, R_1)$, and $c(H, R_2\&R_1) > c(H, R_2)$.

In contrast to the analysis above, this one doesn't seem to be conceptually incoherent.[19] However, despite this, there are in fact plenty of reasons to doubt that it is reasonable to assume that the models' results R_1 and R_2 are confirmationally independent regarding C. For as discussed above, if the models in question share idealizations, uncertain assumptions, omissions, etc., there is no reason to assume that their results are confirmationally independent regarding a hypothesis.

I mentioned the above analysis to stress the fact that any attempt to apply Bayesian accounts, such as Fitelson's, to justify the epistemic import of model robustness must be formulated and defended with clarity and precision. Failure to do so is not going to help us increase our understanding of the epistemic import of model robustness. As far as I can see, there is no apparent reason to think that Fitelson's account cannot be applied in the context of climate model ensembles in a conceptually valid way. However, as extensively discussed in this paper, there are plenty of reasons to doubt that it can be applied in such a way that the assumptions on which it relies can be justified.

6 Conclusion

In this paper, I reconstructed and critically assessed Kuorikoski et al.'s argument for the epistemic import of model-based robustness analysis. In Sect. 3, I argued that a causal argument from coincidence for the epistemic import of model-based robustness

[18] Recall that in my reconstruction of Kuorikoski et al.'s argument, I am attempting to confirm a 'robust theorem', not the common core of the models. Hence why I am not relying on the idea that models' results need to match empirical observations in the target system for them to be confirmatory.

[19] By this I don't mean to suggest that Justus wasn't well aware of the conceptual incoherence involved nor that Justus was making some kind of error in attempting to reconstruct what possibly Lloyd could have had in mind. Indeed, it is clear that the aim of Justus's analysis above was to reveal the incoherence of Lloyd's (2009) approach.

analysis is misleading and should be rejected. In Sect. 4, I reconstructed what I take to be Kuorikoski et al.'s official argument for the epistemic import of robustness analysis. I first argued that the validity of this argument relies on substantial assumptions that have not been made explicit by Kuorikoski et al.; I then argued that, even if its validity is not brought into question, Kuorikoski et al.'s argument is not sound, since the sort of probabilistic independence on which it relies is unfeasible in *all* cases of robustness analysis. In Sect. 5, by revising the notion of probabilistic independence imposed on the models' results, I introduced a prima-facie more plausible argument for the epistemic import of robustness analysis. However, despite this prima-facie plausibility, I argued that it is in fact very hard to justify its soundness in most, if not all, cases of model-based robustness.

As mentioned in the introduction, Odenbaugh and Alexandrova (2011) have also objected to Kuorikoski et al.'s argument for the epistemic import of robustness analysis. According to them, 'robustness analysis crucially depends on showing that the assumptions of different models are independent of one another'; one of their objection to Kuorikoski et al.'s argument is that 'reports of their independence have been greatly exaggerated' (ibid., p. 759). But this objection suggests that the independence on which Kuorikoski et al.'s argument relies merely fails *in practice*, rather than *in principle*; this made it easy for Kuorikoski et al. (2012) to dismiss this objection - their argument relatively unharmed. Whereas I hope to have convinced the reader more forcefully that arguments that rely on some sort of probabilistic independence to justify the epistemic import of robustness analysis are implausible in most, if not all, instances of robustness analysis. In particular, I hope to have shown that it is a mistake to assume that models might behave a bit like measuring instruments merely because this seems to fit well with our unquestioned intuitions. In other words, I hope to have shown that in our attempt to understand if, and when, looking at more than one model of the same phenomenon can help us learn about the world, we must, here as ever, rigorously question our intuitions rather than letting them dictate the kind of assumptions we are willing to accept.

Acknowledgements Thank you to Roman Frigg, Liam Kofi Bright, the members of the working models reading group and two anonymous referees for this journal, for helpful comments and suggestions at various stages of this paper's development.

Funding Not applicable.

Data Availibility Not applicable.

Declarations

Conflict of interest Not applicable.

Research Not applicable.

References

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.
Cartwright, N. (1989). *Nature's capacities and their measurement*. Clarendon Press.
Cartwright, N. (1991). Replicability, reproducibility, and robustness: Comments on Harry Collins. *History of Political Economy, 23*(1), 143–155.
Fitelson, B. (2001). A Bayesian account of independent evidence with applications. *Philosophy of Science, 68*(S3), S123–S140.
Frigg, R., & Hartmann, S. (2020). Models in science. In *The Stanford encyclopedia of philosophy*, edited by E. N. Zalta, Spring 2020. Metaphysics Research Lab, Stanford University.
Houkes, W., & Vaesen, K. (2012). Robust! Handle with Care. *Philosophy of Science*, *79*(3), 345–364.
Justus, J. (2012). The Elusive basis of inferential robustness. *Philosophy of Science, 79*(5), 795–807.
Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *The British Journal for the Philosophy of Science*, *61*(3), 541–67.
Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2012). Robustness analysis disclaimer: Please read the manual before use! *Biology & Philosophy, 27*(6), 891–902.
Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 421–31.
Levins, R. (1993). A response to Orzack and Sober: Formal analysis and the fluidity of science. *The Quarterly Review of Biology, 68*(4), 547–555.
Lloyd, E. A. (2009). Varieties of support and confirmation of climate models. *Proceedings of the Aristotelian Society, 83*(1), 213–232.
Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science Part A, 49,* 58–68.
Mayo, D. G. (1986). Cartwright, causality, and coincidence. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1986 (1), 42–58.
Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analyses in economics and biology. *Biology & Philosophy, 26*(5), 757–71.
Orzack, S. H., & Sober, E. (1993). A critical assessment of Levins's the strategy of model building in population biology (1966). The Quarterly Review of Biology, 68(4), 533–546.
Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science, 87*(3), 457–477.
Parker, W. S. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science, 78*(4), 579–600.
Räz, T. (2017). The Volterra principle generalized. *Philosophy of Science, 84*(4), 737–760.
Schupbach, J. N. (2018). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science, 69*(1), 275–300.
Vezér, M. A. (2016). Computer models and the evidence of anthropogenic climate change: An epistemology of variety-of-evidence inferences and robustness analysis. *Studies in History and Philosophy of Science Part A, 56,* 95–102.
Weisberg, M. (2006). Robustness analysis. *Philosophy of Science, 73*(5), 730–42.
Weisberg, M. (2013). *Simulation and similarity: Using models to understand the World*. Oxford University Press.
Weisberg, M., & Reisman, K. (2008). The robust Volterra principle. *Philosophy of Science, 75*(1), 106–131.
Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. B. Brewer & B. E. Collins (Eds.), *Scientific Inquiry and the Social Sciences* (pp. 124–163). San Francisco: Jossey-Bass.
Winsberg, E. (2018). What does robustness teach us in climate science: A re-appraisal. *Synthese*. https://doi.org/10.1007/s11229-018-01997-7.