Idealizations and Partitions: A Defense of Robustness Analysis

Gareth Fuller & Armin Schulz

Department of Philosophy

University of Kansas

Lawrence, KS 66045

USA

gfuller2@ku.edu

awschulz@ku.edu

Idealizations and Partitions: A Defense of Robustness Analysis

I. Introduction

Robustness analysis (RA) continues to be a controversial method of justifying models. Many scientists, social scientists, and philosophers (Weisberg & Reisman, 2008; Kuorikoski *et al*., 2010) still frequently claim that the fact that a model is robust to changes in its assumptions provides conformation of the model (or some of its assumptions). However, other philosophers and scientists have called this view into question (see e.g. Orzack & Sober, 1993, Justus, 2012). A particularly strong recent challenge to the epistemic status of RA is provided by Odenbaugh & Alexandrova (2011) (see also Odenbaugh, 2011), who argue that RA is unable to provide a compelling validation of idealized models: where the idealizations are not discharged, RA is not epistemically compelling, and where they are discharged, RA is not needed.

In this paper, we show that there is much that is right in Odenbaugh & Alexandrova's argument—in particular, we agree that the idealizations of idealized models need to be discharged for any RA of these models to be compelling. However, we also show that there are more ways of discharging idealizations than Odenbaugh & Alexandrova (and prior critics of RA) have let on, and that once this is taken into account, RA can in fact be shown to be confirmationally compelling. To bring this out, we extend the (widely overlooked) response of Levins (1993) to Orzack and Sober's (1993) criticisms of Levins's (1966) defense of RA to the case of idealized models. The upshot is a conclusion that preserves some of the key insights of Odenbaugh & Alexandrova's arguments—while still showing that RA has some important epistemic benefits.

The paper is structured as follows. We begin, in section II, by presenting Odenbaugh and Alexandrova's criticisms of RA. In section III, we present and clarify Levins's (1993) response to Orzack and Sober's initial criticism of his classic 1966 paper. In section IV, we

develop Levins' (1993) response into a defense of RA of idealized models vis-à-vis the criticisms of Odenbaugh and Alexandrova. We conclude in section V.

II. Robustness Analysis and Idealizations: Odenbaugh & Alexandrova's Concerns

RA has been added to the methodological toolkit of contemporary science as a way of resolving the concern that scientific models, generally, involve at least some false assumptions (Woodward 2006). Given this, trust in the conclusions of these—at least partially false—models would seem to be problematic: the epistemic foundation of the models' conclusions seems weak. However, if these same conclusions can be derived with *different* sets of modeling assumptions, these conclusions appear "robust"—i.e. independent of any of the assumptions found in the various models (see e.g. Levins, 1966, for a classic defense of this kind of view). This, in turn, is then supposed to vindicate the common core of the set of models producing the conclusions, the empirical standing of the conclusions, or both. After all, the fact that the conclusions follow from all of these models seems to show that they do not depend on any of the particular (false) assumptions used in the models, but are essentially the result of the shared *core* of the set of models in question. In this way, the robustness of the result is thought to confirm the "robust theorem" from the core causal mechanism at the heart of this set of models to the result in question.

However, two major criticisms of RA have come to be formulated. First, it has been noted that RA would seem to depend on the models in question—or at least some of their assumptions—being genuinely *independent*; otherwise, it would not be the case that the same conclusion is in fact derivable from *different* models. If the models are not independent, their conclusions would really just be the result of *one* model containing a related set of false assumptions. In turn, this would make the epistemic foundation of these conclusions weak—and thus bring back the initial worry that RA was meant to solve. As it turns out, though, making this notion of model- or assumption-independence precise and plausible is far from straightforward

(Odenbaugh and Alexandrova, 2011, 763; Justus, 2012, Sober & Orzack, 1993). If models share a common set of core assumptions but different auxiliaries, are they independent? What, exactly, does it mean for two sets of *assumptions* to be independent of each other—is that logical independence, confirmational independence, or something else? Hence, it seems that the very foundations of RA are unclear.

However, this concern should not be overstated. While establishing the conditions for the independence of the assumptions and models in RA undoubtedly is an important problem, there is no reason to think that solving it is impossible (see e.g. Justus, 2012, which takes steps towards this issue using a Bayesian framework; see also Weisberg, 2006, Parker, 2011, and Lloyd, 2015). Instead, we here follow Odenbaugh and Alexandrova (2011, 758), who state that the "main reason to doubt the confirmatory power of RA is that there is no reason to think that robust theorems by themselves provide adequate representations of actual causal relations." That is, like Odenbaugh & Alexandrova (2011), we here focus on the question of whether (a part of) a set of independent models—however they are to be characterized—can be confirmed by RA. This is a separate issue from the question of how to characterize model independence. If it turned out that the models at stake in a case of RA are not independent, then, while this would be a reason to question the confirmatory power of RA, this would be a separate such reason. In short: our question is whether RA can provide confirmation at all—*even if* its assumptions could be shown to be independent.

The second criticism of RA, then, turns on the fact that it is not clear that RA can provide any kind of confirmation of parts of the models it is focused on.[1] For example, Orzack and Sober (1993) argue—against Levins (1966)—that RA involves a suspect form of reasoning: if we are

[1] There is some debate as to what exactly the confirmation is about. As Weisberg (2006) notes, the confirmation could be from the core assumptions of the model to a predicted consequence, or it could be the other way around. In what follows, we follow Alexandrova & Odenbaugh (2011) and focus on the latter case; however, the conclusions below can be easily reformulated with a view towards the former account.

sure that all of the models in question are false, then—almost by definition—they cannot be confirmed by any robust result. By contrast, if we are not sure that all of them are false, then this uncertainty seems to remain when their robust consequences are taken into account. If there is a model that is potentially true, then the fact that it derives the same result as models known to be false does not lend it support, for the same reason that a set of all false models cannot confirm a robust result. The general worry underlying this second kind of criticism is that it is not clear why the fact that a model's conclusion is robust tells us anything about whether this conclusion or the models producing it are *true*.

This worry gets particularly acute in cases where the models at stake in the analysis involve idealizations: after all, idealizations are *known* to be false.[2] Since it is now widely recognized that idealizations are a near ubiquitous part of science (Morrison, 2015; Morgan, 2012; Cartwright, 1983; Alexandrova, 2008), this sort of case will be in focus in the rest of this essay.

The concerns with the RA of idealized models are brought out well by Odenbaugh and Alexandrova (2011) (see also Odenbaugh 2011). Their argument, in brief, is this. Consider a robust result of a set of idealized models. Since all of these models still include idealizations, the robustness of the result does not speak to the actual conditions that produce this result (Odenbaugh & Alexandrova, 2011). After all, the fact that this result depends on the idealizations in the models implies that the result cannot provide any confirmation of any of the causal mechanisms represented in any of the models—this was the first horn of Orzack & Sober's (1993) dilemma earlier. There would thus need to be some set of idealizations that could be appropriately removed, through de-idealization, to allow for this kind of confirmation to take place: for then, we could obtain a model that accurately represented the relevant aspects

[2] Odenbaugh and Alexandrova (2011) distinguish two types of idealizations (see also Kuorikoski *et al*., 2010): Galilean idealizations—which remove some possible confounding causal factors—and tractability idealizations—which are mathematical "niceties" introduced to ensure the model functions at all. For present purposes, a closer analysis of this is not necessary, though.

of reality. However, if such a de-idealization is in fact possible (and done), RA loses its confirmatory value. If we can show that a given model is based on assumptions that can safely be de-idealized, then of course its core causal assumptions can be empirically confirmed by the data. This would then be a standard case of empirical, model-based confirmation (see e.g. Sober, 2008, for more on this). In a case like this, though, RA has effectively dropped out of the picture: all the confirmatory work is done by the process of de-idealization; the RA is not adding anything to this.

Put somewhat more formally, consider a set of models **M** = {M_1-M_n}. All the models share a core *C*—a representation of some causal mechanism—and a derivational consequence *P*. Each model is further filled out with a set of auxiliary assumptions A_i, which we assume include idealizations. These auxiliaries are assumed to be necessary for the functioning of the model: *P* cannot be produced without them. In this case, *P* is robust, since each model M_i = (C&A_i) in **M** implies *P*. However, for P to confirm C, we would somehow need to discharge at least some set of auxiliary assumptions A_i. If this cannot be done, then we have no reason to see P as confirming C: after all, P cannot be derived from C alone. However, if some A_i can—somehow—be discharged, RA is no longer needed: for we can then just confirm C in the usual way from P and the discharged A_i.

Of course, it deserves to be noted that there is a lot of discussion surrounding exactly how this kind of the discharging of idealizations works (see e.g. Morgan, 2012; Morrison, 2015). However, this is not what is important here. For present purposes, the key point to note is just that RA seems to do no work in the above sort of cases: if idealizing assumptions can be discharged, RA is not needed, and if the idealizing assumptions cannot be discharged, RA lacks the epistemic foundation to be confirmatory.[3]

[3] It is important to note how this criticism is distinct from concerns about the independence of models. Even if the relevant models are all independent, arguments like those expressed by Alexandrova and Odenbaugh suggest that RA fails to be confirmatory. Since none of the models in question are accurate reflections of the target system, drawing a conclusion about the target system from this set is premature.

However, as we try to show in the rest of this paper, there is more to the confirmational power of RA even of idealized models than the above criticism lets on. To bring this out, it is best to begin by briefly considering Levins's response to Orzack & Sober's original criticism of RA.[4]

III. Levins's Response to Orzack and Sober

In a much-discussed paper, Orzack and Sober (1993) criticize the similarly much-discussed defense of RA in Levins (1966).[5] Levins (1993) then presents a reply to Orzack and Sober's arguments that has not received a great deal of attention, but—so we want to argue here—deserves a second look. Indeed, it turns out that a revised form of this response holds some important lessons for the defense of RA especially of idealized models.

Levins (1993) argues that Orzack and Sober (1993) interpreted his initial (1966) argument incorrectly. RA needs to distinguish the different parts of a model. These parts fall into two categories: a common, empirically plausible core C and a set of simplifying assumptions V_i out of larger superset V. If it then turns out that the same result R can be derived from a large number of combinations of (C & V_i) (for different i), which furthermore "exhausts all the admissible alternatives" in the phenomena, we can conclude that C->R is empirically confirmed (Levins 1993, pg. 553).

What we can derive from the RA is that the result can be achieved from many counterfactual, often simplified, situations. What needs to be done, though, is connecting at least one of these counterfactuals to the real world through discharging assumptions. But this would mean that there is no need for the RA. Regardless of how model independence is determined, there remains this further concern.

[4] Weisberg (2006) also responds to Orzack & Sober (1993). However, he does this by noting that "robust theorems" of the above sort are really only *a part of* RA. The latter is about determining the core, shared structure of a set of models that is responsible for deriving an empirically confirmed result, and then using a variety-of-evidence argument to obtain confirmation for the robust theorem in question. However, our point here is that there is more that can be said on behalf of the confirmational properties of RA. See also Justus (2012).

[5] Levins's presentation of robustness had a slightly different approach, as his focus was on whether the prediction M->*P*—"the robust theorem"—could be considered *true*. However, as also noted by Odenbaugh (2011, 1179), much of Levins's discussion can be appropriately tweaked so as to fit the presentation of Alexandrova & Odenbaugh (2011). See also below.

Put differently, the key idea behind Levins's (1993) response is that the empirical evidence available to us may not be strong enough to allow us to confirm the truth of any particular set of auxiliaries over another. However, this evidence may be sufficient to determine a range of admissible variations, such that some sets of auxiliaries can be seen as possible and others as impossible. With this in the background, he then argues that if the set of core assumptions C and each set of auxiliaries V_i implies R, then we can rest assured that C implies R. After all, the sets of assumptions V_1 to V_n cover all possible, acceptable variations—and hence we know that, whatever case may hold, a model with C will imply R.

Now, there is no question that, as it stands, Levins's response needs further development (see also Odenbaugh & Alexandrova, 2011). For example, as Justus (2012) notes, Levins is not clear in how the C and V_i should be distinguished from each other (which is one of the reasons Weisberg, 2006, presented an alternative defense of RA). Most importantly, Levins does not make clear exactly why (if at all) RA is *confirmatory*: what about the fact that the V_i "exhausts all the admissible alternatives" makes it the case that C->R is being confirmed? Given that idealizations are *all* false, it is not clear how the fact they span a range of "admissible variation" makes it the case that the core assumptions in question should be seen as confirmed by the observed (positive) results of the model. Hence, it is not clear exactly how it is Levins's defense of RA is meant to work. However, it is possible to develop this response further to make this clearer.

IV. Idealizations, Partitions, and Robustness

The key idea behind our extension of Levins's response is that his notion of a range of appropriate auxiliaries can, at least at times, be fleshed out as a range of possible idealized variables. The logical space for a variable can then be partitioned into idealizations that reflect the range of possible values for the variable in consideration. While no individual member of the range of acceptable values needs to be able to be seen as accurate—they can all be

idealizations—collectively, they *can* still span the space of possibilities. Importantly, the fact that they span the space of possibilities may be enough to discharge them collectively—with the result that RA has conformational value after all.

To see this, assume that, like Odenbaugh & Alexandrova (2011), we are interested in modeling some phenomenon with a set of models M containing representations of a core set of causal processes C and a set of auxiliary idealizations A_i. Assume further that the C contain free variables that need to be filled in for them to make any predictions whatsoever. Assume also that the true distribution of the values of these variables is very complex (and perhaps unknown), but that it is known that this distribution is in a given range. Then we may be forced to idealize and set the variables to simple values.

However, if it then further happens to be the case that the same prediction follows from C whatever simple value, within the known range of the true, but complex distribution underlying these values, is chosen—i.e. if the same prediction follows from each A_i—then this provides some confirmation for C.[6] After all, while we know that no individual assignment of the variables—i.e. no A_i by itself—is true, we know that, collectively, they cover the space of possibilities here. While we are not able to discharge any one of these idealizations *individually*, we are able to confirm C by discharging them *as a whole*. In this way, we can challenge Alexandrova and Odenbaugh's (2011) argument by noting that no individual model's idealizations needed to be discharged, but that the set of models as a whole might be used to discharge the assumption. In this case, therefore, the need for RA is not undercut by connecting the idealized models to a truer representation of their target.

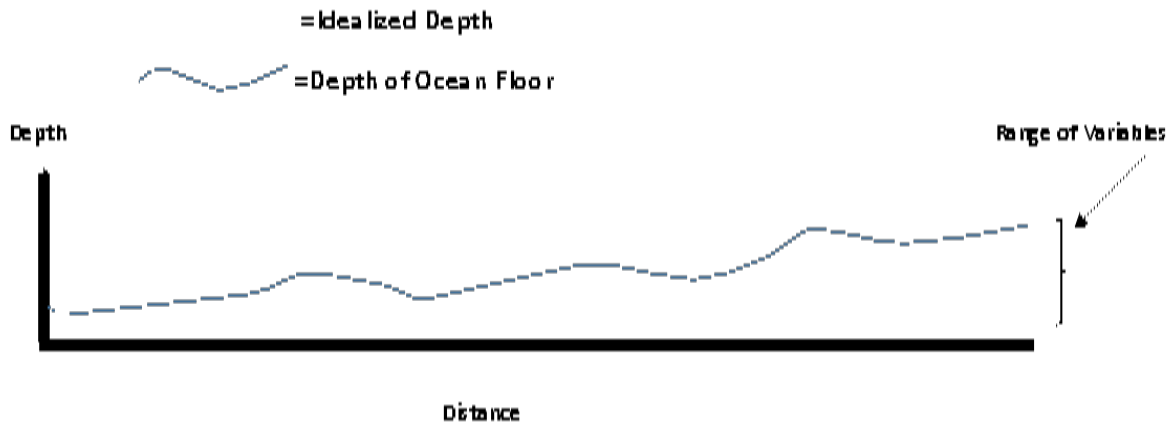
The argument here is best made clear using a stylized example. When modeling the behavior of ocean waves, we may be using a model with a set of core causal assumptions. These core causal assumptions may further need to be provided with a value for the depth for

[6] It is debatable how much conformation it provides. However, this can be left open here; for present purposes, it is just important that it provides *some* confirmation.

the ocean. In deep-water wave models—where the depth is a sufficiently large percentage (e.g. 30%) of the wavelength—we may be able to just idealize this depth by setting it as *infinite*. This does not work for models of wave behavior in shallower water, however: there, we need to set a finite value for the variable representing the depth of the ocean floor.[7] Since the ocean floor varies in depth, though, no single finite value selected here can be an accurate representation of the depth of the ocean floor; every single such finite value can only be treated as an idealization. However, assume that we know that, for the part of the ocean we are modeling, the depths cover a certain range (from between x and y meters, say). [8]

If it then turns out that, whatever idealized value of the ocean floor we set, the same predictive consequence for the resultant wave pattern (amplitude and frequency) results, then that confirms our set of core causal assumptions C. This is so, as the robustness shows that, within the range in question, the exact depth of the ocean *does not matter* for producing the result: whatever shape the ocean floor actually has, all of its depths are featured *somewhere* in our set of models M, and the result of the model holds for all of these depths.[9] Figure 1 makes this clearer.

[7] See Pincock 2011 pg. 99 for a discussion of deep-water wave models.
[8] Note also that the set of idealized values cannot be seen as an estimate of the depth of the ocean floor, as that depth varies. This set of idealized values is literally just the range of the depth of the ocean floor, not an estimate of it at any given point.
[9] Note that if we expected the shape, or change in depth of the ocean floor within our range of possibilities to be causally relevant, this would then be included in our set C, and so the model would be different.



[Figure 1: A stylized example of a set of idealized models of the shape of the ocean floor]

To see why RA is here confirmational, two aspects of the analysis need to be noted. On the one hand, collectively—but not individually—the auxiliary assumptions of these models represent the ocean floor *as it actually is*. Hence, collectively—but not individually—we can discharge these assumptions: the derivation of the conclusion of the model is not called into question by making these assumptions, in so far as these assumptions are taken as a whole. *As a whole*, these assumptions do not misrepresent reality.

On the other hand, since result R is derivable from each M_i in M, and since each M_i is a combination of a shared set of core causal assumptions C and a set of false (i.e. idealized) auxiliaries A_i, there are two options for the status of C that are empirically viable candidates—again, taking into account the set of A_i as a whole:

(A) C is an accurate representation of the true causal process underlying wave formation, and this process leads to the observed wave formation whatever the exact nature of the ocean floor happens to be.

or

(B) C is not an accurate representation of the true causal process underlying wave formation, but it so happens that this misrepresentation of reality still predicts the actually observed wave formation patterns, and does so whatever the exact nature of the ocean floor happens to be.

However, while both (A) and (B) are consistent with the observed data, it is important to realize that a whole slew of interpretations of C are *not* consistent with these data. In particular, what is ruled out is that the derivation of the actually observed wave formation patterns is due to the combination of an *in*accurate representation of the true causal process underlying wave formation together with a (judiciously or fortuitously chosen) false set of assumptions about the depth of the ocean floor. This is due to the fact that, while many sets of assumptions C' can, in principle, derive R if only a subset of the A_i needs to be considered, the RA here shows that R can be derived from *all* A_i.

This can be seen more clearly from noting that reducing the A_i that need to be considered is like reducing the number of data points that a curve needs to fit. There are many more ways of connecting two points (straight line, parabola, etc.) than there are of connecting three such points (in general) (see e.g. Abraham & Ledolter, 2006). Hence, by restricting the A_i that need to be restricted in the analysis, we are freeing up possibilities for causal processes to yield result R. On the flipside, where a result R has to remain derivable under *all* A_i, this thus means that such "gerrymandered" causal structures are ruled out.

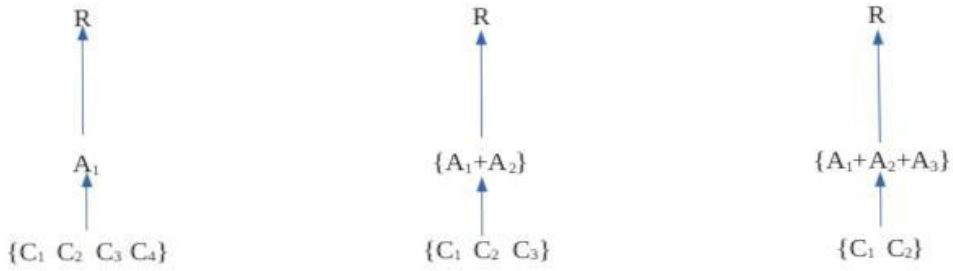
All of this matters here, as it shows that the RA of the present set of models is confirmatory. In general, if a particular wave formation is derived from a false set of core causal assumptions C' with a particular (fixed, unitary) ocean depth, different assumptions about the ocean depth should lead to different such wave formations. Since this is not the case—i.e. since all the M_i have the same result—this raises the probability that this result R is in fact due to the

shared set of causal assumptions C of all the M_i in M: the RA reduces the number of hypotheses that could underlie the data distribution here.

Now, it is important to acknowledge that the probability that C is an accurate representation of the true causal processes underlying wave formation is not raised to 1. As noted above, it is possible that a false set of causal assumptions C' can also combine with the A_i so as to produce R—as in scenario (B) above. For example, it may be that it is in fact the *variation* in the ocean depth that actually causes the observed wave patterns—and not the core causal assumptions C of the set of models in question.[10]

This point, though, should not distract from the fact that the robustness of the models in M (with regard to relevant set of conclusions) *confirms* the set of core causal assumptions C. The confirmation relation is graded and partial, and thus need not lead to full acceptance of any representation (Sober, 2008). What it does is just to *raise* the probability of these representations being true (or our credence in this being so). For this, it does not have to rule out *all* the ways in which the assumptions may be false. It just needs to exclude *some* such ways. This is true here. In the ocean wave example, the robustness of the models in M excludes vast sets of false C': namely all those that can get the observed result only in a subset of the actual ocean depth. Thus, it can be concluded that the RA here raises the probability that C is an accurate representation of reality—i.e. that it confirms it (Sober, 2008). Figure 2 makes this clearer:

[10] Of course, if the variation in the depth of the ocean floor is *suspected* to be causally relevant, then presenting the ocean floor as a single value would be inappropriate: the above set of models then is not a good partitioning of the space of possibilities, and would need to include the variation as well.



[Figure 2: Different causal mechanisms combine with subsets of auxiliary assumption to bring about a given result]

In this way, we can meet the criticism of Odenbaugh and Alexandrova (2011). They argued that we need to discharge our concerns about the idealizations of a model, and that the only way to do so would be with a true set of auxiliaries. As the above shows, though, they are only half right: while we do need to discharge our concerns about the idealizations of a model, it is *not* the case that the only way to do so would be with a true set of auxiliaries. Rather, we can also discharge our concerns about the idealizations of a model *holistically*: the fact that the idealizations *partition* the range of the relevant set of phenomena implies that these idealizations can be discharged *as a whole*. In this way, RA can still be defended as having conformational value (beyond any heuristic usefulness it may have). Note that this is entirely in line with Levins's (1993) response; it is just that all of the A_i here are idealizations. Put differently, the truth here does not lie "at the intersection of independent lies;" rather, it lies *in the range of an idealized partition of the world*.

This point can be applied to more realistic cases as well. Consider Schelling segregation models (Schelling 1971). These models were developed to make sense of how segregation develops in populations. An assumption in these models is that all people in the model have the

same level of preference for diversity, or a lack-thereof.[11] This model can further be run varying the level of preference. In a real community, however, different people will have different preferences for living with others "like them." This may be thought to call into question the importance of the result of the model: what does it matter that a number of unrealistic scenarios lead to segregation? How can this help combat the actual racism pervasive in the world?

However, if we further note that we can *sweep out* the extent of these preferences by varying them continuously as in figure 1, and that we get the same results in all of these cases, this concern can be answered. The fact that all preferences are considered—overall, though not simultaneously—shows that the detailed features of these preferences do not drive the result here. Indeed, the model shows that the core of the causal mechanism underlying the segregation is the fact that there is *some* preference for living with others "like them"—the strength of this preference does not matter. Whatever the *actual* combination of preferences in the world is like, it is covered—holistically—by the fact that the *set* of models contains them all, and has the same result.

To understand this better, two further questions about this argument should be answered.[12] The first question concerns the scope of the present defense. It may be granted that if we can find a range of idealized variables that spans the spectrum of the actual possibilities, the above analysis may go through. However, it may also be thought that this is rarely the case—and that even if it were the case, we are unlikely to know it. For example, some

[11] There are other typical assumptions, such as there only being two "kinds" rather than a broader diversity.

[12] A further question that might be raised here concerns whether our response to Alexandrova and Odenbaugh's argument surreptitiously requires models that lack independence. However, as noted above, we set this question aside here. On the one hand, while, as also noted earlier, the characterization of model independence is far from clear, there is no obvious reason why assuming the ocean floor has consistent depth of x meters and assuming it has a consistent depth of y meters (where x and y are different) are not independent assumptions (for example). On the other hand—and most importantly—we ignore the question of model independence here because, as stated above, the argument presented by Alexandrova and Odenbaugh is meant to show that RA is not confirmatory at all. If the main concern is an issue of model independence, then we can move on from Alexandrova and Odenbaugh's argument as spurious, and the focus should be on defining model independence.

forms of robustness analysis may concern cases that do not allow for a mathematical treatment in terms of free variables that span the range of possibilities. In other words, cases like the ocean wave model above may be the exception that proves the rule: apart from some rare cases like this, the RA of idealized models has no confirmatory value (see also Orzack & Sober, 1993; Justus, 2012).

In response, two points can be noted. On the one hand, there is in fact no reason to think that cases like the wave model example will be very rare or unknowable. To begin with, much of the discussion surrounding robustness analysis concerns what are widely recognized to be mathematical models. This is true for Alexandrova & Odenbaugh's analysis as well as that of Kuorikoski *et al*. (2010), Justus (2012), Parker (2011), Weisberg, (2006), Schupbach (2018), and the classic discussions in Levins (1966), and Sober & Orzack (1993). Furthermore, given the expressive powers of mathematical analysis, many issues should be translatable into a suitable mathematical framework. As far as obtaining evidence for it being the case that the range of idealized variables under consideration spans the spectrum of the actual possibilities: there is no reason to think that this kind of evidence must be impossible to obtain. While it may not always be available, it is reasonable to think that it will be sufficiently often available to make our response interesting and compelling.[13]

On the other hand, we are happy to admit that many of the cases where RA has been appealed to do not meet the requirements above. We are also happy to admit that, in such cases, our defense of RA does not bite. For example, if the auxiliary variables do not cover a well-defined range—e.g. because they are of different kinds, or because their limits cannot be

[13] Note also that Levins (1993, p. 555) responded to the criticisms leveled by Orzack and Sober (and others like them) by arguing that observation is relevant to "the choice of the core model and the selection of plausible variable parts." Odenbaugh (2011) takes this to mean that we need to have empirical evidence for the set of core causal assumptions of the model—and then rightfully points out that, in that case, the idealizations of the model do not really matter, since we have empirical evidence for the representational parts of the model already. However, here, it is the second part of Levins's statement that matters: What we need—and may well have—is evidence for the "variable parts of the model." It is the latter kind of evidence that can give RA a confirmatory role to play.

ascertained—then they cannot be discharged in the above way. However, we want to emphasize that, while the scope of our defense of RA therefore has to be seen to be restricted, there is no reason to think that it will *never* apply either. For instance, as noted above, this above reasoning seems to be apply quite well to Schelling segregation models. More generally, our goal here is not to defend RA at all costs; rather, our point is just that, in a certain set of cases, it does play a confirmatory role.

There is one form of this worry that deserves a closer look, though, as it not only makes the issues here particularly clear, but also addresses an important aspect of much actual modeling work. This form concerns cases where the range of values goes to infinity. Such extremal cases are a common feature of many modeling approaches: for example, many population genetic models assume population sizes is to be infinite (see also Potochnik, 2017).[14] Such cases may seem to be problematic for our arguments, as there is no appropriate range of values to fill in for the variable in question (as there is when the represented ocean depth is finite): there is no upper bound to that range.

However, while it is indeed true that cases like this point to a limitation in the types of case that our arguments can be applied to, it does not imply it can never be applied. On the one hand, some extremal cases can, in fact, be handled by our response: namely, those that can be approached using finite approximations. So, say we have a robust set of models M, some of which assume a finite population size N0, some a finite population size N1 (with N1 > N0), and some an infinite population size (Ninf). To apply our analysis to the case, we can expand this set of models M to replace the model with Ninf with models of population sizes N2, N3, and so on (with Ni > Nj for i>j). In other words, we can treat the infinite case as the limit of a set of finite cases. It is true that this changes the conclusion somewhat: the robustness analysis would then not apply to the models as they were actually formulated, but to a different set of models.

[14] As noted in the example of models of waves, this may even be plausible for some models of wave formation in oceans, where an infinite ocean depth may be assumed.

However, this will often still yield the same conclusion as the initial case: the point of appealing to an infinitely large population size (for example) is just that the actual population is sufficiently large to reduce major chance-based influences. This is well addressed by using a set of models with different *finite* population sizes of different magnitudes.

On the other hand, though, while we do want to acknowledge that it may well be true that this kind of approach will not always work, and that therefore a large number of extremal cases cannot be handled well by our response. However, we also think that, all in all—i.e. taking into account the sort finite approximations just sketched— a sufficiently large set of cases remains to make our analysis interesting and relevant.[15]

The second question that should be answered here concerns how our account differs from other ones in the literature. While several authors have recently provided defenses of RA (see e.g. Justus, 2012; Lloyd, 2015; Parker, 2011), it is especially the account in Schupbach (2018) that deserves a closer look, as this one seems to share some important elements with the account here defended. However, it turns out that these similarities are in fact only superficial.

Schupbach (2018) defends the confirmatory power of RA in terms of ruling out prospective hypotheses or explanations about what produced the result in question.[16] Schupbach begins by arguing that two models are independent from each other if they each rule out different explanations for what produced the results of the models. Given this (or so Schupbach goes on to argue), RA can narrow in possible explanations towards the robust theorem: the more robust a result is, the more potential explanations are ruled out, and thus, the more strongly corroborated the result is.[17]

[15] We thank an anonymous referee for useful discussion of this point.
[16] Note also that Schupbach's account is more general and is intended to apply not just to models.
[17] Incidentally, this is quite in line with Popper's account of corroboration: see Popper (2002, chap. 10).

Schupbach's account thus shares with our account the fact that he grounds the confirmatory power of RA *holistically*: the right explanation is arrived at by ruling out competing alternatives overall. However, importantly, our account differs from that in Schupbach (2018), as our account is not *explanatory* in nature. This matters, as it means that our account can avoid questions over what an explanation is, whether the relevant models can provide explanations (so understood), and whether this leads to an objective or merely subjective account of corroboration. In turn, this is useful, as the latter is a point of major contention in the literature (Skow 2016; Potochnik 2016;). In particular, while many authors conceive of scientific explanations as objective—in the sense that there is a mind- and language-independent fact of whether a model (say) does or does not explain what produces a given phenomenon (see e.g. Woodward 2003; Craver 2007; Skow 2016)—this is not universally agreed (see e.g. de Regt 2009; van Fraassen 1980; Khalifa 2017). Some authors argue instead that explanation is about communicating facts in such a way to an audience that the latter can feel they grasp the phenomenon in question (Potochnik 2016, 2017: Khalifa, 2013, 2017; Grimm, 2010).

This dispute matters here, as it has implications for Schupbach's account of whether RA is confirmatory. In particular, if explanation turns out to be a merely subjective, psychological phenomenon, then, on Schupbach's account, so would the confirmation provided by RA. However, whether this latter claim is plausible is not obvious (Sober, 2008; Brössel & Huber, 2015). More generally, a concern for Schupbach's account is the fact that it ties the confirmatory nature of RA to the nature of explanation. This implies that any uncertainties about the latter will transmit to the former: Schupbach's characterization of the confirmatory value of RA is only as strong as our understanding of the explanation relation. Since the latter is still a majorly disputed topic, this thus introduces much uncertainty into Schupbach's defense of the confirmatory value of RA.

Instead, our account just focuses on the fact that, in some cases, a set of idealized models can *collectively represent* a complex reality. If it is then the case that the same result

follows from each member of this set of models—i.e. if the set of models is robust—this would make it more likely that the set, as a whole, also accurately represents the mechanisms driving this result. This thus shows that the false assumptions of these models can be collectively discharged. Importantly, this conclusion about the confirmatory nature of robust idealized models is derived without appeal to how "explanatory" the relevant models are. In this way, our account has similar benefits to that of Schupbach (2018), but without the drawback of incorporating further controversial concepts.

V. Conclusion

We have argued that it is possible to discharge idealizations in a set of models holistically by using RA. Taking a lesson from Levins, we have shown that in cases where we can partition the world into set of idealized values for a variable and then run these values through a RA, we are able to get confirmation for the core set of causal assumptions in the models in question. As Levins (1993, 554) puts it: "The search for robust theorems reflects the strategy of determining how much we can get away with not knowing, and still understand the system."[18] What we have tried to show is that we do not need to know *which* idealization can be successfully de-idealized. Rather, all we need to know is that the set of idealizations *cover the spectrum of possibilities*. In this way, we can show one of the assumptions of Alexandrova and Odenbaugh's analysis to be mistaken: they assume that the only way to discharge an idealization is with a truth. However as we have showed here, it is at least sometimes the case that idealizations can be discharged as a whole: we do not need to find a model with a reasonably accurate depiction of reality; all that we need to know is that that the set of models we are working with *collectively* contains such a description. In a nutshell: RA can be vindicated where the idealizations it is focused on partition the space of possibilities.

[18] Levins 1993 pg. 554.

Bibliography

Alexandrova, A. (2008). Making Models Count. Philosophy of Science, 75(3), 383-404.

Abraham, B., & Ledolter, J. (2006). *Introduction to Regression Modeling* (1st ed.). Independence, KY: Cengage.

Brössel, P., & Huber, F. (2015). Bayesian Confirmation: A Means with No End. The British Journal for the Philosophy of Science, 66(4), 737–749.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford : New York: Clarendon Press ; Oxford University Press.

Craver, C. (2007). Explaining the brain : Mechanisms and the mosaic unity of neuroscience. Oxford : New York : Oxford University Press: Clarendon Press;.

De Regt, H. (2009). Understanding and Scientific Explanation. In De Regt et al.

Regt, H. W. de., Leonelli, Sabina, & Eigner, Kai. (2009). Scientific understanding : philosophical perspectives. Pittsburgh: University of Pittsburgh Press.

Grimm, S. (2010) The goal of explanation. Studies in History and Philosophy of Science. Part A, 41(4), 337–344.

Grimm, S. (2018). Making sense of the world : new essays on the philosophy of understanding. New York: Oxford University Press.

Justus, J. (2012). The Elusive Basis of Inferential Robustness. *Philosophy of Science, 79*(5), 795-807.

Khalifa, K. (2013) The Role of Explanation in Understanding. The British Journal for the Philosophy of Science64(1): 161-187

Khalifa, K. (2017) Understanding, explanation, and scientific knowledge. Cambridge; New York: Cambridge University Press

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic Modelling as Robustness Analysis. The British Journal for the Philosophy of Science, 61(3), 541-567.

Levins, R. (1966). The Strategy of Model Building In Population Biology. *American Scientist, 54*(4), 421-431.

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. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science, 49*, 58.

Morgan, M. (2012). *The world in the model : How economists work and think*. Cambridge; New York: Cambridge University Press.

Morrison, M. (2015). *Reconstructing reality : Models, mathematics, and simulations* Oxford: Oxford University Press.

Odenbaugh, J. (2011). True Lies: Realism, Robustness, and Models. *Philosophy of Science, 78*(5), 1177-1188.

Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analyses in economics and biology. *Biology & Philosophy, 26*(5), 757-771.

Orzack, S., & 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. (2011). When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science, 78*(4), 579-600.

Pincock, C. (2011). *Mathematics and scientific representation*. Oxford ; New York: Oxford University Press.

Popper, K. (2002). The Logic of Scientific Discovery. London, UK; New York: Routledge Classics.

Potochnik, A. (2016). Scientific Explanation. Putting Communication First. Philosophy of Science, 83(5), 721–732.

-(2017) Idealization and Aims of Science. Chicago, US: University of Chicago Press.

Schelling, T. C. (1971). Dynamic models of segregation. Journal of mathematical sociology, 1(2), 143-186.

Schupbach, J (2018). Robustness Analysis as Explanatory Reasoning. The British Journal for the Philosophy of Science, 69(1), 275–300.

Sober, E. (2008). *Evidence and Evolution : The Logic Behind the Science*. Cambridge, UK ; New York: Cambridge University Press.

Skow, B. (2016). *Reasons Why.* Oxford; New York: Oxford University Press

Skow. B. (2017) Against Understanding (as a Condition on Explanation), in Grimm 2018.

van Fraassen, B., 1980, *The Scientific Image*, Oxford: Oxford University Press.

Weisberg, M. (2006). Robustness Analysis. *Philosophy of Science, 73*(5), 730-742.

Weisberg, M., & Reisman, K. (2008). The Robust Volterra Principle. *Philosophy of Science, 75*(1), 106-131.

Woodward, J. (2003). Making things happen : A theory of causal explanation (Oxford studies in philosophy of science). New York: Oxford University Press.

Woodward. J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13*(2), 219-240.